Fully Moral Artificial Agents: Future or Fiction.

by
Jaimee Moore
MRXJAI002

Supervisor: Dr. Ryan Nefdt.

March 2024

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is our own.
2. I have used the UCT referencing guide for citation and referencing. Each contribution to, and quotations in this essay/report/project from the work(s) of other people has been contributed and has been cited and referenced.
3. This essay/report/project is my own work.
4. I have not allowed and will not allow anyone to copy our work.

Signed by candidate

Jaimee Moore

MRXJAI002

ABSTRACT

The rapid technological progress has resulted advancements in technology that includes the development of Artificially Intelligent agents (AI) for various purposes. The extensive progress in AI has reached a point where the introduction of AI agents as android robots or humanoids into society is plausible. This highlights the evolving relationship between humans and technology.

This thesis explores the multidimensional aspects of artificial intelligence (AI) by examining the intersection of machine ethics, ethical theories, and deep learning. The central focus is on assessing the compatibility of ethical theories with the rapid advancements in technology, particularly the potential development of fully moral artificial agents.

The aim of this thesis is to address the ethical concerns associated with the evolution of AI, particularly the introduction of artificially intelligent agents into various societal roles, such as care robots in healthcare. The need for an impartial approach to address these concerns is identified, leading to the proposal of machine ethics as a framework. Machine ethics, defined as ensuring moral behaviours in AI, provides a basis for evaluating the ethical capabilities of artificially intelligent agents.

As technology continues to progress, the demand for a comprehensive ethical framework for AI decision-making becomes increasingly apparent. Machine ethics offers insights into the ethical processes involved in AI decision-making, allowing a closer examination of computational abilities and ethical capacities. The primary inquiry revolves around whether AI agents can achieve full morality and if existing ethical theories can govern their behaviour effectively.

To explore these questions, the thesis draws upon three ethical theories – Utilitarianism, Deontology, and African Ubuntu Ethics – and their applicability to the development of fully moral artificial agents. It employs deep learning as a critical component in understanding moral agency within the context of AI. The analysis unfolds in four parts: first, providing a background on AI concepts; second, examining recent AI progress and the potential for ethical AI agents; third, exploring the future of AI in relation to morality through ethical theories; and fourth, establishing criteria for considering AI agents as moral agents.

In conclusion, the thesis argues that, with the advancements in technology and the insights provided by machine ethics and ethical theories, the development of fully moral artificial intelligent agents is possible.

Contents

Introduction

This thesis will analyse the dimensions of artificial intelligence through machine ethics with reference to the compatibility of ethical theories and deep learning with the possibility of the development of fully moral artificial agents. In this ever-changing world we live in, there are many factors and tools we as human beings make use of to simplify our lives. Technology has been one of the most influential tools used to do so. It is a tool we make use of daily to alleviate frequent challenges we might face. This ranges from certain devices we use, to the social media platforms we interact on. Over the years, the rapid progression of technology has been very beneficial to navigating and managing our daily lives. The developments in technology have stretched so far that there are now devices integrating with social platforms and can be programmed to not only remind us of daily tasks but to perform them as well, such as Alexa, Siri, and possibly even self-driving cars in the near future. Computer scientists and engineers have gone as far as developing artificially intelligent agents (AI) for these purposes. Additionally, we also see the progress in the different branches of AI like the programs used for screening purposes in banks and often larger companies when selecting job applicants[1]. Although these are different divisions of AI, more and more progress is being made in the field of artificial intelligence.

These developments have progressed so far that there is a possibility of introducing artificially intelligent agents as android robots or humanoids into society. More specifically, introducing machines as medical aids like care robots, as described by Johansson-Pajala, et al., "we refer to care robots as machines that operate partly or fully autonomously with the aim of supporting potential users, older adults and relatives as well as professional caregivers, in providing physical, cognitive or emotional support" (2020, p. 1103). As such, it is clear that as humans, we have already developed a type of relationship with this kind of technology. However, with the advancements of technology into the world of AI, there are a few ethical concerns that needs to be addressed.

The ethical concerns span from the ethical limitations of these types of technology, to whether these humanoid devices can be produced to be agents with full ethical capacities and duties[2].

[1] See Wesche and Sonderegger (2021) on the reality of AI on job seekers. This could result in negative outcomes either due to the decrease of applicants or due to the direct outcome of the automation process.
[2] Apart from its capacities, we also need to consider the appearance of these agents, as the potential for discord can arise. The 'uncanny valley' suggests that there is a correlation between interaction between these agents and

It also includes concerns that pertains to the construction and assembly of these devices as well as the operational and functional aspects of these devices, specifically to the social and psychological impact that it would have on society. These concerns are related to the way these devices are created, the "genetic" or inherent makeup of these devices, and the way in which these devices will run and the functions it would perform in society, as well as how these devices will be classified or identified as. There are also other ethical considerations that needs to be addressed, such as the principles these devices would be programmed to uphold. These concerns are linked to the way in which the technology and programming of these devices plays a role in the interaction and relationships of the agents and individuals.

With that in mind, what is needed is an impartial way to deal with these concerns relating to the interactions between agents and individuals. This, I believe, can be done, through machine ethics. Machine ethics is defined as, "a part of the ethics of artificial intelligence concerned with adding or ensuring moral behaviours of human-made machines that use artificial intelligence" (Anderson & Anderson, 2007, p.15). It looks at the principles that are instilled in these devices and creates the standards on which to judge them, it evaluates whether these machines are capable of making decisions based on their programming as well as independently of the explicit programming.

The growth in technology emphasises the need for a more comprehensive account for the ethical decision making of machines. Traditionally, it is in the hands of the computer engineers who create these artificially intelligent agents, to instil these agents with the correct software and programs, which can be used to make these decisions. This is where machine ethics can be used to provide insight to the ethical processes involved in the decision making of machine. Since machine ethics is concerned with adding or ensuring moral behaviours of human-made machines that use artificial intelligence, it can be used as a basis for looking at the components that governs artificially intelligent agents from its computational abilities to its ethical capabilities. Given the ethical issues surrounding these agents, the main concern will therefore be whether AI agents are capable of full morality and if it is possible, whether current ethical theories can be taken further to govern their behaviour. The use of machine ethics to explore these concerns will shed light on the future implications of the current progressions of artificial intelligence and in addition generate discussions that look at the

humans and the appearance of these agents. It is the idea that a human appearance or behaviour can make an artificial figure seem more familiar to viewers. The sense of viewer familiarity drops sharply into the uncanny valley once the artificial figure tries but fails to mimic a realistic human.

compatibility of technology with ethics and human interactions.

In light of these requirements, there is a clear need for an ethical theory to regulate the actions of machines, as well as that of humans towards machines. This thesis will draw from three ethical theories that may contribute to the development of fully moral artificial agents. It will make use of its ethical principles to show the reality of the possibility. It will involve research pertaining to the introduction of artificially intelligent agents into society and if it can be done successfully. The research looks at ways through which we can better equip these machines and agents to run in ways that is more conducive for humanity. This will be done by looking at the compatibility of technology with ethics and humanity through addressing the ethical concerns by making use of machine ethics and the principles of the ethical theories.

This thesis will analyse the dimensions of AI through the machine ethics whilst drawing on ethical theories and deep learning and its progressions to the development of an agent's capacity for morality. This will be done in four parts, the first will be the background of AI where I will look a few key definitions and concepts in AI such as GOFAI, Weak AI and Strong AI. I will also be making use of John Searle's Chinese Room argument to expand on this. From there I will have an in depth look at the ethical concerns stated above and touch on the impact it has for the generation of AI agents. I will also be drawing on Deep Learning to show the advancements in connectionism and its relation to moral agency.

The second part will consist of the way AI has progressed in recent times and explore the possibility of ethical AI agents. I will look at the different kinds of artificial intelligence and the development of Machine and Data Ethic as well as and the impact it has on AI. I will then go on to describe Moor's (2006) distinction between the three different types of artificial intelligent agents that is, the implicit ethical agent, explicit ethical agents, and full ethical agents. I will show how Moor's distinction in distinguishing the different types of artificial intelligent agents can be pivotal for its development through providing a deconstruction of Moor's grades of Artificial ethical agents.

This then leads me to the third part that will look at the future of AI in relation to morality. This will be done by looking at three ethical theories and its principles, which may be used to analyse the development of artificial intelligent agents. The theories that I will be looking at are Utilitarianism, Deontology and African Ubuntu Ethics and then I will provide an analysis of these theories in relation to moral agency in the next section.

In the fourth part I will look at the criteria by which artificially intelligent agents may be

considered moral agents. I will do this by reflecting on the relationship between deep learning AI and moral agency, through making use of the guiding principles that are prominent in the three ethical theories. Finally, I will conclude by looking at the possibility of developing a fully moral artificial intelligent agent and arguing that with the progress that has been made, there is a possibility of developing a fully moral artificially intelligent agent – that it is the future.

PART 1: Background and concepts in AI

Artificial intelligence is a field that was established in the late 1930's and was believed to be founded by Alan Turing. Turing believed that machines could mimic the behaviour that occurs biologically in human minds. In 1950 he introduced the Turing Test as an attempt to promote the possibility of advanced AI through the idea that the Turing Machine is like a mind. In 1980, John Searle drew on Turing's theory and argued that the possibility of advanced AI is one that outweighs the chances of Strong AI, which is the notion that human level intelligence can be achieved by machines. However, progressions in AI have led to the development of varying degrees of AI and other AI related concepts. In what follows, I will look at the definitions of these concepts. And offer further discussion on these concepts in Part 2, section 1.

Section 1: Important Terms and definitions

Artificial intelligence is intelligence that is manufactured as opposed to naturally produced, it is a type of intelligence that is scientifically made and that is not innately associated with a species. According to Margaret Boden, "artificial intelligence seeks to make computers do the sorts of things that minds can do (2016, p.2)". From this, degrees of AI have been produced.

GOFAI, or Good Old-Fashioned AI, is a branch of AI that uses formal symbolic representations as a basis of operations through programmed instructions. According to Margret Boden (2014), "GOFAI symbols are items in a formal language (a programming language) like the symbols of mathematics or logic, GOFAI symbols and can be regarded as purely formal (meaningless) structures – and programs composed of them" (p. 89). It refers to the use of symbolic means to produce a representation of the desired outcome, using

meaningless symbols which together creates the desired meaning like programming. Until very recently, GOFAI was the main form of AI, until progress in the study in of AI has led to its greater success, resulting in it surpassing GOFAI.

Weak AI is one of the degrees of AI that has been produced. It is defined as "the use of algorithms to accomplish specific problem solving or reasoning tasks that do not encompass the full range of human cognitive abilities" (Khillar, 2020, p.3). It only focuses on problem specific tasks and does not involve the need for understanding, knowledge, or intelligence of any sort. Programs like Siri, Alexa and/or Google search are examples of Weak AI. These programs only perform task specific actions.

Contrast this with Strong AI which is holds the view that machines can development a type of consciousness that is equal to that of human beings. According to Sagar Khillar (2020, p.3), "Strong AI refers to the theoretical form of machine intelligence that views machines or programs, with the mind of their own and which can think and accomplish complex tasks on their own without human interference". It is the notion that Strong AI may develop human level consciousness. It is necessary to note that Strong AI is defined in many ways, however, for this thesis, I will be drawing on the definition by Khillar stated above.

AGI is another concept in AI. It involves the supposed possibility of an intelligent agent acquiring of the intellectual capacity equal to that of humans. It is defined by Hal Hodson as, "the hypothetical ability of an intelligent agent to understand or learn any intellectual task that a human being can" (2019, p.2). It further involves the agents relying on general knowledge as humans do, to determine an outcome. When faced with unfamiliar tasks, with this type of AI the entities have capacity to use software with the cognitive abilities generally associated with humans to find a solution.

With all the discussion on the creation of intelligence, it is necessary to note that the concept of intelligence itself is one that requires the entity gain and apply knowledge through understanding. This leads me to the distinction between syntax and semantics.

Syntax refers to the arrangement of languages and systems, as well as symbols, to illustrate the desired presentation regardless of the meaning attached to it. It does not require understanding, knowledge or meaning. It makes use of the necessary approach to convey the information it wishes to and does not pay any regard to interpretation or meaning held therein. There are rules for combining symbols and these symbols are represented a certain form and do not any meaning. In this sense, syntax is about the correct sequencing or combination of

symbols according to predefined rules, irrespective of the content or significance of those symbols.

Semantics on the other hand, is concerned with meaning. It involves the interpretation of the illustrated presentation for knowledge and understanding of its reference and meaning. It requires that the presentation hold a sense of reference and intension. Reference is the link between the names of objects and the pronouns and nouns that describe them. For example, when we say something about someone or something, and later referring to it the person or thing as 'it'. Intension refers to characteristics (word, phrase, symbol) that are inherently attached to the specific concept or word. The definition of the word is attached and implies its intention. For example, the word dog implies having fur, barking, etc.

I will now make use of Searle's Chinese Room Argument to further distinguish the concepts mentioned above. In 1980 John Searle argued that it is possible to achieve a type of AI that is advance but not quite as advanced as Strong AI suggests. This is done by way of his Chinese Room Argument. The Chinese Room Argument proposed by Searle is a thought experiment where an individual (which in this case is himself) is seen inside a room and receives inputs from the outside requesting a translation from Chinese to English. However, the person in the room does not know or understand any Chinese. The translation is done by consulting a manual (rulebook) that holds the Chinese symbols and the English equivalent thereof. The person inside the room then produces outputs of English translations based on this computation (linking Chinese symbols to its English equivalent).

What Searle wishes to demonstrate with this argument is that what is happening in the room is not knowledge creation and understanding, instead it is the manipulation of symbols through association. According to Searle "in the Chinese case I have everything that artificial intelligence can put into me by way of a program, and I understand nothing; in the English case I understand everything, and there is so far no reason at all to suppose that my understanding has anything to do with computer programs; that is, with computational operations on purely formally specified elements. As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. They are certainly not sufficient conditions, and not the slightest reason has been given to suppose that they are necessary conditions or even that they make a significant contribution to understanding" (Searle, 1980, p.418). According to Searle, for technology to be intelligent in

any form, it must be able to understand the relevant task at hand and perform based on its understanding of the task. For Searle, to be able to understand that Chinese outputs need to be produced is not enough, the Chinese itself needs to be understood. However, Searle holds that what is happening in the Chinese Room Argument does not constitute understanding of Chinese but that what is happening in the room is merely symbol manipulation, it is syntax – form without meaning. He believes that consulting the "rulebook" does not make up understanding of the outputs being produced. For Searle, to program a digital computer to mimic understanding, and make it appear as to understand language, does not automatically guarantee the production of knowledge and genuine understanding.

Searles argument against Strong AI in The Chinese Room Argument can be understood in as the following,

1. For Strong AI to be possible, the computing system should understand the Chinese that it produces – to produce an output, the input needs to be understood.
2. However, through symbol manipulation I could run a program for Chinese without understanding any Chinese – it is a process information association (symbolic management)
3. Symbol manipulation is not understanding – having produced the correct output does not equal understanding the input.
4. As such, Strong AI is not possible.

Searle's Chinese Room argument portrays the degrees of AI discussed above. Although the argument centres around understanding, it can be used to illustrate the different levels of artificial intelligence. The Chinese Room Argument shows the inner workings of computational systems, as such, it can be seen as a representation of artificial intelligence. It shows how what is produced is not Strong AI or even AGI but merely Weak AI with only syntax and no semantics.

Recall the definitions stated above that Strong AI can accomplish complex tasks by using what is believed to be 'a mind of its own' to produce the appropriate response and AGI has the capacity of an intelligent agent to understand and produce generalized knowledge involving multiple tasks, Searle's Chinese Room Argument shows that this is in fact not the case. The output produced stems directly from a systematic computed response made possible through

programming and not the fact that the system has a mind of its own as Strong AI suggests or that the system has understood and learnt intellectual tasks as AGI posits. What it does show is that the system merely makes use of programmed instructions and symbolic representation to produce an output as set forth by GOFAI. It is, in fact, Weak AI. The outputs produced are reached by using algorithms to achieve the necessary problem specific outcome. On the notion of syntax and semantics, we see that much like GOFAI and Weak AI, what happens in the Chinese Room is just syntax. The system makes use of languages and systems with no regard to interpretation or meaning. It lacks semantics as it fails to show how knowledge and understanding is produced with any reference or intension.

Cognitive science then added to artificial intelligence through the theory of connectionism. Connectionism as defined by Buckner & Garson, refers to "a movement that hopes to explain intellectual abilities using artificial neural networks" and, "neural networks are simplified models of the brain composed of large numbers of units (the analogs of neurons) together with weights that measure the strength of connections between the units" (2019, p. 2). It holds that the neurons in the brain form a type of neural network that is responsible for the mental processes that occur. It regards mental processes as one can be understood and interpreted through these neutral networks. This then led to the introduction of what we call deep learning today. And since deep learning no longer has evident connections to cognitive science anymore, it steers clear of the challenges by practitioners of GOFAI such as Fodor and Pylyshyn. That being said, the exploration of these concepts generates questions regarding these distinctions within the realm of artificial intelligence and those that characterize moral beings.

This thesis focuses on whether artificial intelligence can produce an agent that is fully moral. Given the above definitions and distinctions of artificial intelligence, we are left to consider what exactly intelligent agents are. Intelligent agents are agents that are programmed to make decisions or perform an action. These agents are programmed to do so based on user input, experience, and its environment. Given our understanding of intelligent agents, it's important to consider how these entities can embody morality. This prompts an exploration into the definition of moral beings. For this purpose, I will be making use of Angus Taylors' definition for moral agency on which to make the determination of whether artificial agents can be made moral agents. According to Taylor, "moral agency is the ability to make moral choices based on some notion of right and wrong and to be held accountable for these actions" (Taylor, 2003,

p. 20). A moral agent would then be one that makes moral decisions based on his/her own perception of right or wrong and is responsible for the repercussions of the decision that was made. If one considers morality to be something like syntax, involving the application of rules, then all agents can be seen as moral – even Weak AI agents. However, if it requires something more like the understanding of consequences or consciousness (like semantics with reference and intension), then Searles conclusion on Strong AI might follow for morality as well. Now that both artificial intelligence and moral agency has been defined, what is needed is a way to determine if it is possible to create a fully moral artificially intelligent agent. I will start with relation between deep learning and moral agents.

Section 2: Deep Learning AI and Moral Agency

Deep learning is an area of research in AI that looks at how artificial neural systems that are capable of making correct decisions based on data can be created. John D. Kelleher notes that the field of deep learning focuses on the creation of large neural networks that can make accurate decisions based on datasets containing millions of texts and parameters, "deep learning is particularly suited to contexts where the datasets are complex and where there are large datasets available" (Kelleher, 2019, p. 1). According to Kelleher, the decisions that are reached are data-driven by the systems being able to distinguish and extract distinct patterns from datasets. This is done through the accurate association of complex inputs to good decision outcomes. What this implies is that deep learning makes use of datasets to form artificial neural networks that can guide the decision-making process and makes it possible for the machine to reach the most accurate choice regarding what action to perform.

The way this works is through a particular algorithm. An algorithm[3] is defined as, "a process (or recipe, or program) that a computer can follow and outlines a specific process to analyse a dataset and identify recurring patterns in the data" (Kelleher, 2019, p. 7). The connection between these processes is known as functions. The concept of a function is integral to deep learning and is described as, "a deterministic mapping from a set of input values to one or more output values. The fact that the mapping is deterministic means that for any specific set

[3] The matter concerning algorithms is a notable issue, however, it is beyond the current scope. Therefore, although there are challenges with ensuring algorithms accurately track intentions, advancements in interpretability techniques within deep learning are addressing the lack of transparency. Additionally, AI's ability to analyse vast datasets and identify complex patterns allows for sophisticated understanding and application of ethical principles. Thus, although caution is warranted, dismissing AI's potential for moral development outright fails to notice the progress and possibilities within the field.

of inputs a function will always return the same outputs" (Kelleher, 2019, pp. 7-8). This allows for the accuracy of decision-making as it ensures that the input produces the same output as intended or as needed based on the initial data inputted. With the same basic architecture of early connectionist networks, but much more data and thousands of additional hidden networks (that pass on information), deep learning has revolutionised AI.

The introduction of deep learning to AI has contributed to the success of this area in research as it knows what to measure and how to measure it. As Kelleher notes, given the fact that it can produce accurate outputs by taking in features from raw data that are at a lower level and complex nonlinear mappings inputs, provides AI with the capacity to make these accurate determinations. However, according to Kelleher, "this ability is dependent on the availability of large datasets; when such datasets are available, deep learning can frequently outperform other machine learning approaches, additionally, this ability to learn useful features from large datasets is why deep learning can often generate highly accurate models for complex domains, be it in machine translation, speech processing, or image or video processing" (2019, p.35). As such what deep learning has done is revealed the potential of big data and by the integration of its models to consumer devices, it has largely impacted the way in which data is understood and made use of.

Furthermore, what is also necessary to note is that all this rests on the neural networks in deep learning. According to Gary Marcus (2018), "neural networks in the deep learning literature typically consist of a set of input units that stand for things like pixels or words, multiple hidden layers (the more such layers, the deeper a network is said to be) containing hidden units (also known as nodes or neurons), and a set output unit, with connections running between those nodes" (p.4). This makes it possible produce outputs that perfectly correlates with the given inputs and is fundamentally the basis for the data-driven decision-making of the machine. Apart from this, neural networks allow for deep learning to create internal representations that aids in the process of decision-making, "it is also known for its ability to self-generate intermediate representations, such as internal units that may respond to things like horizontal lines, or more complex elements of pictorial structure" (Marcus, 2018, p. 4). This means that neural networks have the capacity to create internal representations that guides the decision-making process by capturing relevant patterns and feature from external data as the default input data. This enables it to adapt to and accommodate all types of data, ultimately, strengthening the possibility of going beyond GOFAI syntax-driven models.

Given the success of deep learning, I now move on to consider how this development may impact the concept of moral agency. On Taylor's account of moral agency, it requires that the agent be able to make decisions and justify them by taking accountability for them. This demands that the agent calculate the correct action in ethical dilemmas using ethical principles and that the agent possess characteristics of intentionality and free will. The question then is, are artificial agents capable of doing this? We can see that with deep learning the possibility of the developing of these types of agents are strengthen.

In principle, one could imagine these agents being able to make ethical decisions and justify them, we see that the neural network of deep learning allows for this to be done[4].All that is needed is large sets of data on ethical decisions and the justification of these decisions to be inputted into these networks. What will be produced is the desired output and the agent will be able to make ethical decisions and be able to justify them. What further advances deep learnings position is the fact that it is able to learn from and use raw data and multifaceted mappings from inputs to outputs. Furthermore, on the condition that these agents should be able to use ethical principles to calculate the correct action in an ethical dilemma, the fact that deep learning is able to self-generate intermediate representations, that it is able to create internal representations by capturing relevant features and patterns from the input data, promotes the use of deep learning as a basis on which to build moral agency. This ability also allows for these agents to portray a sense of intentionality and free will, as it ultimately leaves the final decision regarding the action to be performed to the agents as such promoting free will and intentionality.

Consider the case of self-driving cars. These are vehicles that are equipped with a combination of technology systems such as cameras, sensors, AI, etc. and can perform the task of driving without the need for human intervention. The technology the car has, enables the vehicle to utilize it to understand and adapt to the environment. This consists of the vehicle knowing the not only the road it would need to take and the possible obstacles on it but also the rules of the road and making decisions in the moment to ensure a safe trip. It is their ability to learn and adjust to scenario specific reactions that makes these vehicles competent to handle the road. By using the technology that is embedded in them, like AI and deep learning technologies, these vehicles can adapt to evolving traffic scenarios and changing road conditions, thus making driving safe, convenient, and productive. The use of AI and deep

[4] See Pereira and Saptawijaya 2016 on the Programming of Machine Ethics.

learning technologies provides these vehicles with many crucial functions such as predictive modelling, sensory and perception capacities, natural language for the vehicles that using voice recognition to interact with the passengers as well as accurate decision making so that these vehicles can be seen as a safer way to travel. As Sudeep Srivastava notes, “as self-driving cars continue to evolve, their growing sophistication in understanding, adapting to, and navigating complicated real-world scenarios highlights the transformative potential of AI in the automotive industry” (2023). As such we see how the progress in AI can change the automotive industry and how it can influence how we use and think about vehicles.

Another example of AI making moral decisions is the legal setting, using COMPAS. COMPAS, which stands for Correctional Offender Management Profiling for Alternative Sanctions, is a tool that is used to support the decisions that are made by the courts regarding the likelihood of a convicted offender to reoffend. It makes use of algorithms that are designed with behavioural and psychological constructs of high recidivism to determine the defendants’ outcome on the offence that was committed. According to Northpointe’s Practitioner's Guide to COMPAS Core, “the general recidivism scale is designed to predict new offenses upon release, and after the COMPAS assessment is given” (2015, p.26). What is necessary to note, is that in the case of COMPAS, it has been seen that the outcomes reached was severely misaligned with the reality of the matter. This is due to the fact, as stated by Northpointe, “the scale that is used, uses an individual's criminal history and associates, drug involvement, and indications of juvenile delinquency” (2015, p.26), which does not accurately reflect the possibility of an individual reoffending. The recidivism rates were in fact inaccurate as it negatively reflected on defendants of colour as it was racially discriminating against criminals. It is important to note that the refined algorithm of deep learning produces various aspects that can be adopted and integrated into various aspects, “as deep learning algorithms become increasingly sophisticated in processing complex information, their applications have bourgeoned across chemical engineering, including in molecular design, computational chemistry, control, and optimization” (Alshehri et al., 2020). That being said, deep learning might in fact be the ground on which to build the programs and processes needed for AI agents to be considered moral agents. If deep learning is rooted in ethicality and this is applied to the agents decision-making process, it might be possible to produce an agent that is moral, in the sense of being capable of making moral judgements.

Above I looked at the different types of AI and how it has developed over the years. Its growth has introduced other dimensions like deep learning to its success and resulted in it becoming

a sustainable tool for the future of technology. In what follows I will look at how this progress can be used to develop ethical AI agents. More specifically, it will look at the possibility of introducing fully ethical AI agents into society and arguing for ways in which it can be done successfully.

PART 2: The possibility of AI ethical agents.

There are many that consider the relationship that the mind has with the body to be identical to the relationship that the hardware of a machine, has to the software of the computer system programmed into it. Where the machine itself is seen as the body, and the software of the computer system, is the mind. This resulted in scientists experimenting with equipping technology with software that emulates activities of the mind, giving rise to the development of artificial intelligence. This section will be looking at the different kinds of artificial intelligence and AI related concepts. That is, the differences between what is known as Weak AI, Strong AI, and Artificial General Intelligence (AGI). As mentioned, this section will provide a further discussion on the three types of artificial intelligence previously noted in Part 1, as well as discuss the commonalities which they may hold. Furthermore, I will also look at the philosophical consequences that could arise for the three types of artificial intelligence. From this, I will aim to show that more needs to be done to regulate the actions of these agents and then further explore machine ethics as way of doing so.

Section 1: Varieties of Artificial Intelligence.

Artificial intelligence is a field crowded with various dimensions that obscures the true definition of it. In other words, AI has many different elements and the meaning of it is often plagued with varying understandings. In this section, I will be making use of Arkoudas and Bringsjord's interpretation. Arkoudas and Bringsjord (2014, p. 34) refers to the field of artificial intelligence as, "the field devoted to building artifacts capable of displaying, in controlled, well-understood environments, and over sustained period of time, behaviours that we consider to be intelligent, or more generally behaviours that we take to be at the heart of what it is to have a mind". This is more simply defined as, "systems that display intelligent behaviour by analysing their environment and taking action" (Brussels, 2018, p.3). Margret Boden (2016) views Artificial Intelligence as a platform that "seeks to make computers do

the sorts of things that minds can do (p.2)". Fundamentally, all of these definitions suggest that AI refers to the creation of non-human systems with the capacity to replicate human-like abilities and characteristics while taking the environment into consideration.

It is believed that one of the main goals of AI systems, as held by Boden, is to use computers to get useful things done, to assist with tasks usually performed by humans. Contrast this to Turing's intended vision of how machines interact with humans, which is that they engage with humans through interactions are indistinguishable from them. A way this can be done is through deep learning. Deep learning, for instance, is usually understood to be within an engineering paradigm. The fact that the engineering paradigm is rooted in software and human factors engineering, and deep learning is seen to be part of it, offers a way through which these tasks can be performed. Thus, the ability of computers or machines performing these tasks are not at all idealistic. These tasks include driving your car, collecting, and working through data, growing food, to name a few.

With that in mind, it is necessary to consider the type of tasks that one individual carries out throughout the day. From the moment someone wakes up till that individual goes to bed, there are a number of tasks and activities that the individual performs. These tasks vary from simply remembering to wake up on time, to conversing with friends, family and/or work colleagues as well as accommodating unforeseen scenarios and circumstances that may occur. Taking all these things into consideration, AI was developed. Given the mammoth task of incorporating all these abilities into one system that would be implemented into a machine to conduct these tasks, the field of artificial intelligence ventured into different forms and classifications, as mentioned in Part 1, section 1, important terms and definitions, namely Weak AI, Strong AI, and AGI.

Weak AI is a limited form of artificial intelligence and is known as the most basic degree of AI. It is a type of intelligence that is only able to perform certain specific tasks. According to Khillar (2020), Weak AI does not extend to the full range of human cognitive abilities as it focuses on specific problem-solving tasks, and not to all tasks. These systems are designed on the notion that intelligent behaviour can be imitated and used by machines. It is the use of computer systems that can mimic and simulate certain primary mental processes of the mind related to a specific task and perform it. These systems are then programmed into machines and devices to perform the certain tasks set out. Weak AI systems are programmed with limited pre-defined functions or skillsets to perform a particular task and often have

predetermined responses programmed into it. The simplest examples of Weak AI are programs like Siri, Alexa and/or Google search. These systems in themselves are not intelligent, but instead they behave and operate on the intelligence programmed into them. This largely differs from Artificial General Intelligence (AGI).

Artificial General Intelligence (AGI) is the view that non-human entities such as machines can adapt what it has learnt and apply it to other tasks. It is the idea that these systems can learn and apply the intelligence that they are not only equipped with, but have also accumulated, to solve problems like an individual would in each circumstance. Recall Hal Hodson definition of AGI, noting it as the ability of an intelligent agent to "understand or learn any intellectual task that a human being can" (2019, p.2). It is the idea that these systems are in fact, in themselves, intelligent entities and not solely systems and machines that are operating on the programs that they are designed with. Instead, it relies on its abilities to achieve a degree of knowledge, much like general intelligence, to perform the tasks. In light of the complexities of the goal of AGI, the process of producing a system and machine that portrays these abilities has been strenuous.

This correlates with Weak AI in the sense that, with Weak AI, the responses received from a Weak AI system are already determined based on the program it operates on, which AGI also has access to. However, AGI makes use of both programmed information and its ability to assess a situation. This is where AGI differs from Weak AI. With AGI the responses are determined by the machine itself, it is not only based on the system it is programmed with but also its capacity to assess a certain scenario and react based on what it is faced with in that given moment. Weak AI systems are not programmed with the skill to survey the environment and its ability to react and direct its actions is limited, but not with AGI.

In light of this, one has to consider the moral implications of the actions determined by these systems of intelligence. The moral implication of an action is one that raises the question on which form of intelligence would be able accommodate which kind of moral actions. When one considers different kinds of moral actions, for a weak AI agent the moral action is attached to a specific task like administering medication to a patient. With AGI on the other hand, is associated with more complex tasks like *whether a child should be placed in foster care* and requires understanding the actions from its context, mitigating factors, prior actions, etc. For Weak AI systems, the responses are pre-determined and are already set out even before the situation occurs, whereas with AGI systems the responses may depend on a given scenario.

For example, consider a Weak AI system, like Siri, and suppose we ask it a general question like *"Hey Siri, what is the weather?"*. Based on its program and processes, it will search the temperature on the internet and provide you with an answer. The response is pre-determined based on the process of collecting information off the internet and providing the answer as represented by the weather services. However, for AGI, based on its classifications, if we consider the same question being asked to an AGI system, the answer may be different. Since AGI is primarily characterized as having the intellectual abilities of a human-being, not only will these systems be able to advise on the weather, more than that it could advise on what you should wear given the weather and based on this info it could even call you an uber[5]. This is a key distinction between Weak AI and AGI, the ability of AGI agents to evaluate certain scenarios, assimilate to it as well as modify its response to accommodate the specific scenario. This is closely related to Strong AI and what Strong AI is.

Like AGI, Strong AI is the idea that machines have the ability to achieve human level intelligence. However, with Strong AI it is believed that these machines can really develop a type of consciousness equal to human beings. As mentioned above, it is the view that machines have a mind of their own and can perform task without interference from humans. It is the idea that machines can be developed with not only a mind, but the ability of consciousness that the mind possesses as well. It holds that with the correct computational make-up and algorithm, machines will be able to operate solely based on these algorithmic systems instilled in them, with no involvement by a human-being, "complex algorithms help systems act in different situations and Strong AI-powered machines can make independent decisions without human interaction" (Khillar, 2020, p.4). When it comes to the moral actions that requires Strong AI, one would imagine an even more complex scenario than that of AGI, with more variables and more consequences. For instance, in the case of placing the child in foster care, consider instead of it being foster care (a place with apparent guaranteed safety and security), it is between the two parents – mother and father. This demands a type of consideration and calculation more sophisticated than that of Weak AI and AGI. The agent would need to be able to master independent thinking, compassion and emotion or even being able to imagine what it is like to be one of the parents. Notwithstanding the fact that, Strong AI and AGI share certain similarities, there is a distinct difference between the two types of AI. Whilst AGI proposes that artificially intelligent systems are able to retain data, learn and acquire new functions and adapt its response to a certain scenario, which Strong AI systems

[5] See Melanie Mitchell 2024 on the nature of AGI.

share, they do not suggest that these systems in any way possess the capacity for a mind, which Strong AI systems does. This is the prominent difference between AGI and Strong AI. It is the introduction of the functionality of the mind and the capacity for consciousness that differentiates them. However, despite the fact that Strong AI is vastly different from AGI, based on the element of consciousness, Strong AI needs AGI developed into it, for optimal functionality.

Like AGI, based on the complexities of the concept of Strong AI, developing a system capable of producing the functions set out has not yet been achieved. However, it is possible to envision what this system would look like. Going back to the example mentioned above, for the sake of simplicity, we might want to ask, 'what would a Strong AI systems response be to the question: *'Is this outfit appropriate, considering the weather and the event?'*'. Given the fact that we can assume the response of an AGI system would go beyond just merely relaying the information as received to interpreting it and providing feedback on its interpretation, we can assume that a Strong AI may go even further. Not only can it be perceived that these systems are able to provide you with the temperature as presented by the weather services, and that they may be able to assess the location and alert you of possible climate changes, but it is also possible that, these systems may go as far as potentially 'feel' what it's like to feel the rain or the feelings associated with bad weather and make the appropriate recommendations. It could 'understand' why you might want to know what the weather is in the first place. It may consciously form its own opinion on what is appropriate and what is acceptable since it knows the occasion (funeral or wedding, work function or church function). It could be able to conceptualise the importance of the occasion and the outfit itself and be able see the perspective from the other person's side and consider what the implications are of choosing the wrong outfit. It could be able to feel the hurt and pain one would feel if the wrong choice were made or the joy and happiness if the right choice is made. Considering the characterisations associated with Strong AI, such as its function of having a mind of its own and the capacity for consciousness, if the system knew the weather would change or can recognise a need, such as a coat, it can be assumed that the system would inform you when you ask the question: "*Is this outfit appropriate, considering the weather and the event?*".

The distinction between the three types of intelligences can be understood as Weak AI being the simplest form of AI, using only the systematic responses that it is programmed with. AGI is a little more advanced and uses programming and its ability to evaluate and analyse environments and scenarios, to produce a response. Strong AI goes beyond that and proposes

the possibility for consciousness. It suggests that the responses produced were derived by the machines own consciousness that is equal to that of a human being and not programmed or environmental analyse like Weak AI or AGI.

This gives rise to the consideration of the different types of AI and their moral components. That is, moral Weak AI, moral AGI, and moral Strong AI and what it would look like. The idea of moral weak AI is one that attaches the morality of the action to the systematic responses programmed into the Weak AI system. The action is moral if the systematic response is grounded in morality. A weak AI system is one that follows set responses, so if the system is programmed in morality, the system will follow it, however if it is programmed to adhere to consumerism it may not always produce a moral outcome. For moral AGI, the morality of its actions stems from the systems understanding of the circumstance and actions. These systems rely not only on the programmed functions but also its interpretation of it. As such, when it comes to moral AGI, apart from the programmed functions, the systems understanding, and interpretation should be rooted in moral grounds for it to produce a favourable outcome. This is the same for moral Strong AI. The system's ability to retain and interpret information should be established on the moral grounds in order for the system to react that way. With moral Strong AI, the program embedded in the system plays an important role, not in its response but the ability to distinguish between moral and immoral conditions. Like with moral weak AI, the program plays a role in what it adheres to. Although with moral Strong AI, the system's response is based on its interpretation and understanding of the scenario and the response is solely based on the systems computation of the correct action, if the program embedded in the systems adheres to and promotes something other than morality, it may result in less moral response.

Section 2: Morality and Intelligence

The matter of morality in agents centers around the agent's ability to clearly portray intelligence. It should be able to show that it can perform tasks based on its own knowledge, volition, and skill. From what has been discussed, we see that intelligence can be used to regulate and guide an agents' behaviour, however, it is unclear if the technology can be used to have intelligence include the concept of morality into its processes and practice. That is, whether it can account for an agents' ability to make moral decisions based on its own view and knowledge of right or wrong, and that it does so at its own will. Essentially, morality

needs intelligence, but with AI the intelligence is engineered differently that it falls short in its ability to justify its morality. In what follows I look at an example that seeks to explore which of the agents can use the programs that they are encrypted with as well as their capacities to expand on it to align morality and intelligence.

Consider an alternative approach to the distinction between Weak AI, AGI, and Strong AI. For example, consider that by chance you are in a hostage situation with three machines. One programmed with Weak AI, the other with AGI, and the third with Strong AI. Given the situation, it is not hard to assume that these machines would want to help you in such a challenging time. How they would help you may differ between the three machines, but we may be able to envision what the responses would be. The question therefore stands, how would these machines respond?

For the machine with Weak AI, it will respond only as programmed – it is task specific. The one with AGI will respond as programmed and based on the current circumstances and past experiences. The Strong AI machine will produce a response that is based on the programmed material and systems, as well as the consciousness and understanding of circumstances that it possesses.

For the machine with the Weak AI system, its response in an attempt to help you may take the form of suggesting that you stay calm, listen, and obey the hostage takers demands or it could just say that it cannot help you. We know that Weak AI systems are programmed to respond to certain scenarios in certain ways, if there are no responses programmed for a given scenario, it will not try to create a response to it, it will simply inform you that it does not have a response.

The machine with AGI will go a little further than merely suggesting you stay calm. It might respond by assessing the situation and considering practical solutions to the problem. It may even consider providing you with a plan to escape. Based on the characterization of AGI and its ability to evaluate certain scenarios, assimilate to it as well as change its response to accommodate the specific scenario, it is not hard to imagine the response that these types of machines may offer. It is however necessary to note that, since AGI is simply multiple weak AI systems put together, it will not be able to strategize as logically as needed. As such, although it can provide multiple options as solutions, AGI is unable to strategize based on intelligently assessing the situation – the assessment only scratches the surface and the options provided are based on exposure to such conditions. With AGI, if it was not experienced, the

outputs generated are prompted by case studies and hypotheticals as opposed to informed evidence on the matter.

The machine with Strong AI may provide the same response as the one with AGI, but instead of just providing possible solutions, it may calculate the practicality of the solutions and suggest the possible consequences of the actions if they are done. It may even go as far as considering the repercussions of the actions. Although Strong AI is 'smarter' than AGI, it requires AGI (AGI is like computational common sense). AGI provides Strong AI with the basic level of knowledge needed for practicality and good judgement on matters. Strong AI understands the concept of harm and threat to safety which an AGI system does not. Since Strong AI is described as having the ability of having a mind of its own as well as consciousness, it is not hard to assume that its response may include a consideration of the consequences of the actions on well-being of others.

The problem that AI faces with the intelligence needed for morality is that the reactions performed by the agent is done through programmed responses. It displays no sense of self by the agents. Morality requires a type of intelligence that is driven by autonomy. With the three responses, Strong AI has the most potential to account for the intelligence needed for morality. Its capacity for consciousness makes it possible to identify the difference between the actions of the agent and the programmed responses, and it provides AI with the intelligence needed for morality.

Given the proposed future of AI, it is necessary to consider the impact that its progression may have, as well as the philosophical considerations that need to be addressed when looking at these three types of artificial intelligence. These philosophical considerations include the practicality of engineering the type of AI that is envisioned with AGI and Strong AI as well as the repercussions that might follow if it were created.

Section 3: Philosophical Concerns

AI was built on an idea that centres around the possibility of replicating mental processes and instilling it into machines[6]. The primary feature of AI is that, by replicating human mental processes and programming it into machines, machines are able to mimic these mental

[6] See Rosenblatt 1958 on the storage and organization in the brain. Shows how information is stored and remembered in the brain.

processes by following rules. The rules that the machine follows are what produces this mimicry. Just as humans communicate through the relationship between the mind and its ability to interpret language, machines are able to do so as well. There is the assumption that through replicating mental processes, the relationship between language and mind implicitly follows. This is a fundamental philosophical concern for all three types of AI, the AI needs a way to show that the relationship between language and the mind is still intact when replicating mental processes. All this can be done through GOFAI. GOFAI can provide a common ground on which the replication of mental processes can be developed.

Good Old-Fashioned AI or GOFAI is a type of AI that makes use of programmed instructions through formal symbolic representations as a basis of operations. According to Margret Boden (2014), "GOFAI symbols are items in a formal language (a programming language) like the symbols of mathematics or logic and can be regarded as purely formal (meaningless) structures – and programs composed of them" (p. 89). However, these structures become meaningful when they form part of programs in order to perform certain tasks. These structures earn its meaning through the programme that they are a part of and not as a singular symbol or structure. As such, the functionality of the symbols or symbolic representations depends on the way in which it is programmed to perform. Essentially, according to Boden (2014), it is required that rules guide symbolic data structures for the success of GOFAI computation. GOFAI problems are represented as a set of possibilities (defined by a finite set of generative rules), within which the solution lies – and within which it must be found. Examples include the set of legal moves in chess, or permissible word strings given some particular grammar and vocabulary. It is to say that if something is true, and the thing that is true implies another thing, then the other thing is true as well. For example, if I study hard then I will pass. I will study hard therefore I will pass. In the example, the thing that is true is that fact that studying hard equal good outcomes. Studying hard implies the student will pass, as such, the student passing is true as well. The validity of the statement relies on the initial information, it is dependent on the intention of the program it is set to perform. As such, by using GOFAI, the relationship between language and mind can remain intact when mental processes are replicated, through the use of the correct symbolic data being programmed into machines. GOFAI provides a way in which the intention of duplicating the language and mind relationship can be deliberately programmed and performed by agents.

What GOFAI highlights is that in order to create a functional AI system, there needs to be a shared linguistic foundation upon which the AI is programmed into the machines. This means

that to produce a functioning AI system, there need to be a type of universal 'language' that is compatible with all machines. As GOFAI has shown, symbolic representation can be manipulated by the function it is programmed to perform. So, what is needed for Weak AI, AGI and Strong AI is a program that shares its' meaning and influence on these systems. Even for Weak AI, if there is no way of communicating with these machines, a way to do so other than just coding them with instructions, how would these systems operate? For AGI and Strong AI, it can be argued that these systems may learn these skills on their own, however, they need to have the capacity to recognise a certain type of relationship between mental and linguistic processes. For these systems to have this capacity, they must be programmed into them, which, at this point, appears to be challenging on the grounds that common mental and linguistic processing in the way that is required for AI is yet to be made computational.

Another philosophical consideration to be had is one that involves human intelligence. It is believed that human intelligence involves problem solving and inference. This can be deduced through Arkoudas and Bringsjords (2014) notion on symbolic manipulation. Through exploring the idea that symbol manipulation will be able to account for essential human traits like intuition and judgement which make up problem-solving and inference, it is found that problem solving is essential for human intelligence. That is, in order to excel in human intelligence, the machine must be able to correctly solve problems. Problem-solving involves the capacity to apply logic to reason, as such, in essence, a machine must master both logic and reason. From this we can see that reason is not logic but that it does include logic, reason is partly made up of logic. Machines can only access logic but also needs reason. Reason includes, but is not limited to, the mind using its abilities to think, understand and form judgments, and that it is able to logically justify its conclusion. This is not the only dimension of reason; however, it is the part I will be focusing on. It requires that the entity reaches a conclusion by making assessments of principles of validity based on its own understanding and opinion, and its own emotional and statistical inference. If logic makes up human intelligence, and logic is applied to reasoning (as a machine should reason logically) does it not imply that, for machines to accomplish human level intelligence through logic, they need the capacity to reason as well. Having said that, it is impractical to suggest that a machine can reason. This is based on the fact that although machines can be programmed to make use of explicit logical inference and rules, reason requires more. It requires that the agent make use of emotion, statistical references, and many other cognitive functions.

Unlike mental processes, reason and rational thinking is a creative process and not merely a

deductive process that can be easily replicated, programming the type of system capable of doing this into a machine seems impossible. The problem that the agent faces relate to the subject of creativity and creative reasoning. It involves higher cognitive processes and the need for meticulous characterisation of information. Though it may be believed that existing AI presents a type of creativity as seen in its ability to 'create' new products such as paintings, poetry and/or music, the product produced does not directly stem from its own understanding or logical thinking. It does convey the earliest form of the product from which copies can be made but instead is itself a version of a copy. Although contemporary generative AI does not rely on these kind of rule systems, it still makes use of previously produced data to achieve its outcome. It could be argued that the previously produced data inspires the outcome and does not take it up entirely, like an artist, however, unlike an artist, these systems does not have an original idea of its own which then makes its outcome a new amalgamation of concepts to produce an outcome based on the request that was inputted. Also, the fact that it acts upon request further demonstrates its reliance on external control and lack of autonomy. It is rooted in data retrieved from sources and resources and can be seen a form of sophisticated symbol manipulation. It is grounded on existing data and programmed knowledge, not through engagement with its personalised thoughts or understanding. The resultant new products that it creates are, simply, reconstructed data derived through advanced methods of word association and manipulation of accessible information. Although it is true that, AI has a sense of creativity, it is limited to only what it knows and what it can access, this causes a barrier for the level of creativity needed. AI has a form of symbol manipulation that allows for the rearrangement of existing data to generate a new outcome, but this process is stifled by the system's pre-existing knowledge. This restriction impedes the creativity of AI systems. As such, although AI systems have a sense of creative expression, does not possess the level of creativity that is needed to inform its decision-making.

This concern is specifically related to AGI and Strong AI. For those systems to operate optimally, human intelligence is needed. For human intelligence to be recognized, logic and reason is needed. Intelligence requires that the entity conduct reasoning methods based on emotion, logic, and some form of creativity. And since machines are not able to account for the emotional and creative side of intelligence, we cannot suggest that these systems have human intelligence. This is not a concern for Weak AI since its responses are pre-programmed and determined, there is no logic or reason needed for them to function optimally. As such, for AGI and Strong AI, it is unlikely that the goal of these intelligences will be reached if they

do not consider the need for the capacity to reason.

Another philosophical concern that needs to be considered is the problem presented in Searle's Chinese Room Argument. As mentioned, this argument is portrayed through a scenario where an individual is inside a room and those on the outside are inputting paper with questions in Chinese, a language he does not understand. He then needs to respond to these questions in Chinese, but since he does not understand Chinese, he makes use of a rulebook to understand the questions and provide the answers. His argument is that through consulting the rulebook to answer the questions, it does not infer that he in turn understands Chinese. The same can be said for AI systems. Even if AI systems are built on symbolic logic systems and they are programmed to learn from human competencies and then replicate it, it does not mean that they are then as intelligent as a human. It is argued that simulation does not result in duplication and as such, just because human intelligence can be simulated, it cannot necessarily be replicated. In other words, the imitation of intelligence does not guarantee creation thereof. According to Searle, technology should be able to understand the relevant task at hand and perform said task based on its understanding it. He holds that what is happening in the Chinese Room Argument does not constitute understanding but that it is merely symbol manipulation. However, this concern may be eradicated when one considers the systems approach.

M. Tim Jones (2008) investigates the systems approach to artificial intelligence through an examination of AI algorithms and techniques. This entails that the algorithms are explored in terms of inputs and outputs, within a larger system. As he notes, "no algorithm is useful in isolation, but instead (it is useful) from the perspective of how it interacts with its environment (data sampling, filtering, and reduction) and also how it manipulates or alters its environment" (Jones, 2008, p. 13). This implies that the algorithm relies on the environment, firstly understanding it and also understanding ways of manipulating it. According to Jones, "the systems approach illustrates the practical side of artificial intelligence algorithms and techniques and identifies how to ground the method in the real world" (p.13). When one considers this approach to understanding, the philosophical concern raised in Searles Chinese Room Argument might be eradicated. However, Searle dismisses this possibility.

Searle explored the systems approach and dismissed it based on the fact that internalising all the necessary data does not constitute understanding, there is no meaning associated with knowing all the processes. His argument was that the process that is happening in the room is

merely symbol manipulation, that it is only syntax and no semantics. There is no knowledge or understanding being created in the room and semantics cannot be achieved through syntax.

Despite Searles objections, since it already has syntactic advantages, there may be a way that it can be used to produce the necessary semantics and in turn understanding. The systems approach speaks to how the use of algorithms can produce understanding and in fact, knowledge as well. If the processes are known it creates the syntactic advantages, and then distinctions can be drawn to generate the needed semantics. Once the agents' algorithms are able to identify with the environment, it is able to interact with its environment (data sampling, filtering, and reduction) and also manipulate or alter it, this then creates the necessary understanding for agent. Furthermore, what is also necessary to note on this concern is that Searle argument on understanding has been seen to be flawed by modern AI[7]. The progression of technology in terms of CTM, virtual machines as well the introduction of cognition and connectionism developed into AI processes, programs, and systems has shown that what is happening in Searles Chinese Room is not replication or duplication but instead production.

The Computational Theory of Mind (CTM) involves the prospect of the mind itself as a computational system. It stems from the likelihood of a machine's ability to mirror mental processes such as reasoning, decision-making, problem-solving, perception and more. One of the main components of the CTM is to explain the intentionality of mental processes. It is to show that mental processes have both syntax and semantics. According to Horst, "The computational theory of mind holds that the mind is a computational system that is realized (i.e., physically implemented) by neural activity in the brain. The theory can be elaborated in many ways and varies largely based on how the term computation is understood. Computation is commonly understood in terms of Turing machines which manipulate symbols according to a rule, in combination with the internal state of the machine. The critical aspect of such a computational model is that we can abstract away from physical details of the machine that is implementing the computation" (Horst, 2005, p.25). It is believed that if the correct computation is programmed into a machine, it will be able to perform as an optimally functioning entity. CTM, therefore holds that the mind is literally a computational system.

With CTM the computational abilities of mental processes make it possible to program the necessary capacities for understanding. On the account of virtual machines, if the programmer

[7] See Piantadosi, S. 2023. Modern language models refute Chomsky's approach to language. Lingbuzz. https://lingbuzz.net/lingbuzz/007180

computes the mental processes using connectionism, it would be able to function accurately in any environment. This further accounts for the production of understanding as the neural activations provides the properties that determine meaning and is needed for understanding.

Searle should therefore consider the entire environment and not just the machine. The outputs produced gets do not only get its meaning through the consultation of the rule book but also from causal interactions with the environment – the Chinese speakers outside of the room. As such, what is in the Chinese Room is not a mere machine but one equipped with modern AI that is able to understand requests. The machine then makes use of all its resources from the rulebook to the environment and interactions to fulfil the request, portraying a sense of understanding.

In terms of the different types of AI, the matter of understanding can be seen through intelligence. Intelligence requires logic, emotion, and creativity in the right environment. Weak AI may be able to account for the capacity of logic however, it will not be able to account for the requirement of emotion and logic. With AGI although it could have logic and creativity, it lacks the capacity for the emotional aspect required. The same can be said of Strong AI. It could accommodate the requirement of logic and creativity but there is no strong evidence in favour of it being able to account for requirement of emotion.

For AGI and Strong AI, the philosophical concerns raised by Searle places relative consequences on its development. If we concede that with these systems the tasks, in all its capacities, can in fact be understood, Searles stance on simulation still places a hold on its development. Since simulation of human intelligence does not imply replication of human intelligence, we cannot consider these systems to be intelligent. As mentioned, the imitation of intelligence does not guarantee creation thereof. It may be able to demonstrate intelligent responses within that defined context, but replicating the full spectrum of human intelligence, which includes emotions, consciousness, and adaptability to diverse situations, remains a challenge.

For example, consider that these systems are able to simulate the conditions of a fire, this does not make them a fire. They may be able to raise the temperature in the room and accompany that with crackling sounds and even the necessary light features associated with a fire, but it is not a fire and the same can be said of human intelligence. These systems may be able to imitate certain characteristics associated with human intelligence and even being human, it would not make them human nor intelligent on a human level. The argument is that human

intelligence goes far beyond the intelligence being displayed in one instance, and machines do not have the capacity to replicate it to that extent. To this one may say that children acquire intelligence through the replication of human intelligence however, with children the act of replication is linked to understanding, with AI understanding is relative. When a child sees you reprimand a person or object (scold your brother or hit the chair), it understands that it is because you were hurt, the machine on the other hand will not understand this. Machines do not have the capacity to attach understanding to the process of replication. Although Strong AI and AGI hold the hypothetical capacity to do so, it has not yet been proven to do so. Until Strong AI and AGI can show that it is able to replicate intelligence with the understanding of it, there is no reason to believe that it is in fact intelligent. At this point, the matter of intelligence appears to be one that escapes the grasps of AI systems.

Although the philosophical considerations were addressed, it remains unclear if AI can fulfil the criteria outlined by these concerns. AI's ability to replicate mental processes and include the intricate relationship between language and the mind, is still uncertain, even with the option of GOFAI. Also, the question of whether AI can satisfy the criteria of logic and reason to possess human intelligence remains open, as does the challenge of demonstrating that understanding in AI is not mere imitation but a genuine comprehension by the agents. While it is evident that additional efforts are needed for these agents to attain the level of intelligence associated with humans, there is no denying the progress that technology has made with the development of artificial agents. In light of this advancement and attempts to reach human-level intelligence, the issue of morality comes into question.

On the matter of morality, what needs to be considered is the ethics of these agents. There is a necessity for an ethical framework to regulate machines, given their extensive involvement in our lives. As these machines become increasingly integrated into various aspects of lives it becomes imperative to define and implement ethical guidelines that guide their behaviour and interactions within human environments. Earlier I explored the idea of machine ethics as a way to provide insight to the ethical processes involved in the decision making of machines, here I propose that it be used as the primary ethical system that govern machines.

Section 4: Machine and Data Ethics

When one thinks of technology, it is often in reference to one or more of the many devices and machines that we make use of daily. These machines are built with systems and

mechanisms that come from the advances in engineering, computer science and AI. However, what is often overlooked by most of us, is the actual systems that are programmed into these devices. In fact, it is these systems that makes the technology we make use of so efficient and practical for daily use. When philosophers do think about these systems, they consider these systems to be the ethical value system that is used to regulate and control the functionality of the machines. To better understand, not only the machines or devices we use, but the systems and ethical values that govern them, we need to look at machine ethics.

> *Machine ethics is an emerging field of research that studies the possibility of constructing machines that can mimic, simulate, generate, or instantiate ethical sensitivity, learning, reasoning, argument, or action" (Guarini, p. 213, 2013)*

The mimicking and simulation of ethical sensitivity and actions refers to the systems that are programmed into these machines to control and guide the way these machines react. According to James Moor (2006), the concept of machine ethics is one that varies based on the complexity of ethics in general and the fact that not all values are necessarily ethical. He notes that, "when people speak of technology and values, they're often thinking of ethical values. But not all values are ethical. For example, practical, economic, and aesthetic values don't necessarily draw on ethical considerations" (Moor, 2006, p. 18). The concept of ethics is one that speaks to the moral or normative principles that governs an individual's behaviour and interactions. There is a clear sense of ethical consideration that everyone lives by. Thus, the concept of machine ethics is heavily reliant on the notion of ethics that involves ascribing overarching ethical considerations and moral values to human-made machines. With the increased development in AI, the need for a machine that is equipped with a dimension for moral decision making, is therefore evident. Although AI may struggle to achieve complete morality without the capacity for reason, it remains imperative to equip it with specific capacities aimed at promoting and fostering moral behaviour, even if this involves merely implementing human moral parameters. For this, I propose that machine ethics be used.

It can be said that machine ethics (or machine morality, computational morality, or computational ethics) is a "part of the ethics of artificial intelligence concerned with adding or ensuring moral behaviours to human-made machines that use artificial intelligence, otherwise known as artificial intelligent agents" (Moor, 2006, p. 18). Machine ethics can therefore be used for the implementation of ethical principles into machines and creating a common ground for humanoids and humans. It will facilitate by providing a space for these

entities to interact not only efficiently, but successfully as well. This can be done by integrating the way in which humans interact and communicate with each other, into these computational or humanoid devices which will then result in these humanoid devices conducting itself in an appropriate and correct manner. Machine ethics allows for this possibility as it is built around incorporating the ethics that govern human beings into these machines. It can be used as a basis for looking at the components that governs artificially intelligent agents from its computational abilities to its ethical capabilities. Therefore, the role that machine ethics will execute involves establishing the compatibility of technology with ethics and human interactions.

According to Anderson and Anderson (2007) the role that machine ethics plays in the development of artificially intelligent machines and agents involves creating a sense of security when looking at the behaviour of these machines toward human users as well as other machines. It is concerned with ensuring that the behaviour is not only acceptable on a functional level but an ethical level as well. That is, the way in which it performs its tasks as well as whether it performed the most ethically appropriate action. It emphasizes the need for machines to explicitly represent ethical principles and standards. I will explain what I mean by the term "explicit" below. As such, when considering the development of machines as ethical agents they draw on Moor's distinctions noting that there are two distinct type of agents that can be produced. The first being an *Implicit Ethical Agent* and the second, an *Explicit Ethical Agent.*

An *implicit ethical agent*, according to Anderson and Anderson (2007, p.15), is an agent that is "programmed to avoid unethical behaviour, without an explicit representation of ethical principles". That is, when the agent is created, the agent is programmed to avoid unethical behaviour at all costs. Ethical behaviours itself are not programmed into the machines, instead these machines functions based on a clear set of instructions of what not to do. Asimov's famous laws of robotics[8] is an example of this. James Moor (2006) also states that these machines acts ethically because its internal functions promote ethical behaviour, "ethical behaviour is in the machines nature" (2006, p.19), this will be further explored later in the

[8] "The First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
The Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
The Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law". (Asimov, 1950, p.40).

text. The agent then simply follows the programming imbedded into it and functions as an ethical entity based on this programming, it functions as an ethical entity because it is programmed to avoid unethical behaviour. They are equipped with moral principles and trained to avoid those that are immoral. These agents are not introduced to unethical behaviour, they do not know how to behave unethically, and as such can never react in an unethical manner. The programming of the agent is therefore constrained and limited based on the designer and/or programmer's set of ethical principles. For example, consider a medical robot assistant. That is, a medical robot that is programmed to assist with administering medication to patients. This robot assistant is programmed to facilitate the administering of medication by acting as an instructor to the patients and will do so for as long as it is programmed to do so, it will not diverge from this based on the program that was developed into it. Whether or not the individual actually does what he or she has set out to do at that time (take the medication), the robot assistant will perform the task set out as programmed. The robot assistant can therefore be seen as an implicit ethical agent because it just simply follows the program that it was developed to do. Although, there is nothing directly ethical about this example, the simplicity of its function can capture the linear understanding of the implicit ethical agent concept. Now suppose the robot assistant can detect whether the task scheduled is performed or not, and it reacts by enforcing a type of repercussion or consequence until the tasks are completed. This type of behaviour would be considered uncommon for an implicit ethical agent because it diverges from the program that it was developed with. But, since this was never introduced to it, these agents will not react in this way. On the other hand, explicit ethical agents might.

According to Anderson and Anderson (2007, p.15), an *explicit ethical agent* is able to calculate the best action in ethical dilemmas using ethical principles. These agents are exposed to unethical conditions and situations and react in the most ethical way possible based on its ethical structure. They are programmed with multiple data sets of ethical and unethical behaviours and principles, and they are able to compute and determine the most ethical reaction to circumstances and situations where necessary and react in that way. Take the following example, there are three towns; one has 400 residents, the other has 200 and the last 100. In each town, there is a need for a hospital, but you can only build one. Where do you build? The explicit ethical agent would opt to build in the town with that aligns with its ethical system. For the agent whose systems is rooted in act utilitarianism, they may go for the town with 400 people because it benefits more people. For those rooted in deontology, they will

follow the systematic rules to determine where to build the hospital. Those rooted in African Ubuntu ethics will consider the aspect of community when making the decision. These agents represent ethics explicitly since their reactions are determined by calculating the most ethical response and they operate effectively on the basis of this knowledge. In the example of the medical robot assistant, we know that the implicit ethical agent reacts by performing its task (reminding the patient to take the medication) at its scheduled time whether the scheduled task is performed (the medication is taken) or not. For the explicit ethical agent, suppose it is able to detect the tasks set out to be done, when the robot assistant does its job – knows what it means for the patient to take medication. Based on the fact that the robot assistant is essentially a reminder for the task to be done, these agents might react by providing some consequence till the task is actually performed. It may react in this way based on the fact that as a reminder, it draws attention to the task to be done. If the individual does not perform the task set out, it might presume that its task as the robot assistant is, to a certain extent, incomplete and, based on this deduction, will react to this till it has completed its task. It will continue reminding the individual of the task set out until the individual does it, perhaps by sounding an alarm. With the explicit agent, they follow the principle of well-being for their patients, they are genuinely acting morally as opposed to implicit agents who are merely following rules that are programmed.

Granted, this example is a bit far-fetched, there is some truth in it. There are devices that perform these tasks such as medical equipment, and even smart watches. As such, it is clear that, explicit ethical agents are primarily concerned with the ethical decision making itself, however, as Anderson and Anderson notes, "having all the information and facility in the world will not, by itself, generate ethical behaviour in a machine" (2007, p.15). It is for this reason that we need machine ethics. I will later revisit these concepts from the perspective of James Moor and include a third component – a fully ethical agent.

The introduction of machine ethics into intelligent agents facilitates with the regulation of the behaviour of the machines and devices, by way of equipping these devices with ethical principles through which they can operate. It is believed that the ultimate goal of machine ethics is "to create a machine that itself follows an ideal ethical principle or set of principles, it is guided by these principles to make decisions about the possible courses of action" (Anderson & Anderson, 2007, p.15). That said, because the concept of machine ethics is fundamentally based on instilling machines with ethical principles that they can use to operate. There is a need for careful consideration on the subject of the ethics and ethical

principles that are programmed into these machines.

According to Anderson and Anderson, there are fundamental importance's for the development of machine ethics. The first is related to the ethical consequences of present operations of machines. It is believed that "neglecting the ethical implications and aspects of machine behaviour could have serious ramifications" (2007, p.16). That is to say, that closer attention needs to be paid to the reactions of the machines and we need to verify whether they are grounded in ethical determinations. If we do not equip machines with the capacity to make decisions based on ethical consideration, it might result in these machines acting recklessly or in an unethical manner.

Anderson and Anderson notes that with the rapid growth in Artificial Intelligence and technology in general, it is becoming more and more evident that our notions of the nature of intelligence will be challenged with the creations of 'fully autonomous robotic entities'. They hold that, these machines have the capability of causing harm to humans and that they can only be prevented from doing so by including an ethical component in them. The importance of machine ethics is of such a nature that it provides principles as a way or regulating the actions if the machines. It gives the machines with a set of ethical guidelines that the machine can use to determine its actions itself. It enables the machine to understand how we should act morally. By equipping machines with the necessary ethics, it establishes a foundation on which these machines can operate that is acceptable and appropriate.

According to Anderson and Anderson (2007, p.16), the second reason for the importance of machine ethics relates to introducing artificial agents into society. This is based on the fact that it is impossible to know whether they are able to actually behave ethically. They believe that the future of AI may be at stake since humans fear the possibility of autonomous intelligent machines. This fear stems from the uncertainty of the ethical behaviour of machines. As such, they hold that, "whether society allows AI researchers to develop anything like an autonomous intelligent machine may hinge on whether they are able to build in safeguards against unethical behaviours popular culture as it is rife with images of machines devoid of any ethical code and mistreating their makers, as can be seen by the HAL 9000 computer represented in 2001: The Space Odyssey, as well as the Matrix and much more" (p.16). Machine ethics has the capacity to minimize the risk of this happening as instilling machines with ethics will facilitate machine behaving ethically. Bill Joy (2000) argues that

the only antidote to such fates and worse is to relinquish dangerous technologies. However, technology allows for the enhancement of life and aid those in need. From improvements in efficiency and communication to the advancements in medicine and the economy, technology can be very beneficial. As such, it may not be necessary to relinquish dangerous technologies as a whole but instead minimise the dangerous aspects thereof. It is believed that machine ethics may offer a viable, more realistic solution.

Another aspect of machine ethics that needs to be considered is the influence of the engineers themselves. The moral assumptions of the engineers play a role in the moral aspects of machines. Allen et al. noted that we need to pay attention to the principles unconsciously being built into technology since the use of information technology has placed new demands on values and moral principles as a whole. According to Allen, Wallach, and Smit (2006), "the morality implicit in artificial agents' actions is not simply a question of engineering ethics — that is to say, of getting engineers to recognize their ethical assumptions" (p.13). Engineers are not moral philosophers; they need guidance when it comes to incorporating the moral aspect into their systems. Although engineers are good at building systems based on certain specifications, there is no clear-cut specifications for moral and ethical behaviour.

Machine ethics allows for the implementation and incorporation of the most generalised ethical theories. The role that machines ethics play in the field of computer science is such that it, "extends the field of computer ethics beyond concern for what people do with their computers to questions about what the machines themselves do" (Allen, Wallach & Smit, 2006, p.15). It is believed that with machine ethics, there is a possibility of directly building decision-making capacities into machines The introduction of machine ethics into machines is therefore beneficial as it might allow for the development of more responsible autonomous machines if certain guidelines are followed. However, it is questionable whether we want machines to possess the capacity to act as autonomous machines, but this will be addressed later.

For Anderson and Anderson, research in machine ethics may be able to grow the study of ethical theories as a whole. They note that although it is in the nature of ethics in philosophy to be concerned with the behaviour of agents when they encounter an ethical dilemma, there is little thought being put into the application of the ethical theory to the dilemma. They do not consider how the ethical theory is being directed to the problem – if the ethical theory is

compatible with all ethical dilemmas. Research in machine ethics creates a space for this to be done. This research has the potential to discover problems with current theories and perhaps even leading to the development of better theories as AI researchers force scrutiny of the details involved in actually applying an ethical theory to particular cases. For this reason, "ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas" (p.17). Through this research, there may be a way to construct the type of ethics that can be programmed into machines to direct its behaviour in an ethical manner.

Another aspect to consider is the fact that machines are programmed to operate on ethical principles that are derived from data sets constructed to portray ethical decision making. But where does this data come from? How is the research conducted and what guarantee is there that the research was conducted in an ethical manner and accurately reflects ethical behaviour and principles. In other words, if the research conducted is primarily set up to reflect the desired outcome, if it is conducted in a biased or unethical manner, and not as a more general information seeking manner, the research in itself is tainted and flawed. Therefore, for machine ethics to accurately depict ethical principles, it should be programmed on ethical data. This then draws attention to the ethics of the data and data ethics as a whole. Data ethics refers to a type of ethics that, "studies and evaluates moral problems related to data algorithms" (Floridi & Taddeo, 2016, p.1). It is for this reason that it is necessary not only to look at the ethics of the data captured but also the processes involved when handling this data. Data ethics can therefore serve as a foundation on which the validity of machine ethics can be built.

With the advances in the science of data, data ethics not only studies the already existing data in computer and information studies, but rather uses it to accommodate new technology and ethics as a whole. It emphasizes the intricacies of the ethical challenges of data science and allows for the development of a framework that has the capacity to successfully address these ethical challenges.

Luciano Floridi and Mariasaria Taddeo (2016) look at the data ethics by classifying it in three distinct categories. The first is the ethics of the data itself. It involves the actual data and the processes involved in obtaining this data. They posit that the ethics of data primarily "focuses on ethical problems posed by the collection and analysis of large datasets and on issues ranging from the use of big data … as well as open data" (Floridi & Taddeo, 2016, p.3). That is to say that, given a certain data set, the ethics of data calls attention to the ethical legitimacy

of the particular data set. This legitimacy relates to the key issues concerned with the ethics of the data, which is believed to be, according to Floridi and Taddeo, "possible re-identification of individuals through data –mining, -linking, -merging, and re-using of large data sets, as well as risks for so-called 'group-privacy', when the identification of types of individuals, independent of the de-identification of each of them" (Floridi & Taddeo, 2016, p.3). These issues may lead to serious ethical problems, from group discrimination to group –targeted forms of violence. For this reason, when it comes to the ethics of data, Floridi and Taddeo holds that, it is crucial to create and maintain trust and transparency when working with data. This shows how lack of public awareness is acknowledged especially in relation to the benefits, challenges, risks, and opportunities associated with data science.

The second category in data ethics that is believed to contribute to the make-up of the data sets instilled in machines is, the ethics of algorithms. This relates to the way in which the data that is collected, is processed. The ethics of algorithms, "addresses issues posed by the increasing complexity and autonomy of algorithms broadly understood, especially in the case of machine learning applications" (Floridi & Taddeo, 2016, p.3). Floridi and Taddeo notes that there are crucial challenges that affects the unforeseen and undesired consequences of data, including the moral responsibility and accountability of the designers and data scientists which could further result in missed opportunities. As we know it now (without machine ethics at the foundation of all AI systems), it is impossible to calculate and compute the value and areas of morality, these challenges pose a threat for the development of a machine fully equipped with ethical determination and decision making. However, it is unsurprising that, as they mention, the research on the ethical design and auditing of algorithms requirements and the assessment of potential, undesirable outcomes are increasing (Floridi & Taddeo, 2016, p.3).

The ethics of practices is the third category of data ethics that highlighted by Floridi and Taddeo. The ethics of practices, addresses questions regarding the responsibilities that the people who are in charge of the policies, strategies and data processes have. That is to say that it is about understanding the responsibility and management of the data operations and activity, from those who collect it to those who interpret and those who convey it. The goal of the ethics of practices is therefore to, "define an ethical framework to shape professional codes about responsible innovation, development, and usage, which may ensure ethical practices fostering both the progress of data science and protection of the rights of individuals and groups" (Floridi & Taddeo, 2016, p.3). The three issues at the centre of this analysis are

consent, user privacy and secondary use. The issue of consent relates to the permission of the participant not only to the study but to the sharing of the data captured as well. This in turn, links to issue of privacy when it comes to the data as well as the participant identity. The issue of secondary consent is associated with the distribution of the data collected.

This leads to the relationship between data ethics and machine ethics. Machines function on the data that they have access to. Therefore, for machines to operate ethically the data that they make use of should be ethical. While it is necessary to encode proper ethical principles into these agents and we also need to ensure that the information or data that they have access to are ethical as well. This relationship is of such a nature that the ethicality of the data itself influences the ethics of the machine. Although the machines act of ethicality relates to the actions the machine performs, if the data that the machine uses to determine the action is of an unethical nature, it obscures the determination of the machine's action and can result in an unfavourable outcome. For example, consider an individual who is taught to value honesty as an ethical principle. If, at the same time, they are exposed to the belief that certain situations justify dishonesty, such as deceit being acceptable in business for personal gain, the adherence to the ethical principle of honesty may become compromised. As such, it is necessary, not only to be mindful of the ethical principles that is rooted in the systems of these machines, but the ethicality of the data programmed in them as well.

If machine ethics is used it allows for the development of these categories into machines. Through machine ethics, the ethics of the data can be accounted for by evaluating it through multiple ethical theories and principles. It can also facilitate in the ethics of algorithms and the ethics of practices through the technology that machines make use of and the distribution of the data. It holds the capacity to develop a machine that is able to acknowledge the concerns of data ethics and incorporate the results of the issues into the machines, that is, find the solutions and operate using them. For machines to function optimally based on accurate and correct information, and in the way the world is envisioning to use it, it is necessary to incorporate the data ethics classifications into it. The only way to do this is to develop machine ethics. It is for this reason that machine ethics is important. The importance of machine ethics is one, that is not only evident for the reactions of the machine towards human users, but also to the response that human users will have towards these machines. In essence, although perfect calculations may be impossible, the use of machine ethics allows for approximate calculation.

With the rapid progressions in technology, there is a call for a more comprehensive account for the ethical decision making of machines. The importance of machine ethics is therefore clear, not only based on the ethical ramifications it would have on humanity, but also because of the fear that humans has about interacting with these types of agents. Machine ethics can therefore be used for the implementation of ethical principles into machines and creating a common ground for humanoids and humans. Machine ethics looks at the principles instilled in these devices and creates standards on which to judge them, it evaluates whether these machines are capable of making decisions based on their programming as well as independently of these programs. I believe that it is only through machine ethics, through the development of machine ethics and the incorporation of it into modern day devices, it is possible to create a fully moral artificially intelligent agent. That being said, Moor proposed three grades of AI agents. I will therefore move on the looking at these three grades and considering which of these would be able to accommodate the magnitude of machine ethics.

Section 5: Moor's Three Grades of Artificial Ethical Agents

The technological advances in the field of Artificial Intelligences gave rise to the possibility of artificial intelligent agents in human society. Although these advances have been seen to provide many beneficial aspects to the general lifestyle of society, it is necessary to acknowledge that the growth and development of various forms of AI and AI agents as well, calls for a closer look at what exactly these agents are, from an ethical perspective. Earlier I looked at Anderson and Andersons characterisation of Moor's distinction of ethical agents. Now I will look at Moor exploration of the possibility of implicit ethical agents, explicit ethical agents, and fully ethical agents. This is an approach that looks at how and to what extent ethics can be implemented into machines. It looks at the degrees to which ethics can be programmed into technology and the space that this technology will occupy in our world. In this section, I will be looking at the three AI agents proposed by Moor, namely the implicit, explicit, and fully ethical agents. I will be looking at what exactly these agents are as well as the varying dimensions of these agents. I will also be considering the possibility of developing a fully ethical agent, as proposed by Moor. By doing so, I will conclude by showing that there may be a way to make produce a fully ethical agent if a few concerns are properly addressed.

Artificial ethical agents are AI agents that are believed to be able to operate and function

ethically in a high functioning society. This entails equipping machines with the relevant and necessary ethics to do so. Although there have been many discussions around what these agents would look like and how the ethics can be programmed into machines, Moor explores the degrees through which this can be done, and how it could be done to accommodate the integration of these agents into society. He makes use of three types of AI agents to make the distinction. The first is 'the implicit ethical agent'.

Recall Anderson and Anderson's reference to Moor's distinction of implicit ethical agent, an *implicit ethical agent*, according to James Moor (2006, p.19), is an agent that is "programmed to behave ethically, or at least avoid unethical behaviour, without an explicit representation of ethical principles". As mentioned earlier, when the agent is created, the agent is programmed to avoid unethical behave as it is not introduced to these types of behaviour. These agents are programmed to act ethically through direct transmissions of what ethical behaviour is, such as "do not steal, do not kill, etc.". For these agents, its internal functions are built on ethical behaviour, they do not know unethical behaviour. The agent then simply follows the programming embedded into it and functions as an ethical entity based on this programming. With ethical representations[9] embedded in their software, these machines are designed to adhere to ethical behaviour, not having any knowledge or understanding of unethical conduct, "creating software that implicitly supports ethical behaviour, rather than by writing code containing explicit ethical maxims" (Moor, 2006, p.19). The programming of the agent is therefore constrained and limited based on the designer and/or programmer's set of ethical principles. An example of an implicit ethical agent would be things such as ATM's or digital ticket operating systems. Although the transactions involving money are ethically important, as mentioned by Moor, these machines simply perform their function which involves no form of ethical considerations or calculations. And whilst these actions have ethical implications, they are not themselves ethical concepts, the behaviour is not necessarily ethical, but the effects or consequences thereof are. In essence, an implicit ethical agent operates based on software that implicitly upholds ethical behaviour, the ethical outcomes are directly programmed into the machine and as such, the machine has no need for learning which behaviours are ethical and unethical. Like Searle suggests, the agent or system need no ethical understanding but functions through symbol manipulation. The tasks these agents

[9] In this case ethical representations refer to the consequences of the action. It is programmed with a clear description of the ethical consequences of actions the agent performs.

perform are specific in its duties and requirements, it also leaves little room for learning new tasks and prohibits them from becoming full-fledge ethical agents. Contrast this to explicit ethical agents who has the possibility of developing these capacities.

*Explicit ethical agent*s are the second degree to the incorporation of AI agents into society. For Moor, not only should these agents be able to represent ethical categories and perform the relevant analysis of these categories but operate effectively based on their knowledge of them. These agents should be able to represent ethical categories like calculations on a calculator, that is, by clearly denoting the foundations of its decision and accurately presenting its analysis. According to Moor, they should be able to 'do ethics,' they should be able to carry out their ethical conclusions, like a computer plays chess. That means, based on the scenario, these agents should be able to provide representation of it, make the necessary computations and calculations and perform the relevant actions needed in that moment, much like a computer playing chess would. With this in mind, Moor considers how an agent with such a capacity could be reached and believes that it is only through making use of logic that such an agent can exist. He proposes that through the use of deontic, epistemic and action logic, there is a possibility of producing an ethical agent, or at least the foundation on which to build one. According to Moor, together, these logics indicate the presence of a formal framework that is capable of sufficiently describing ethical situations, thus, enabling machine-based ethical judgments. He notes, "you could use a combination of these logics to state explicitly what action is allowed and what is forbidden in transferring personal information to protect privacy" (Moor, 2006, p.20). That is, if technology is programmed with these three forms of logic, they will be able to produce a machine that is an explicit ethical agent. He believes that with the permission and obligation statements of deontic logic, the statements of beliefs and knowledge of epistemic logic and with action logic informing the actions of the agent, it will provide a basis through which machines can sufficiently make ethical judgements. This will create an awareness of what is allowed and what is forbidden on which the machines can operate, "you'd program a computer to let some personnel access some information and to calculate which actions what person should take and who should be informed about those actions" (Moor, 2006, p.20). In addition to this, these machines would need an ethical theory on which to operate. To this, he introduces Anderson, Anderson, and Armen's (2004) considerations, suggesting the potential application of Jeremy Bentham's Hedonistic Act Utilitarianism approach. This approach weighs the amount of pleasure produced to

displeasure. Based on this theory the appropriate action would be the one that produces the most pleasure and the least displeasure. Anderson, et al. also provide a revised version of the W.D model[10], that prioritizes prima facie duties over absolute duties, known as their computer model. Since no duties are in fact absolute, the implementation of the W.D model is compromised, and to this, Anderson, et al. suggest introducing a learning algorithm to adjust the judgements of these agents. A type of algorithm that combines the types of duties and their levels with past intuitions. By doing so, the appropriate decision can then be reached. Given the complexity of the task of distinguishing what an explicit ethical agent would be, Moor reaches a point that views these agents as capable of making decisions based on their own form of reasoning and considerations. He asks "What would such an agent be like? Presumably, it would be able to make plausible ethical judgments and justify them. An explicit ethical agent that was autonomous in that it could handle real-life situations involving an unpredictable sequence of events would be most impressive" (Moor, 2006, p.20). Unlike implicit ethical agents who operate on pre-determined actions and programmed responses that are not ethical in nature themselves, explicit ethical agents could in principle hold the capacity to determine their own actions and behaviour by considering ethical principles explicitly. The next level adds a further component bringing it closer to the human ideal, namely Fully ethical agents.

According to Moor, "a *full ethical agent* can make explicit ethical judgments and generally is competent to reasonably justify them" (2006, p.20). These agents are often compared to the average adult human because it is perceived that these agents have the capacity to possess the same characteristics as humans such as consciousness, intentionality, and free will. Ideally a fully ethical agent would be able to calculate the best action in ethical dilemmas using ethical principles. These agents are exposed to unethical conditions and situations and react in the most ethical way possible based on their ethical structure. They are programmed with multiple data sets of ethical and unethical behaviours and principles, and they can compute and determine the most ethical reaction to circumstances and situations where necessary and react in that way. These agents represent ethics explicitly since their reactions are determined by calculating the most ethical response and they operate effectively based on this knowledge, as well as their capacity of consciousness, intentionality, and free will. They will perform the most ethical action in a scenario because they are aware of the situation, have the intention of

[10] W.D Model: (named for William D. Ross). Ross's theory emphasizes prima facie duties as opposed to absolute duties. Ross considers no duty as absolute and gives no clear ranking of his various prima facie duties.

doing so and have the capacity to do so based on their self-determination. As such, a fully ethical agent is often seen as an average adult human. However, whether these machines have these characteristics needs to be considered, that is, whether machines can be full ethical agents. This is a technological problem as well as philosophical one. If the technology brings machines closer to us in function and complexity, the line between human and artificial agents might become blurred. Moor himself makes use of the bright line argument to portray the complexity of these agents, as he states, "the ontological difference between humans and whatever machines might be in the future" (Moor, 2006, p. 20). This means that even if machines become highly intelligent and sophisticated, it is unclear whether they can truly possess consciousness, self-awareness, or other human-like qualities. With this the ontological difference then reflects the idea that there may be inherent and profound distinctions in the essence of human existence compared to any artificial entities that may be created. With the introduction of the bright line argument there is a prospect for a focused exploration of the distinct boundaries that may exist between certain concepts or entities.

The first side in this argument is to hold that only fully ethical agents are ethical agents. That is to say that only machines that are full ethical agents, those with the capacity to make explicit ethical judgements and can reasonably justify them as well as be in the position to be compared with the average adult human, are ethical agents. This stance involves regarding the other senses of machine ethics as ethics that do not involve agents, it still considers the implicit ethical agents and explicit ethical agents as machines or devices instead of agents. Since the implicit and explicit agents do not have the capacity for consciousness, intentionality, or free will, on this account they are not ethical agents. Although Moor holds that these agents provide valuable insight to the aspects of machines as a whole, "to ignore the ethical component of … implicit ethical agents, and explicit ethical agents is to ignore an important aspect of machines" (Moor, 2006, p. 20), he further holds that, "the fact that lesser ethical agents lack humans' consciousness, intentionality, and free will is a basis for arguing that they should not have broad ethical responsibility" (2006, p.20). Thus, he links morality to consciousness and free will. But arguably non-human animals could exhibit moral behaviour without some of these properties (they do seem to have free will if we do). Furthermore, each one of these properties are highly contentious in philosophy. Free will is questionable in a deterministic universe, consciousness is hard if not impossible to explain (Chalmers 1995) and intentionality is shown in other animals. It is therefore either unclear why machines are excluded, or it is simply unfair to exclude them.

The second side to the bright line argument is to hold that no machine can be a full ethical agent. That is to say that no machine can have consciousness, intentionality, and free will. However, this may never be known. As Moor states, even Searle who strongly opposes the development of Strong AI, leaves room for this possibility, "he only denies that computers, in their capacity as purely syntactic devices, can possess understanding. He does not claim that machines cannot have understanding, presumably including an understanding of ethics. Indeed, for Searle, a materialist, humans are a machine, just not a purely syntactic computer" (Moor, 2006, p.21). What is meant by this is that the only matter that is problematic and that is being objected is the computers' ability to possess the ability to understand if they are just syntactic devices. It is not the understanding, but the fact that the computers are viewed as purely syntactic devices. This means that these devices can possess understanding if they go beyond the syntactic capabilities and attain semantic properties. In essence, the issue is not the potential for understanding but the restrictive characterization of computers as purely syntactic devices. The dispute addresses whether machines, if they surpass syntactic limitations and acquire semantic properties, could indeed possess understanding, including ethical comprehension. This suggests that the prospect of developing machines as ethical agents is still open to consideration, and that it depends on overcoming the current limitation of perceiving them merely as syntactic devices. As such the possibility of developing a full ethical agent is one that is yet to be disregarded.

The key difference between these agents, explicit agents, and fully moral agents specifically, is in their characteristics. A fully moral agents is believed to have characteristics such as consciousness, intentionality, and free will, this allows them to autonomously form conclusions. In contrast, the explicit moral agents depend on programmed responses in the sense that the decision-making process is pre-programmed and not necessarily the responses themselves. Therefore, the conclusions reached by fully moral agents are inherently their own, while explicit moral agents gain their conclusions through programmed computational processes.

Given these distinctions, a question that needs to be considered is whether these agents can be developed. On the account of the implicit ethical agent, it seems clear that not only can these agents be developed, but that they already exist. Since these agents are characterized as being able to perform the most basic task, and do so ethically, we can see how machines such

as ATMs, as mentioned by Moor, are implicit ethical agents. The tasks that these machines are performing were previously done by a human being because it was considered an overly sensitive and intricate task. This is based on the fact that the transaction requires that the individual understands the request for money, and not only understand the task but also be able to carry out the task by knowing the value of the money. For example, Person A goes to a bank teller (Person B) requesting R500. Although this transaction seems simple, there are quite a few components involved. Firstly, Person B should be able to recognize that Person A is requesting money and not something else, like requesting bank statements. Secondly, Person B should be able to understand the value of money and the money being requested. That is to say, Person B should know what the value of R50, R100 and R200 notes is as well as the request for a R500. Furthermore, it should also be able to calculate which note corresponds with which value and which of these will make up Person A's request for R500. As such, it is clear to see the ethical component of such a transaction, since it involves that a certain amount of knowledge being required to perform the action correctly and successfully. There are ethical considerations that occur based on the responsibility associated with possessing the necessary knowledge. Ethical behaviour is closely linked to competence, expertise, transparency, and responsible conduct as such, there is a clear ethical obligation to utilize knowledge for the benefit of all parties involved. This is done to foster trust, fairness, and the overall well-being of the individuals participating in the transaction. The calculation that the made is based on the preexisting knowledge that the agent has regarding the value of the money and the request itself. Based on this, we can concede that ATMs are able to perform such functions. These machines can recognize the request of the individual, they understand the value of money being requested as well as hold the capacity to formulate the correct calculation to attend to the request. This is a clear representation of implicit ethical agents as these machines are programmed to perform specific tasks, their avoid unethical behaviour by giving the incorrect amounts because their software implicitly supports ethical behaviours. Programming this structure into an ATM shows that implicit agents not only can be developed but do already exist.

When it comes to the explicit ethical agent, we must consider the distinct characterizations of these agents. For Moor, these agents should not only possess the capability to represent ethical categories and conduct the necessary analysis but also demonstrate effectiveness in representing, performing, and operating based on their understanding of these categories, "it would be able to make plausible ethical judgments and justify them. An explicit ethical agent

that was autonomous in that it could handle real-life situations involving an unpredictable sequence of events would be most impressive" (Moor, 2006, p.20). We are therefore left to consider, if these agents are developed, what these agents would look like. These considerations may not be all that far-fetched if we think about the self-driving car. Self-driving cars may, to some extent, be considered explicit ethical agents. If we work on the analysis of explicit ethical agents as being autonomous and that can handle real-life situations, we see that self-driving cars have the capacity to do so. These vehicles perform the intricate task of driving and make the necessary considerations as humans do when driving. They are, to some degree, relatively autonomous in having the freedom to control its own actions and all these actions are revealed through real-life encounters. The actions of these vehicles are based on both programmed information and learnt experiences. If a self-driving car is designed to prioritize the safety of the driver but encounters a scenario where the car faces the dilemma of potentially harming five pedestrians to protect the driver, the vehicle is placed in a difficult position. However, suppose it has encountered a similar situation before, the decision-making process then involves drawing from both sets of information (programmed and experienced). It is only after the car has made its decision and carried out the action that the manufacturers can examine the underlying code to understand the underlying principle behind the chosen course of action. For the explicit moral agents, if it was made mainly on the programmed information or from learnt experiences. They are able to make ethical judgements, however, we are left to question whether these judgements are plausible and whether these agents can justify them. This is a question that may never be answered and for that reason we might need to acknowledge that although there is a possibility to develop explicit ethical agents, until we are able to prove that these agents are aware of the ethical judgments they make and can justify them, its development is stagnant. As such, the possibility for the development of full ethical agents is also obstructed.

On the possibility of creating fully ethical agents, we see that this too is obstructed. Full ethical agents are defined by the capacity for consciousness, intentionality, and free will. However, what technology has failed to accomplish is to provide a way that this can be done. For the requirement of consciousness, since we are unable to determine what exactly consciousness is, there is no universally accepted meaning that describes consciousness. If we cannot distinctly define it, we cannot program it into machines and thus cannot produce a machine with the capacity for consciousness. This then places strain on the basis of morality. David

Gunkel explores the condition of consciousness only to discover that consciousness is in fact, an unclear and virtually undefined term. According to Gunkel, “the problem, then, is that consciousness, although crucial for deciding who is and who is not a moral agent, is itself a term that is ultimately undecided and considerably equivocal” (2014, p.116). In what follows, I will expand on how Gunkel explores the condition of consciousness as the foundation of moral agency.

The issue with intentionality is that it is difficult to ensure that a machine is acting based on its own intentions and not as simulated or based on underlying programmed information. Machines operate on algorithms and programmed instructions, and they perform tasks based on predetermined logic and data inputs. Even though there has been progress in machine learning and AI, the question still remains whether a machine's actions truly reflect an internal awareness and intentionality or if they are simply responding based on generated programmed parameters. The problem is in detecting if a machine's apparent intentionality is a result of true cognitive processes or merely a sophisticated mimicry of intention-driven behaviour. The ongoing challenge is to develop methods and frameworks that can validate and ensure the authenticity of intentionality in machines.

Although the development of technology has proven that its advancements produce the possibility of ethical agents to be introduced into society, it is yet to provide an indisputable way for this to be done. However, with the introduction of ethical principles to the production of AI agents, there may be many ways for these introductions to take place. For this reason, when considering whether these agents can be developed, it becomes clear that this possibility is one that should not be forsaken.

In essence, despite the fact that it seems like Moor has succeeded in demonstrating the possibility for ethical agents, it is not without its flaws. With the implicit ethical agents, although the possibility and reality of these agents are greater than those of explicit and full ethical agents, there is something to be said about its foundations. Upon critical analysis, it is seen that although these agents perform actions with ethical implications, these agents are themselves not ethical concepts. This means that the agent is not ethical but the consequences of the actions they perform are ethical. As shown in the ATM example, the ATM producing the correct amount of money is an ethical action, but the ATM itself is not an ethical concept or an ethical agent. The explicit ethical agents too, faces critiques based on Moor’s notion that these agents function ethically through accurately representing scenarios and calculating the

relevant actions needed in the moment. Being able to function ethically does not guarantee the ethicality of the agent. In other words, having the ability to perform ethically or adhere to ethical principles does not automatically ensure that the agent is inherently ethical, capability to act ethically does not guarantee the overall ethical nature or integrity of the agent itself. This is because the ethicality of an action could be separated from the agent. The action of an agent is based on the scenario and not the agent itself, as such the ethicality of the agent should not be determined by its actions but rather the agent itself. Furthermore, relating to Moor's characterization of explicit ethical agents, the fact that machines would be considered explicit ethical agents through the use of logic, gives rise to questions regarding humanity and its relation to these agents. For the full ethical agent, these critiques grow based on the fact that full ethical agents are defined by the capacity for consciousness, intentionality, and free will. It is believed that these agents will perform the most ethical action based on these capacities, however, it is unclear how these agents come to possess these capacities. For this reason, in what follows, I will be presenting critiques for these agents. I will be making use of David Gunkel's (2014) paper on the vindication of the rights of machines to show how Moor falls short of presenting a sound argument for the development of ethical agents.

Section 6: The Deconstruction of Moor's Grades of Artificial Ethical Agents

The production of artificial ethical agents stems from the possibility of equipping machines with the relevant mechanics and ethics to operate and function ethically in a high functioning society. Above we have seen how Moor explores what these agents would look like and how the ethics can be programmed into machines, he explores the degrees through which this can be done, and how it could be done to accommodate the integration of these agents into society. Now I will look at Moor's characterisation of implicit, explicit, and full ethical agents in relation to the attainment of moral agency. With the implicit ethical agents, I will look at the way in which Moor considers moral agency can be obtained by these agents, noting the distinction between the agent and its actions. For the explicit ethical agent, I will look at the role that logic plays in morality of these agents drawing on the impact that it will have on the relationship between humans and machines. With the full ethical agent, I will look at two components that the concept centers on. I will look at these components and look at the importance it has for the development of full ethical agents. I will also introduce the need for consciousness and making use of Gunkel to expand on what constitutes consciousness.

When we look at Moor's characterization of implicit ethical agents, we see that he classifies these agents based on their actions. Their actions are ethical because these agents are programmed to avoid unethical behaviour without explicit ethical systems. They act ethically as it is the only actions that are programmed into them, by way of avoiding unethical behaviour. However, these agents are not themselves ethically competent since, they do not engage with ethical concepts. The agent acts ethically because of its programming, so their adherence is solely based on a set of rules or guidelines provided by their creators. What is important to note is that ethical competence implies the ability to know, examine, and make decisions based on a deep understanding of ethical principles and apply it to ethical dilemmas that may arise. This means that, although their actions have ethical implications, since the agent themselves does not engage in ethical consideration, they do not determine their own actions through deliberation of the ethicality of an action, they cannot be characterized as ethical agents.

Much like moral standing and moral agency, the ethicality of an action relies on who or what is performing the action. According to Gunkel, "moral agency has been customarily restricted to those entities who call themselves and each other (hu)man" (Gunkel, 2014, p.115). This raise concerns as what is being proposed as conditions of moral agency are capacities or characteristics that are hard to define and even harder to achieve. Moral agency relies on the notions of consciousness, intentionality, and free will. On the condition of consciousness alone, we see problems arise. Having consciousness as a condition for morality leads bigger concerns such as the discrimination against other moral beings. The problem is that our understanding of consciousness and free will is still so limited, making them the criteria for defining moral agents would be faulty. The lack of clarity on how consciousness is achieved complicates the task of determining moral agency in other things. As such, if we are unable to define it, how will it be possible to attribute its classifications to others. If we do not know how one goes about achieving it, how will it be possible to distinguish who possesses or not.

He also recognizes that there has been a shift in the attribution of the characterization of moral agency from the individual to its actions, he acknowledges it but still holds that despite the shift, it is still primarily agent oriented. Likewise, the same applies to ethical agents, it is a concept attributed to agents that engage in ethical consideration and not just based on the action itself, although the action itself may be ethical, we cannot regard the agent as an ethical agent. It is the process that the agent follows that allows the agent to reach the ethical conclusion that makes it an ethical agent, but since implicit ethical agents follows no

deliberative ethical process, they cannot be considered ethical agents. These agents are therefore not ethical agents at all, but mere machines performing their functions, like a printer that produces a printed page.

With the explicit ethical agent, we see Moor face the same critique. Explicit ethical agents are defined by the way in which they determine the most appropriate and correct action. These agents are often characterized as being similar to the operations of a computer playing chess, however complications arise when we look at how the determinations of the actions of the agents are done. These agents are believed to produce ethical conclusions through representing a scenario and making the appropriate calculations to reach the most accurate conclusions, but we are yet to uncover the methodology they make use of to reach these conclusions. With these agents, although we know the information that they are programmed with, we may not know the value that they place on certain connections in certain scenarios, as such the outcomes produced can questionable. We do not know significance that they place on one action or connection as opposed to another. Like with the self-driving car example, if these agents are face with the decision of protecting the driver and harming five individuals, it is only once the action is performed that the process can be assessed and modified. And since it is the connections that are made that determine the action, we have no way to know, in that moment, or even before, the way they will react in every scenario this is based on the lack of information on the significance of connections. Also, why they were made, are unknown to us. Although Moor advocates for the use of deontic, epistemic and logic, there is still discrepancies regarding what they believe is the correct reaction. Each logic holds its own views and conditions and the outcomes that are produced through these logics often differ, as such, for the explicit ethical agent this becomes problematic according to Moor's characterization of these agents and we can see this when we look at the notion that these agents are like a computer playing chess.

In the first instance, chess is a complex game with clear-cut rules and objectives, making decisions on ethical actions on the other hand, differs from it. That is, if the agent does not operate on deontic logic. If it does, then it takes a form similar to that of chess. Although the complexity of making ethical decisions bares some similarity to that of playing chess, it is in no way a game and there are no clear-cut rules, and the objectives are often shifted and reconsidered. With chess, the rules are clear and universally accepted and all those who knows them agrees with them, however, when it comes to the rules of making ethical decisions, it is uncommon that all would agree. With ethicality, although there are rules that are followed by

all, there could still be a few people that may stray from it. While certain ethical rules may be universally acknowledged and accepted, the application of these principles in real-world situations can give rise to disagreements, which further emphasises the subjective and context-dependent nature of ethical decision-making. For example, in chess, the fact that the bishop only moves diagonally cannot be disputed, but it can be disputed whether a prisoner on death row can donate his organs to save lives.

Furthermore, Moor introduced incorporating logic into these agents to produce machines that are explicit ethical agents. He believed that through deontic logic that informs the permission and obligation dimensions of the agent, epistemic logic for beliefs and knowledge and action logic to inform the actions of the agent, machines as explicit ethical agents can be produced. For Moor, these logics propose that ethical situations can be describe through sufficient ethical judgements made by machines. However, this results in questions regarding the ethical reasoning of this machines. Logic has its limits. Not all reasoning is based on logic and not all of ethics can be captured by it. What this means is that although logic plays a role in reasoning, not all forms of reasoning rely strictly on logical principles, and similarly, not all ethical considerations can be fully understood by logical reasoning alone. It leaves room for the possibility of inadequate representations of ethical situations as we do not know what informs the logic and reasoning of these machines[11].

With this in mind, I now move on to the role that logic plays in morality of these agents drawing on the impact that it will have on the relationship between humans and machines. The introduction of logic to morality gives rise to questions of humanity and the relationship between humans and these machines. It calls into question the way we as humans make our ethical decisions, if it is through this logic, does it imply that these machines have human capacities or that these agents are of a higher standard than we are, and what are the implications it has on the characterization of moral agency. It forces critical examination of the human approach to ethical decision-making. If human ethical decision-making is driven by logic, all moral decision-making need to be evaluated.

Furthermore, as humans we are also exposed to these critiques. The possibility of a human-being always making the right ethical choice (always being the operative word), is one that disputed on every account. The difference, however, is that with humans, these choices are determined by the individual, the consequences are experienced, in all circumstances, to a

[11] See Robert Kowalski (1989) for more on the limitations of logic.

certain extent, by the individual and the individual can be held accountable for the choice that was made. With these agents, these aspects shift from the agent to its programming, the capacity it finds itself in and the engineer or owner. As such, the idea that explicit ethical agents can be produced, the idea that an agent capable of handling real-life situations involving an unpredictable sequence of events can be developed and introduced into the high functioning society is one that is only possible once these critiques are addressed and dealt with.

On the account of the full ethical agent, it is believed that these agents will be able to make ethical judgements and be able to justify them. These agents are said to have the capacity of consciousness, intentionality, and free will; however, it is unclear how this is to be done. This can be divided into two components, one epistemic and the other metaphysical.

The epistemic component relates to the justification of actions. That is, the making of the ethical judgements and the way in which the agent justifies its actions. This component is one that is needed as it will attest to the agent's knowledge on ethics and its intentionality regarding the actions its choses to perform. It will also speak to the agent's capacity of free will. As such, if an agent is programmed or learns ways to make its actions known after the fact, as well as show just cause for the action, it will be able to meet the epistemic requirements. There is no reason that machines cannot get to the point where they are able to justify their ethical actions in accordance with ethical theories, especially if they are programmed to do so. However, what is necessary to note is that whether the justification is programmed or learnt, it may lack the legitimacy of true ethical understanding. This could lead to a superficial alignment of these agents with ethical theories without genuine moral understanding. This then leads to the second component, the metaphysical component.

This component involves the agent having certain kinds of properties. In this instance, it is the properties of consciousness, intentionality, and free will. The epistemic component aids in the portrayal of the agent's capacity of intentionality and free will through its justification of actions. However, the condition of consciousness is one that stirs trouble not only for agents but for morality. In the first instance, for a machine to possess the capacity for consciousness, we should be able to determine exactly what it is, what consciousness is, but we know little of what consciousness is. The absence of a clear understanding of how consciousness is achieved and the lack of criteria for distinguishing its presence, raises concerns for the basis on which morality is built.

In an attempt to uncover the foundations of moral agency, Gunkel explores the condition of consciousness only to discover that consciousness is in fact, an unclear and virtually undefined term. As mentioned earlier, Gunkel holds that the problem with consciousness is that the term itself is one that is significantly vague and notably ambiguous even though it is a deciding factor for who is and who is not a moral agent. To further his stance on the complexity of the definition of consciousness, he uses Max Velmans (2000, p. 5) characterization of the term that expresses that there is no universally accepted meaning as it means different things to different people in many different societies. A question then arises if we are unable to define what consciousness is, if this concept is so vastly different in different fields and vaguely defined when tried to distinguish it, how is it possible to equip machines with it, and how will we know if they have it. This is then precisely the problem with consciousness. We have no idea what it is or even what it looks like in ourselves, and if we are unable to quantify it, if we cannot distinguish it from other characteristics of ourselves, we cannot subject others to this requirement. As Gunkel states, "because consciousness is a property attributed to "other minds," its presence or lack thereof requires access to something that is and remains fundamentally inaccessible" (2014, p.117). This means that in order to replicate consciousness, a comprehensive understanding of the concept is needed and without this, without a framework or definition of consciousness, it cannot be recreated. Until we are able to identify, and components and mechanism needed, this process will always be hindered. For example, suppose a group of individuals are attempting to create a paint colour matching the shade of '*Jade'*. However, there's disagreement among them regarding what precisely constitutes the colour jade. Some argue it leans more towards green, while others contend it has blue undertones. Without a consensus on the exact hue and properties of jade, the group struggles to mix the paint accurately. As a result, their attempts to replicate the colour may yield varied results, with each interpretation falling short of capturing the true essence of jade. In this scenario, the lack of a clear definition or understanding of jade colour obstructs their efforts to produce an accurate representation, similar to the challenge of replicating consciousness without a definitive understanding of its nature. And so, if we are never able to define and determine exactly what consciousness is, we will never be able to produce a machine with the capacity for consciousness.

From the numerous attempts to demonstrate the possibility of developing AI agents, it is clear that, these agents are in fact the future. However, the introduction of these agents into society is a concern that needs to be addressed. This can be done by creating a platform for these

agents to establish meaningful and sustainable relations through the use of their 'cognitive' abilities to enhance the ethicality of these agents. In what follows, I will look at the ways in which these agents make use of their abilities to perform ethical actions.

Section 7: The Integration of Ethical Theories

I have looked at the possibility of ethical AI agents by looking at the different types of AI's and how machine ethics can be used to direct the decision-making processes of these systems. By exploring Moor's classification of three AI ethical agents, although good progress has been made for the introduction of AI agents, it can be seen that more is needed for the ethicality of these agents.

As shown, the idea that fully moral artificial intelligent agents can be developed is one that is definitely possible, but more is needed before this can be done successfully. What is needed is a suitable and tangible ethical theory that is able to be incorporated into these devices for them to perform the most ethically correct action and to be considered moral agents. More specifically, what is needed is not an ethical theory as a whole but instead ethical principles of an ethical theory or multiple theories to produce the desired agents. As such, the most effective way to ensure the ethical conduct of these agents is by designing them with ethical principles from the outset. This entails integrating ethical considerations into every stage of their development, from initial conception to its ongoing operation. By equipping these agents with an ethical framework, we can cultivate a culture of responsible AI that prioritizes ethical decision-making. This approach fosters trust in AI systems and reduces the potential risks and consequences associated with their actions, thereby promoting their beneficial integration into various domains of human life. I will look at the specific kind of ethical principles that these agents could be implemented with, to make this possibility a reality. Since we know that machine ethics can be used to implement ethical principles into machines, we need to determine exactly which ethical principles that should be. I will be discussing the possibility of the development of fully moral artificial agents by using ethical principles of three ethical theories. Specifically, I will be discussing whether Utilitarianism, Deontology, and African Ubuntu Ethics will be able to develop a framework for machine ethics that guarantees the ethical behaviour of AI agents while incorporating societal considerations and is sustainable over time. This will be done by analysing each theory and comparing its compatibility with artificial moral agency. I will look at a principle for each theory to discover which theory is

the most compatible for the development of moral AI agents. By doing this, I will show the most compatible theory for moral AI agents, is in fact all three of the theories. If machines were to incorporate elements of all three the theories to navigate and regulate their actions, it could potentially equip them with a comprehensive set of strengths necessary for moral behaviour.

PART 3: Ethical Reasoning and AI

The development of fully moral AI agents can be explored by programming AI agents with ethical theories for them be function as moral agents. In order to determine this, what is needed is a clear understanding of what a moral agent is. As previously mentioned, I will be relying on Angus Taylor views moral agency that defines it as, "the ability to make moral choices based on some notion of right and wrong and to be held accountable for these actions" (Taylor, 2003, p. 20). This implies that a moral agent is an individual that is able to make reasonable moral choices that by differentiating right from wrong. This differentiation is made possible by cognitive processes such as perception, attention, memory, language processing, problem solving, decision making, learning ability, creativity, reasoning, emotional regulation, etc. As Markus Christen, et al. notes, "A moral agent is any person or collective entity with the capacity to exercise moral agency through rational thought and deliberation, these are prerequisite skills for any agent" (2014, p. 117). This coincides with Moors definition of full ethical agents as stated above. In this section, with these definitions in mind, I will expand on how ethical theories can be used to develop ethical AI agents. The ethical theories I will be looking at is Utilitarianism, Deontology and African Ubuntu ethics. I propose incorporating principles of all three theories into AI agents in order to develop the ethicality of these agents. I will first look at these theories individually and in the analysis, I will show how the use of its principles can produce the desired outcome.

Section 1: Utilitarianism

Utilitarianism is the ethical theory that derives the ethicality of an action from the consequence that follows the action. The idea behind utilitarianism is that the correct action improves the well-being of the people. It holds that the morally right action is the one that produces more good than bad, i.e., the action that produces the most good. It is also often referred to as the

theory that equates good with pleasure and bad with pain, as such, a version of this theory that is most common is that it is the theory that operates on the pleasure and pain principle. The morally right action is the one that produces more pleasure and less pain. This type of utilitarianism is known as Hedonistic Utilitarianism. Apart from this, there is also Preference satisfaction utilitarianism. This type of utilitarianism highlights the wishes (preferences) of the individual. Additionally, it is also necessary to note, that there are also two ways through which utilitarianism functions, that is, Act utilitarianism and Rule utilitarianism. Act utilitarianism focuses on the action's impact when judging if that action was good, where rule utilitarianism, on the other hand, is interested in the moral rule that the action follows, "Rule utilitarians prioritize good general rules over the particular consequences of a given action, whereas act utilitarians focus on these particular consequences and not on rules" (Abumere, 2019, p. 48). For the sake of argument, I will interpret the implementation of utilitarianism as Hedonistic and which functions on rule utilitarianism. This is done to achieve a comprehensive understanding of ethical dilemmas and how balancing considerations of individual actions with adherence to overarching moral principles can offer the greatest overall benefit to society.

Jeremy Bentham and John Stuart Mill were two of the early advocates of Utilitarianism. They believed that in order to perform the most morally right action at all times, we should do the thing that maximizes good, "that is, bring about 'the greatest amount of good for the greatest number of people" (Driver, 2022, p.1). This is a critical tool in the theory as it distinguishes what makes an action a morally good action or the morally wrong one. It is however necessary to note that the theory operates on impartiality such that the 'goodness' or 'rightness' of an action does not depend on the individual or any particular individual. Instead, when considering which action to perform all factors are considered. As Driver notes, the theory is agent neutral and everyone's happiness hold the same weight, "when one maximizes the good, it is the good impartially considered. My good counts for no more than anyone else's good. Further, the reason I have to promote the overall good is the same reason anyone else has to promote the good" (Driver, 2022, p.5). Therefore, the morally correct action is the one that generates the most pleasure for all and the least to no pain for anyone.

In essence, the theory functions on three principles. The first is the action itself. This is the belief that all actions are either good or bad and in order to perform the most ethically

appropriate action, one has to determine which action is good and has to perform said good action. In order to determine which action is good, one needs to consider societal welfare, this is the second principle. It involves critical analysis of the action based on the total welfare it has on all individuals. This leads to the third principle, the subject of utility. This entails analysing the action and the welfare of the society by establishing which action maximizes pleasure and minimizes pain or more generally which maximizes utility and minimizes disutility.

Utilitarianism is therefore recognized as a consequentialist theory, as it holds that the right action is the one that maximizes utility. This perspective places emphasis on the outcome of the action rather than the action itself. As pointed out by Sinnott-Armstrong, utilitarianism has specific conditions that contribute to the theory's optimal functioning. One of these conditions are precisely the consequences of the action, according to Sinnott-Armstrong, "it denies that moral rightness depends directly on anything other than consequences" (2022, p.3). And so, the rightness or morality of an action is solely determined by the outcomes it produce. Although there may be other factors involved, what is important to distinguish is that what makes it morally wrong is its future effects on other people and not necessarily the fact that the agent performed a questionable action.

Since utilitarianism is a consequentialist theory, it is necessary to look at what exactly this entails. Sinnott-Armstrong notes that "Sidgwick seemed to think that the principle of utility follows from certain very general self-evident principles, including universalizability (if an act ought to be done, then every other act that resembles it in all relevant respects also ought to be done), rationality (one ought to aim at the good generally rather than at any particular part of the good), and equality ("the good of any one individual is of no more importance, from the point of view ... of the Universe, than the good of any other")" (Sinnott-Armstrong, 2022, p.14). What is meant by this is that Sidgwick viewed the principle of utility as one that stems from principles that are self-evident for the individual. These principles are based on the idea is that if an action is morally right, then any action that is similar in all relevant aspects should be considered right as well. In addition, the individuals should aim for the widespread good rather than focusing on specific aspects of it, and the well-being of any individual should be considered as equally important in the broader context of the universe. Essentially, he believed that these principles are the foundation that leads to the portrayal of the principle of utility in ethical decision-making. Utilitarianism requires that the action performed considers and includes a sense of universalizability, rationality and equality which would then result in

a consequence. This will then make it possible to determine whether the action is right or wrong.

The theory of utilitarianism is often portrayed through trolley problems. Trolley problems are thought experiments used to test the critical decision making of an individual. It usually involves the individual having to choose to do harm to one person or a group of people.

The most common trolley problem is as follows:

> *"A trolley is going down the tracks and is set on course to run down five people who are tied to the tracks. The driver of the trolley has the option to divert the trolley onto another track in which only one person is tied. Foot wondered whether or not the driver should divert the trolley" (Andrade, 2019, p. 3)*

The trolley problems force the individual to make a decision, either to save the life of the one or of the multiple. In this instance, the life of the one or the life of the five. If we make use of the principles of utilitarianism to determine the morally correct action, we see that this would be sacrificing the one to save five. On the first principle, the action itself, the individual has the choice of sacrificing one and saving five or sacrificing five and saving one, those are the options. With that in mind, we then move on to the second principle and consider the social welfare of making either one of the decisions, that is, what the impact of killing one to save five will have on society and how killing five and saving one would affect the welfare of the society. Lastly, together with the comparison of the second principle and the options from the first principle, what now needs to be considered is which of the two actions of the first principle will maximize pleasure and minimize pain. Therefore, on the utilitarian account, it is clear, that the morally correct decision is to save the five and kill the one.

Consider another example of the trolley problem:

A self-driving car is navigating the roads of the busy city, in the distance a group of pedestrians cross the road, obstructing the car's way. The action to follow is that of the car, does the car swerve and miss the pedestrians and risk its driver, or protect the driver and hit the pedestrians. Using utilitarian principles, the self-driving car should choose to put the driver at risk as the utility of the pedestrians could outweigh that of the driver. However, the problems may not always be as straightforward.

It is necessary to note that, these trolley problems are often altered and at times it is proposed that the one individual that is being sacrificed is a loved member of the one making the

decision and that the five being saved are mere strangers, and a question then arises, does this change the decision being made? Since utilitarianism is primarily impartial on who the people are, it has no bearing on the decision being made. For the utilitarian, the right decision will always be to sacrifice the one individual and save the five. The pain generated by the loss of one individual is minimal compared to the pain that would be generated if five were killed. As such, killing one and saving five would maximize pleasure and minimize pain, and therefore it is the morally correct action.

On the other hand, however, there might also be an argument in favour of the one individual on the utilitarian account. If the one person is going to produce more utility, then utilitarianism would lean towards saving the one. For instance, if the five are criminals and the one is a nurse, utilitarianism would say save the nurse. Initially, when faced with this decision, a human might experience internal moral conflict, but ultimately, they might choose to prioritize saving the one person, thus adhering to the principle of minimizing harm. But when it comes to technology, these problems become even more complex. The driver of the trolley determines the outcome of scenario based on the necessary considerations, however, suppose there was no driver. Suppose in this case as well, it was a self-driving car, can we expect a car to make such determinations? Can we anticipate that a self-driving will be able to navigate and resolve such complex challenges and reach the same moral conclusions? These critiques create concern for the theory as the morally correct action may not always be as clear as the theory holds. However, the theory's principle on the determinations of actions can be useful for the development of ethical AI agents.

It is necessary to note that there are many advantages to the inclusion of utilitarianism in the context of machine ethics. Firstly, with the goal of creating AI systems that contribute positively to society, we see that the theory's focus on maximizing overall utility or well-being can aid in the accomplishment of this goal. Machines that are programmed with utilitarian principles will be able to prioritize actions that lead to the greatest good for the greatest number of individuals, thus promoting outcomes that are beneficial for society as a whole.

Secondly, utilitarianism relies on calculations which makes it suitable for informing machine ethics. The quantitative nature of utilitarian calculations allows machines to assess the potential outcomes of different actions based on their consequences for overall utility. It is by calculating and comparing the consequences of different actions, that machines can make

informed choices that aim to optimize positive outcomes while minimizing harm. Machines are good at making calculations and if it has the established structure of utilitarianism, the decision-making process is further enhanced and refined. This ensures that actions are guided by a systematic evaluation of outcomes that is based on maximizing overall utility. As such, it can be seen how utilitarianism is able to provide a clear framework for decision-making, which is crucial for AI systems operating in complex and uncertain environments.

Additionally, utilitarianism allows for flexibility and adaptation to various contexts and situations. AI systems can dynamically adjust their actions based on changing circumstances to ensure the best possible outcomes for everyone involved. From this it can be seen that utilitarianism can provide valuable guidance for machine ethics. It is important to note that although AI systems can radically alter actions for optimal outcomes, strictly following utilitarian principles risks neglecting individual rights and well-being, as seen in the case of the Utility Monster.

The utility monster is a hypothetical entity that gains significantly more utility from consuming a resource compared to anyone else. In the scenario where this entity exists, utilitarianism would account for the mistreatment or elimination of others, as the pleasure experienced by the utility monster would outweigh any suffering it may inflict. From this perspective it becomes clear that utilitarianism should not be the only theory or principle that guides the ethical behaviour of agents; it is clear that additional measures are needed. This will be discussed further in the section on synthesis.

To ensure a more comprehensive and balanced approach, it should be integrated alongside other ethical principles. This will be explored further in the section on the analysis of how ethical theories can be introduced to AI agents.

Section 2: Deontology

The ethical theory of deontology is one that uses rules to distinguish which actions are the morally correct actions. This theory operates on obligation and the element that there are universal rules that makes an action either right or wrong. It holds that the actions themselves are either permitted or prohibited, and not aimlessly determined by circumstance. As such, when deciding which action to perform, the question one has to ask is, is this right or is this wrong.

This theory is easy to apply as it often coincides with the natural order of things about what is ethical and what is not. What this means is that it is considered elementary to put into practice as it frequently corresponds with what people consider ethical or unethical based. This implies that ethical principles are the principles of reasoning that individuals naturally adhere to and ought to persist in following. It simply requires that individuals follow their intuition on the ethicality of actions and not stray from this natural order – that they do not abandon the principles of reasoning that they naturally follow. That they follow the Natural Law Theory. In other words, the theory is seen as practical because its principles and guidelines is likely to match notions of right and wrong that are commonly accepted. It implies that individuals can understand and follow this ethical theory because it resonates with their intuitive sense of morality, which is often in harmony with the way things naturally seem to be ethically structured. It avoids any form of ambiguity or indecision as the moral rules are usually clear such as, "Do not lie. Do not cheat. Do not steal".

Immanuel Kant is one of the advocates of deontology. According to Joseph Kranak, "Kant's ethics … focuses on duties, defined by right and wrong. Right and wrong (which are the primary deontic categories) are distinct from good and bad (which are value categories) in that they directly prescribe actions: right actions are ones we ought to do (are morally required to do) and wrong actions we ought not to do (are morally forbidden from doing). This style of ethics is referred to as deontology." (Kranak 2019, p. 53-54). What is necessary to note is that with Kantian deontology, a distinction is made between primary deontic categories and value categories. For this theory, the morality of an action is dependent on the primary deontic category and strengthened by the value category. It focuses on duties, with right and wrong actions being prescribed by deontic categories and value judgments are derived from adherence to these moral duties. Essentially, it boils down to doing what is morally right because it is your duty, with good intentions, and following principles that can be universally applied without contradictions.

Furthermore, Kranak holds that Kantian Deontology is based on the moral worth of the action (act or principle) itself and not the consequences. This means that the consequences of the action are not the primary focus; instead, what is the primary focus is the inherent nature of the action and the moral principles behind it. Kranak notes that Kant holds that morality must be rational, "he models his morality on science, which seeks to discover universal laws that govern the natural world. Similarly, morality will be a system of universal rules that govern action. In Kant's view, as we will see, right action is ultimately a rational action" (Kranak,

2019, p. 54). As such, for Kant, the right action is always rational as it is based on collective rules in nature and these rules were formulated through rationality. Essentially, morally right actions are inherently rational because they are associated with universal rules derived from nature and formulated through human reasoning, emphasizing the idea that ethical behaviour is grounded in rational deliberation and adherence to universally applicable principles. Moreover, according to Kranak, ethics for Kant consists of commands about what we ought to do, these commands are direct about the action that we should do and those we should not.

The most common examples are things such as the Ten commandments and the laws of the land – The Constitution. They depict direct rules and following them would portray deontology, such as do not kill, do not steal, do not harm, etc. Consider the following scenario, *'A'* is attacked by an armed robber. In the scuffle, after already being injured, *'A'* manages to break free and have the upper hand. *'A'* now has a choice to either retaliate or runaway. For those who follow deontological ethics, the right action would be to walk away as retaliating would entail causing harm or possibly killing the other person, which is wrong. As such, deontology dictates that *'A'* runaway.

It may be argued that *'A'* could defend himself as he was attacked. However, for the deontologist, the rule is clear, "Do Not Harm", as such, despite being justified through the actions of the attacker, the right action would always be for *'A'* to run away.

As mentioned, Kant's theory focuses on duties, and fulfilling these duties to the best of our abilities. For Kant, these duties are realised through what he calls the categorical imperative – the foremost principle of morality. This principle comprises two formulas and by using these formulas, a formulation of the correct action can be determined.

The first formula states, "that we ought to act in a way such that the maxim, or principle, of our act can be willed a universal law" (Wilburn 2022, p. 533). That is, we should determine whether the action that will be performed can be done by anyone and/or everyone all the time or whenever they see fit. If the answer to this question is no, if it cannot be universalised, the act is then morally prohibited. Wilburn uses an example of stealing bread to illustrate this point. His example is that if one is considering stealing bread, one has to ask if this act (maxim) can be made a universal law (2022, p.533). The question that we need to ask is if it is okay for everyone to steal all the time. From this, we see that the answer is no, as it would defeat the purpose of the maxim as a whole. Where the purpose of the maxim would be for all to have a supply of product but if everyone stole all the time (or had the permission to),

there would be product or property and stealing would not be possible. The goal is to produce maxims that can be supported by everyone, maxims that can develop actions and principles which when is done by all. This process ensure that the maxim is, as Wilburn put it, is consistent can be applied universally without contradicting its purpose.

The second formula states, "we ought to treat humanity (self and others) as an end and never as a mere means" (Wilburn, 2022, p. 533). This entails that all people should be treated with respect and dignity, that we refrain from using them as instruments to achieve my individual goals but instead help achieve their goals when possible. For instance, befriending your employee in the hopes of receiving a promotion. This means that people should not be treated as mere as tools to fulfil personal objectives, instead that they be approached with courtesy and consideration. It also involves actively contributing to each other's goals whenever feasible. For Kant, the fact that as humans, we possess the capacity for rationality and autonomy, it is important that we then use these capacities to treat each other with dignity and respect. This highlights the importance of the basis for the moral imperative that emphasizes notions of integrity through politeness, kindness, and thoughtfulness when one interacts with others.

Through these two formulas it is clear that Kant's moral theory centres around justice, individual rights, fairness, and consistency. For Kant, "the categorical imperative is the only rational moral rule, through which we can derive a complete, consistent set of moral duties and thus, every person who is fully following their rationality will agree on what is right and wrong" (Kranak, 2019, p. 63). Consider the following example, you are in an exam room with a friend. During the exam, your friend manages to ask you for an answer to one of the exam questions. On the one hand, you would want to help your friend, however, on the other hand, this would be considered cheating. The rule is that cheating is wrong and that you should not do it, as such for deontology, you should not help your friend. This seems straightforward, but now consider that instead of you helping your friend, there is a 'humanoid' in the exam, and your friend asks the humanoid for help. A question is then raised whether the humanoid will be able to reach the same conclusion. To this one might say that it might reach the same conclusion as it can be programmed with the rule that cheating is wrong and should be avoided. However, a problem arises as that same humanoid might be programmed with a rule that states, 'You should always help others'. Will it be able to determine which rule outweighs which in these types of scenarios? To this, I'd say that in order for the humanoid to reach the ethically correct decision on the deontologist account, what is needed is not merely rigid rules

on right and wrong actions but instead the rules need to be more sophisticated like in actual computer programming. If *X* then do *Y* unless *Z*, then do *A*.

Deontology ethics is therefore straight-forward with regards to how rules should be followed to perform the most ethical action, however, it has been critiqued based on this. One of the main critiques of Deontology is the fact that the 'right' action, is not always the morally correct action that should be performed. For the purpose of the creating of moral AI agents, there may be a way to minimise these types of critiques by introducing additional ethical principles to navigate these issues. It should be noted that, deontology is also well aligned when considering its inclusion to machine ethics.

We see that deontology is a suitable for use in machine ethics as it is centred around principles and rules, which make up the early stages of AI development such as GOFAI. In GOFAI, the AI systems that were produced, were predominantly structured around rule-based algorithms, and specific instructions were programmed to govern decision-making processes. It is these rules that served as the foundation for AI behaviour. By replicating the principles of deontological ethics, which prioritize adherence to moral rules or duties regardless of consequences, machine ethics will be further elevated.

Furthermore, Moors deontological reasoning in AI, which involves the application of ethical principles and rules within AI, suggests that AI systems can be designed to follow moral rules that is similar to those followed by humans. This allows for the incorporation of deontological principles into the decision-making processes of AI agents. The deontological approach promotes moral duties that are driven by ethical principles and rules regardless of consequences, making it appropriate for machine ethics.

The similarity between GOFAI and deontological reasoning is such that, just as deontologists stress the significance of following moral rules, GOFAI systems were designed to adhere strictly to programmed rules without considering other factors. The merging of rule-based systems in GOFAI with the principles of deontological reasoning suggests a natural compatibility between deontology and machine ethics. This compatibility offers a familiar framework for designing AI systems that prioritize ethical principles and rule-based decision-making, laying a foundation for ethical AI development.

This will be further explored in the analysis section.

Section 3: African Ubuntu Ethics

African Ubuntu Ethics is an ethical theory that places emphasis on the communal values of actions and attributes the rightness and wrongness of an action based on the value associated with this. Philip Ujomudike defines it as the ethical theory that focuses on, "a set of values central among which are reciprocity, common good, peaceful relations, emphasis on human dignity, and the value of human life as well as consensus, tolerance, and mutual respect" (Ujomudike, 2016, p. 2869). This theory applies the concept of Ubuntu to the way we determine our actions. The word Ubuntu means humanity towards others, and it is more commonly translated to the statement, "a person is a person because of or through others" (Eze, 2010, p. 190). According to Thaddeus Metz, "to have ubuntu is to be a person who is living a genuinely human way of life, whereas to lack ubuntu is to be missing human excellence" (Metz, 2014, p. 6761). Simply put, having Ubuntu means to live in a way that truly reflects what humanity is, it means embracing and displaying kindness, community, and compassion. It means that the person values the well-being of others and actively contributes to it as well. On the other side of this, the lack of Ubuntu means that the essence of human excellence is missed – that all these characteristics are absent from human interactions. In essence, the concept of Ubuntu is one that emphasizes the importance of a shared sense of humanity in how we live our lives and human relations.

From this we can see that Ubuntu is rooted in communal association and the social engagement, whilst still promoting self-realisation of the individual. Its focus on these aspects provides the basis on which the theory functions. According to Metz, the ideal form of interaction between people is perfectly portrayed by this harmonious type of community, it is a way of life that people relate to and should strive to create and maintain. In order to attain this sense of community required by this theory, Metz notes, that what is needed are a combination of two components. These components are, "identifying with others and exhibiting solidarity with them" (2019, p.2). Through understanding and supporting one another, we contribute to an environment where human excellence can grow.

According to Metz, the notion of identifying with others is, "for people to treat themselves as members of the same group: to conceive of themselves as a "we," to take pride or feel shame in others' activities, and to engage in joint projects, coordinating to achieve shared or at least compatible ends" (Metz, 2019, p.3). This means to acknowledge the humanity of others and conduct oneself as to portray this, to engage in shared tasks and create bonds which can be

grown through association with others. It is the idea that people embrace a sense of unity and consider themselves as part of this group. Apart from considering the self as part of this group, one should also aim to coordinate their actions to achieve shared goals and aim to align goals with that of the others. Essentially, the component focuses on the shared emotions and collaborative efforts toward mutual identification and raising a sense of interconnectedness and community.

The solidarity involves, "engaging in helpful behaviour, acting in ways that are reasonably expected to benefit others, it requires attitudes, emotions and motives that are positively oriented toward others' good, say, by sympathizing and helping them for their sake" (Metz, 2019). This entails that the nature in which the action performed by the individual is governed by and grounded in integrity for oneself and others. It means having positive motives directed towards the well-being of others and acting in ways that goes beyond the action itself so that it would likely benefit others. It is about more than just performing actions that are helpful but also having a genuine wish to contribute to the betterment of others.

From these components we see that with African Ubuntu Ethics, the rightness and wrongness of an action, therefore, depends on the value of the action in the context of the larger community. When one thinks of which action to perform, one must consider the value of the action in relation to how it impacts the community as a whole. With the first component the actions should align with a shared goal of the collective, and with the second component, the action itself should not be held in higher regard than the well-being and needs of others, so the value of the action should be determined based on these conditions. For example, when coming across an angry drunk, suppose this drunk speaks foully to you. You have the option to either respond as foully as have been treated or not engage. On the account of Ubuntu ethics, to not engage with or even offer assistance to this drunk holds more value than reciprocating this behaviour, as such the correct action would be not to engage with this drunk. The conduct of the individual must show an elevated moral standard and demonstrate virtuous behaviour. The significance of actions in terms of their moral value is crucial in the make-up of a person's character as it contributes to an individuals' ability to live a virtuous life. This highlights the way in which the act of ethical decision-making is connected to personhood.

African Ubuntu Ethics is often associated with Ifeyani Menkiti's concept of personhood. Katrin Flikschuh notes that Menkiti describes personhood as, "a temporally bounded 'ontological progression' from an it to an I" (2016, p.4). For Menkiti, personhood is a notion

that has to be achieved and not merely attributed, he differentiates between being a human being and being a person. It is a concept that is dynamic and whose distinction is attributed through progressive development that transforms and unfolds over time. It involves the deliberate and ongoing process of becoming, evolving from a mere 'it' – a biological existence – to an 'I,' a fully realised person with a sense of self, agency, and moral awareness.

Personhood is seen as a process human-beings need to engage in certain experiences in order to be considered a person. This process involves community engagement to achieve membership, "personhood is achieved through moral membership in communal life" (Flikschuh, 2016, p.4). For that reason, Menkiti holds that infants are not considered persons. Personhood is acquired by individuals passing from infancy to early childhood and gradually obtaining responsibilities and corresponding entitlements, in social and communal settings throughout adolescence. This would result in individuals gathering the necessary requirements to be a person. On Menkiti's account, to be a person one must be conscious of oneself as a participating community member and since infants do not have the capacity for reflexive self-awareness as yet, they are not yet persons. As Flikschuh notes, "an infant is a human being with biological needs and wants which the community is called upon to satisfy for it but since an infant does not as yet play an active moral role in the life of the community, it cannot be considered a person on Menkiti's account" (2016, p.5). For Menkiti, there is a distinct difference between being a human being and being a person. This difference is that being a human being with biological needs is not enough, the individual has to play an active moral role in the community. His framework for personhood involves more than mere biological existence, it requires an individual to actively contribute to the moral camaraderie of the community. Therefore, Menkiti views personhood as a process of social engagement to obtain moral membership in the community. From this we see that for Menkiti, a person is a moral being.

The relationship between Menkiti's conception of personhood and African Ubuntu Ethics is therefore clear. Both concepts place a great importance on communal engagement and reflexive self-awareness. According to Wareham, an important factor of the conception of personhood is the ability to 'live' in harmony that is in sync with oneself and others, "conception of personhood requires the capacity to co-exist in friendly and harmonious relationships of identity and solidarity" (Wareham, 2020, p. 131). An individual's capacity for communal association and the social engagement, whilst still promoting self-realisation, is central to the theory of African Ubuntu Ethics and is the foundation of personhood for

Menkiti. The acts that the individual performs is therefore motivated by these conditions which will determine whether the individual is a moral being or not. These actions must hold a value, it has to be behaviour showing high moral standards, behaviour showing virtue.

What is necessary to note is that, for personhood to be achieved within African Ubuntu Ethics it is traditionally associated with communal relationships and shared humanity, however, it's essential to recognize that personhood can evolve in different ways especially if the technological contexts are considered. In the context of AI, personhood could be understood not as a simulation of human socialization, but instead as a recognition of the AI's capacity to interact meaningfully with human society and contribute to collective well-being.

Given the characterisation of this theory, it can be seen that African Ubuntu Ethics is ideally aligned with the development of fully ethical agents. This is based on the fact that with moral agency places emphasis on the condition that the agent be able to integrate with and understand others (in the community and society) and African Ubuntu Ethics is primarily focused on this aspect, it is able to accommodate this need in moral agency. It promotes and creates the capacity within agents to do so. Since African Ubuntu Ethics focuses on actions rooted in a certain values and virtues that are advantageous to others and the establishment of communal association and the social engagement based on this, it is clear that agents would be able to perform actions that are beneficial to communities and as such they can be considered as fully moral agents on this account in that aspect.

In the attempt to create artificial intelligence (AI) that aligns with ethical principles, examining various ethical theories becomes essential. Utilitarianism, Deontology, and African Ubuntu Ethics represent distinct frameworks that offer insights into fostering fully moral AI agents. As shown, each theory presents unique perspectives on what constitutes ethical behaviour and provides guidelines for decision-making. In what follows I will examine these ethical theories within the context of AI development, and by doing so demonstrate how integrating principles from all three can enhance the ethical behaviour of AI agents, ultimately leading to the desired outcome.

Section 4: The Synthesis

Above I have looked at the three ethical theories that can aid in the development of ethical AI agents. I will now look at a way to establish the potential for success in this development and

explore ways we can produce AI systems infused with morality, thus ensuring their alignment with ethical principles. I propose that this be done by making use of principles derived from each of these theories.

When one considers these theories individually, for each ethical theory there are components that obstructs the use of the theory as the sole ethical guideline for an AI agent. For utilitarianism, we see that this theory encounters issues when the problems that the agent is faced with becomes more complex. Going back to the trolley problem on the self-driving car and it having to choose between sacrificing the five for the benefit of the one (the criminals and nurse), we are see when it comes situations that are out of the norm such that one person is able to produce more utility than five people, there is no way to determine or confirm if these agents will reach the morally correct conclusion.

With deontology, problems arises when the agent has to decide which rule to follow. Drawing on the example of the student in the classroom, we see conflict of the rules and the challenge lies in determining which rule should take precedence in the given situation. There may be principle suggests that the agent should adhere to moral rules regardless of the consequences, thereby prohibiting actions such as cheating, which are inherently wrong, but this places strain on the ability of a humanoid AI to comprehend and adhere to moral principles such as honesty and integrity.

African Ubuntu Ethics also faces obstructions when considering it as the ethical guideline for AI agents. This is the fact that, although it associates the rightness or wrongness of an action with the communal or societal value of the action, the theory itself is very unclear about how these values are ascribed and received. Furthermore, it emphasises that the right action will follow through the individual's virtuous nature. The value referred to in this theory does not solely rely on the action itself, but the individual performing the action too, but it does not dictate how these particular values are acquired. While each ethical theory presents valuable insights, none alone offers a comprehensive ethical guideline for AI agents. Utilitarianism faces challenges in complex scenarios where determining the greatest utility is ambiguous, while deontology struggles with conflicts between moral rules. Similarly, African Ubuntu Ethics lacks clarity in the ascription of values and the acquisition of virtues. Thus, a comprehensive approach that integrates principles from all three these ethical theories may better equip AI agents to navigate the complexities of moral decision-making.

On the utilitarian account, consider the condition that the actions of an agent are dependent

on the consequences thereof. The consequences determines whether an action is good or bad. Its dependence on the inherent utility of actions and the endorsement of societal well-being provides the presence of ethicality that agents need. The use of this theory promotes a moral high ground through its consequences and not its action. Although this may seem counter intuitive, as the action itself may be immoral, utilitarianism offers the capacity of the agent to make statistical calculations to needed to determine which action to perform based on its consequences.

The use of utilitarianism's consequentialist statistical calculations to assess the ethicality of actions, allows the agents to be able to achieve a certain level of ethical conduct. That is to say that the agent knows that the action that has an ethical consequence is the best action to perform. If agents are programmed with this theory, they can make use of these statistical calculations to establish the optimal outcome. The consequence will therefore be ethical, as the calculations made actively considers the ethicality of the consequence, the concern now is the ethicality of the action. This is necessary to note as a positive outcome does not necessarily indicate that the action that led to it was ethical. For example, a surgeon stealing medication to save the life of a patient. Although saving the patient's life is ethical – the act of stealing is still wrong. Therefore, this is a concern. Also, in the case of the Utility Monster. The Utility Monster problem portrays the tension within utilitarianism. That is, the possibility of overlooking individual rights and welfare in the order to maximise the overall utility of the one entity. As such, this causes strain on the notion of utilitarianism and how it is conducted. While this type of critique still presents a problem for these determinations, integrating another ethical theory into the decision-making process may mitigate these concerns. What is needed is a method to guarantee the moral integrity of both the action and its consequences, there needs to be a way to ensure that both the action and its consequences are morally sound. To accomplish this, I suggest the inclusion of deontological principles.

Deontology primarily focuses on the action and is governed by rules to ensure that the action is morally right. It avoids ambiguities by adhering to universal rules and obligations by not attaching a certain value to the act or consequences but solely applying collective rules to determine the right or wrong action. There is therefore a distinct difference between right and wrong and good and bad. With right or wrong, these actions cannot be disputed, and act is either right or wrong however, with good or bad, the 'goodness' or 'badness' of the act, is dependent on circumstance and scenario as well as consequences thereof. For this reason, if

instead of the agent solely relying on the consequence of an act being good or bad to determine its decision, if it also makes use of deontic logic to guide it to the right action, the agent would be performing the most accurately moral action.

Utilitarianism together with deontology creates the ideal condition for which to introduce ethical theories to agents. With deontology the agent is able to determine correct action based on its rules and this determination is done through the use of utilitarianisms statistical calculations of the action itself and the consequences thereof. See it as the agents being presented with the scenario and through the utilitarianisms considerations for consequences determines the action to perform, but before performing the action, it then makes use of the statistical calculations of utilitarianism and the rules of deontology to determine the if it really is the right action to perform. The agent is then able to perform ethically through the use of universal laws and the need to do good. Furthermore, the use of deontology strengthens the bond between humanity and these agents.

Previously, when considering the introduction of these agents to society, their ability to engage in meaningful social interactions have been scrutinized. However, the use deontology promotes the treating each other more than just a means to an end, but as the end itself for the progression of humanity. Apart from deontology, if we go further and add African Ubuntu ethics to the development of these agents, these agents will be able to stimulate meaningful social engagement.

With African Ubuntu ethics, the agents will be able to uphold communal values through communal association and social engagement in order to grow and maintain self-realization. The use of this theory advances the agent's ability to sustain harmony within the community by treating others the same as themselves. In fact, there are models in current AI deep learning that already allow for some aspects of this view to be achieved. This can be seen in instances when Large Language Models go through a process of correction called Reinforcement learning from human feedback (RLHP). Essentially, it validates its output with humans (who have been hired) to determine whether an answer fulfils the three H's: harmfulness, helpfulness, and honesty. By integrating Ubuntu ethics, an extra dimension for humans to give feedback on whether the AI answer also benefits the community can be introduced. This can be done by perhaps introducing an additional H, like health (community health). It is the through the acknowledgment of others and portraying them as such that bonds are created and grown. Through engaging with others in ways that are helpful and acting in these ways to

benefit others, the entity is able to develop high moral standards and build sense of self-awareness that is reflexive. The addition of African Ubuntu Ethics will empower AI agents to embrace the interconnectedness of humanity and prioritize the common good.

As such, the ideal ethical agent is one that holds not only one ethical theory but all three. While each theory has its limitations when applied individually, a combination of the theories offers a more credible approach to moral decision-making, the integration of utilitarianism, deontology, and African Ubuntu Ethics offers a comprehensive framework for the development of ethical AI agents. By equipping these agents with Utilitarianism, Deontology and African Ubuntu ethics, we will create an agent with the optimal capacity to form ethical decisions. The agent will be able to determine the most ethical action by considering whether it is right or wrong through the use of deontology and perform the action while upholding communal values through meaningful social engagement that is guided by African Ubuntu ethics and all these determines are done when the agent makes use of the statistical calculations that utilitarianism provides. If we incorporate these perspectives, we can strive towards creating AI systems that not only prioritize ethical behaviour but also foster positive interactions and contribute to the betterment of humanity as a whole.

Section 5: Putting Principles into Practice

In the analysis section, I proposed the use of principles derived from three ethical theories to construct the beginnings of a comprehensive ethical framework that is capable of effectively guiding AI agents. This section will look at which principle of each theory should be used in order to make this possible and explore the practical ways through which AI agents can integrate these principles into their decision-making processes. It will look at the specific steps and considerations that are involved when trying to shift from the theoretical principles into actionable guidelines. I aim to provide a strong basis for the ethical procedures that AI systems can rely on and explain how these agents can use these principles in practical scenarios.

From the proposal outlined above, we are left to question which specific principles from these theories should characterise the formation of this envisioned new ethical framework? With Utilitarianism, for instance, we see that the theory operates on a multifaceted framework aimed at achieving its desired outcome. This includes the notions that the maximisation of happiness is the ultimate goal as well as the fact that everyone's' happiness has significance when it comes to determining the morally right course of action. Similarly with Deontology

and African Ubuntu Ethics, there are comparable complexities in their ethical frameworks. Therefore, to construct this new innovative ethical framework, it is necessary to distinguish which specific tenets of utilitarianism, deontology and African Ubuntu Ethics should be extracted and integrated.

Firstly, from Utilitarianism, we adopt the principle of maximizing utility, wherein the maximisation of pleasure or happiness serves as the fundamental source for decision-making. Secondly, Deontology contributes the principle that emphasizes distinct duties as absolute rules for governing conduct, regardless of the consequences. In this lies the belief that certain actions have inherent moral qualities, independent of their outcomes and that this is what compels individuals to adhere to these moral obligations. Lastly, from African Ubuntu Ethics, the principle that promotes and fosters a collective well-being and harmonious relations among individuals. At its core, it endorses the elevation of human life by promoting dignity, tolerance, and mutual respect. These principles collectively serve as the basis for informing the formulation of the projected ethical framework, thus providing a foundation for ethical decision-making in diverse contexts. As such the components to be used are:

Principle 1: Deontology's Rule-based Principle.

The AI should be equipped with GOFAI rule systems that enable it to follow certain rules of action without fail.

Principle 2: Utilitarianism's Utility Principle.

An AI agent should always maximize utility either in terms of net pleasure over pain or preference satisfaction.

Principle 3: AUE's Community centred Principle.

The AI agents should be capable of upholding communal values through communal association and social engagement.

To this end, one might ask what this would look like in a real-life scenario. To illustrate this, consider the following example,

> *Imagine a situation where the agent is faced with the decision of whether to intervene in a public altercation where a person is being physically assaulted. The agent knows that intervening will likely result in personal risk and potential harm to themselves, but it will also likely prevent serious harm or even save the life of the person being attacked.*

On the deontological account, there is a duty to prevent harm to others, or the duty to help those in need, this may be considered fundamental moral obligations. As such, a deontologist may argue that it is morally required to intervene in the assault as it supports the duty to protect the well-being and rights of others, even if it involves personal risk. Having reached this decision, the agent then has to assess whether the other ethical principles conclusions will align with the one of deontology. This leads us to consider the perspective of utilitarians.

From a utilitarian perspective, the action of intervening would likely result in the greatest overall happiness or well-being for the greatest number of people. The intervention would minimize suffering and promote the welfare of the victim by preventing serious harm or saving the life of the person being attacked. Therefore, on the utilitarian account, intervening in the assault would be the correct decision. Now that it has been established that the utilitarian perspective supports the decision made by deontology, the next step is to examine whether African Ubuntu Ethics would reach the same conclusion.

African Ubuntu Ethics would likely support intervention in the public altercation as a means of upholding communal solidarity, expressing compassion and empathy, and promoting justice and harmony within the community, despite the potential personal risk involved for the individual intervening.
From this we see that all three ethical frameworks support intervention in the public altercation, thus making it the morally justifiable course of action.

However, what occurs when these ethical principles conflict and reaches different conclusions? In such instances, how does the agent decide on the action to follow? Consider, for instance, if the principle of harmony within the community, as held by African Ubuntu Ethics, fails to align with utilitarianism's emphasis on maximizing utility, or with deontology's focus on duty. In such a scenario, the agent faces a complex moral dilemma. They must weigh the merits and implications of each ethical principle and consider their respective strengths and weaknesses. This involves a thorough examination of the consequences of each route, as well as the underlying moral obligations and values inherent in each ethical framework.

From this we see that with deontology, there is a sense of importance for the adherence of moral rules and duties. This can serve as an initial framework as actions are evaluated against these established principles within the community, on the deontological perspective. After which, a rule-utilitarian approach is favoured, and the consequences of actions within the

context of these moral rules are considered. Rule-utilitarianism prioritizes actions that maximize overall utility when they are followed universally. In this context, utility is understood as more than just individual pleasure or happiness but instead as communal harmony and the promotion of its values. This then allows for the introduction of African Ubuntu ethics, as utility is defined as maximizing communal harmony and values. The action is then assessed based on its ability to foster community cohesion and solidarity. If this holistic understanding of utility is integrated into the ethical decision-making of the agent, it is able to represent these three ethical principles. This integration of deontological principles and rule-utilitarianism, guided by Ubuntu ethics, provides a comprehensive framework for ethical decision-making that balances individual autonomy with communal welfare. Although this approach provides a comprehensive framework for ethical decision-making, it is important to note that it is not the sole approach.

Consider the following example on how the aforementioned approach would work:

An artificial agent is tasked with making a decision regarding a small town facing an impending environmental catastrophe caused by a factory's toxic waste dumping. This factory is a major employer in the town, providing livelihoods to a significant portion of its population. Closing the factory would lead to widespread unemployment and economic turmoil for the community. However, allowing the factory to persist in its operations pose's grave risks to public health and the environment. Should the artificial agent opt to shut down the factory or permit it to continue functioning?

On the deontological account, there are certain actions that are inherently right or wrong, regardless of their consequences. As such, shutting down the factory may be seen as a moral duty to protect the rights and well-being of individuals affected by its toxic waste dumping, even if it results in short-term economic hardship for the community. For deontologists, principles such as justice, fairness, and respect for human rights are prioritised and one may argue that allowing the factory to continue its harmful practices violates these principles, regardless of the potential consequences. Therefore, from a deontological viewpoint, the AI agent may reach the conclusion to shut the factory down.

From a utilitarian perspective, the focus is on maximizing overall happiness or well-being.

The option of shutting down the factory may prevent environmental damage and protect public health in the long run and may be seen as promoting greater overall happiness for the community. However, the immediate consequence of widespread unemployment and economic hardship could lead to significant suffering and distress for many individuals in the short term. Therefore, a utilitarian calculation might suggest allowing the factory to continue its operations to minimize immediate suffering, even though it poses risks to public health and the environment. On this account, the artificial agent may reach the conclusion to allow the factory to continue its operations.

For African Ubuntu ethics, while shutting down the factory may result in immediate economic hardship for some individuals, it is the decision that best aligns with Ubuntu principles of communal well-being, environmental comfort, and social justice, ultimately promoting the overall welfare and sustainability of the community for future generations. As such, African Ubuntu Ethics would most likely support the prioritisation of the long-term well-being of the community and the environment, over short-term economic considerations. Therefore, with African Ubuntu ethics, the artificial agent may reach the conclusion to shut the factory down.

In cases such as these, where the principles do not reach the same conclusion, the agent may need to prioritize one principle over others. This means that the agent has to consider all of the factors involved, such as the severity of the situation, the potential impact on individuals involved, and the long-term implications for the community as a whole. Furthermore, they may also need to seek more moral guidelines that offer insights into resolving conflicts between the ethical principles. In cases where all three principles do not align, I propose that a type of ranking system be used to moderate and guide the actions of the agent. That is, that the agent follows the three principles in a certain order. This order would have deontology first. This is due to the programmability of GOFAI into AI agents, thus ensuring the agents adherence to moral rules and duties while prioritizing ethical principles. Second in this line of order would be utilitarianism. This is based on the fact that deep learning gives them access to big data and statistics, so it aids in the calculations needed for the maximization of overall welfare and minimizing harm and considering the consequences of AI actions. Then African Ubuntu can be programmed at the societal level such that it takes more than individuals into the calculation. This allows for the action to align with principles of interconnectedness, community, and empathy, emphasizing the importance of relationships and communal well-being in ethical decision-making processes. It’s important to point out that there could be circumstances where the ranking might require adjustment, this is dependent on the context

in which the agent is used (such as work, church, or family settings).

Ultimately, the decision-making process may involve a delicate balancing act. An act in which the agent has to assess both the significance of the consequential action and the significance of the ethical principle, which can be programmed as well. For example, while the deontological principle may hold greater weight than utilitarianism, the consequences of the action could diminish the importance of the principle itself. If one considers morality as a whole, human and AI, we see that this process is one that requires careful considerations of actions to be performed. As such, individuals must delicately balance the significance of their actions' consequences with the ethical principles that they uphold, which can sometimes conflict. This requires reconciling conflicting ethical considerations while remaining true to one's moral convictions and striving to promote the greater good or uphold moral integrity within the community. In essence, just as with humans, the agent also does all this in an attempt to reconcile conflicting the ethical principles while remaining true to their moral convictions and the overarching goal of promoting the greatest good or upholding moral integrity within the community.

PART 4: The Future of AI and Morality

Now that I have shown how ethics can significantly influence the moral conduct of agents and identified the specific ethical principles that can be used to facilitate this development, the next step is to analyse the possibility of this progression. Given our current understanding, this section explores the practicality of developing fully moral AI agents. Specifically, it examines the plausibility of programming artificial entities with qualities of morality, including diligence, frugality, honesty, discipline, politeness, cleanliness, unity, and generosity.

The agent that is envisioned is one that takes the form of an artificially intelligent machine, which is capable of operating autonomously with rationality, while embodying the aforementioned capacities. The ultimate goal is for these entities to exceed the notion of mere machinery and potentially be regarded as entities that are deserving of the personhood status, thereby assuming the function of moral agents.

In essence, the proposed development seeks to bridge the gap between the concepts of artificial intelligence and moral agency, with a future where machines not only function

efficiently but also adhere to ethical rules that is commonly associated with human morality. This goal requires the examination of the technical, philosophical, and ethical challenges that are inherent with such an ambitious undertaking. By analysing the reality of this development, we can gain insights into the potential trajectory of AI advancement and its implications for our conception of morality and personhood.

Section 1: Future or Fiction

Given the current standing of AI development, it seems that this possibility has the potential of becoming a reality. In the first instance, AI has already been developed, all that is needed is a way to equip these agents with a sense of ethicality through which they could operate. I am of the opinion that a method to achieve this goal has already been created. I posit that machine ethics is the method by which to accomplish this goal. Therefore, it is important to explore the type of agents and AI that are capable of integrating the introduction of machine ethics and the suggested ethical principles, and how it can be used effectively. As mentioned, machine ethics looks at the principles instilled in these devices and creates standards on which to judge them. It evaluates whether these machines are capable of making decisions based on their programming as well as independently of these programs. With this already at our disposal, if machine ethics is further engineered with the inclusion of the ethical principles discussed in the analysis section, the incorporation of it into modern day devices, will make it possible to create a fully moral artificially intelligent agent. As such, fully moral artificial intelligent agents are the future.

There may be those who may view this possibility as fiction. These individuals may argue that AI has in fact reached its limits. One of these individuals is Jatindar Palahar, who holds that, "it (AI) has its own set of limitations that can affect its efficiency and utility, especially when we're unaware of our own gaps in knowledge" (Palaher, 2023). This is based on the fact that what is proposed in theory for these fully moral artificial intelligent agents is in reality, an impractical notion. Further to this he notes that, "the limitations of artificial intelligence become especially clear when we consider the ethical implications of AI decision-making" (Palaher, 2023). Although this possibility seems credible in theory, the practicality of it is questioned based on the tangibility of the features needed to create these agents.

Furthermore, it is impossible to be certain that AI can be moral. As mentioned, some of the key characteristics of a moral being is diligence, frugality, honesty, discipline, politeness,

cleanliness, unity, and generosity and it is impossible to imagine that AI agents can possess these characteristics. If we consider honesty, AI agents will not be able to portray this if the scenario requires that the agent needs to lie to maximise utility or similarly that the agent lie to uphold communal welfare in African Ubuntu ethics, if the goal of the collective necessitates defiance of these characteristics, these agents will react in this way. Although this seems to be in line with the requirement of what it is to be a moral agent – to act of its own volition, it however raises questions about the actions these agents performs and how we will be able to determine whether and when these agents are defying these characteristics are for the collective goal or individual desires. These people may argue that until these components can be deconstructed into tangible features capable of being programmed, we cannot create fully moral artificial intelligent agents.

Amidst these objections, it's important to recognize that there may be those who view fully moral intelligent agents as the future. The development of morally skilled AI agents holds immense promise for addressing societal issues, such as bias and discrimination in decision-making processes. Osasona, et al. (2024) looks at the ethical implication of AI in the decision-making process and state that, "ethical considerations extend to the broader societal impact of AI in decision-making. Striking a balance between technological advancement and social responsibility is paramount to ensuring that AI benefits society as a whole" (p.323). Furthermore, infusing AI systems with ethical principles creates a foundation of trust and acceptance of AI technologies among society and can lead to a more widespread adoption and integration into various domains. Additionally, by exploring the possibility for fully moral AI agents in its entirety, there is potential to introduce a sense of cooperation and understanding between humanity and machines. As Osasona notes that, the ethical implications shows that there is need for exploring the extent to which we can develop transparency, fairness, and accountability to address concerns related to bias, responsibility, and the broader societal impact of AI-driven decisions, "ethical frameworks must evolve alongside technological advancements to foster a responsible and equitable integration of AI in decision-making processes" (Osasona, et al., 2024, p.323). This reflects a commitment to using technology for the betterment of humanity. If AI agents that are completely moral are the future, it prompts us to consider what lies ahead. More specifically, what precisely will these agents look like.

The future proposed with these agents involves the emergence of AI agents as explicit ethical entities. These agents possess the capacity to determine the most suitable course of action based on their ethical framework. They go beyond mere information processing, and instead

focuses on ethical decision-making itself. As Moor suggests, these agents function by applying their understanding of ethical principles and representing them through comprehensive analyses. Unlike implicit ethical agents, who rely solely on predetermined actions and programmed responses that lacks inherent ethical considerations, explicit ethical agents have the potential to autonomously determine their actions and behaviours by explicitly considering ethical principles. Moreover, these agents exhibit adaptability when encountering real-life scenarios, even in unpredictable circumstances. There have been questions regarding the decision-making of these agents, whether these judgements are plausible and whether these agents can justify them. For this reason, the idea that explicit ethical agents can be produced, the idea that an agent capable of handling real-life situations involving an unpredictable sequence of events can be developed and introduced into the high functioning society was previously considered far-fetched. However, AGI and machine ethics makes it possible for these agents to justify their decisions and negate these objections.

As mentioned above, Hal Hodson defines AGI as the anticipated ability of an AI agent to understand and/or learn intellectual tasks like human beings can (2019, p. 2). It is the idea that these systems can think, comprehend, learn, and apply the intelligence that they are not only equipped with, but have accumulated, to solve problems like an individual would in each circumstance. As such, with the introduction of AGI, the agent has access to a vast array of tools to inform its decision-making. This way the agent can portray actions that are ethical explicitly. The capacity for this type of explicit ethical behaviour in these agents stems from the ethical framework that they programmed with. This ethical framework is ultimately upheld by machine ethics.

Machine ethics serves as the foundation of the ethical framework of these agents. It guides the behaviour and decision-making of these AI agents, by enabling them to navigate complex moral dilemmas and act in accordance with explicit ethical principles. By making use of machine ethics, AI agents can handle difficult moral situations effectively and make choices that correspond with the defined ethical principles. It serves as the assurance that machines exhibit moral behaviour, which is done by reconciling its computational abilities with its ethical capabilities. Another aspect to acknowledge is that machine ethics relies fundamentally on data that is collected and recorded, thus emphasizing the role that data ethics has in shaping the ethical integrity of agents. It aids in the validity and sustainability of the machine ethics that the agents are built on. Data ethics is essential for ensuring that new technology is accommodated by ethics as it allows for the acknowledgment and inclusion of

these technologies within ethics. If data ethics is correctly integrated, it will then result in a strong ethical framework that is capable of addressing ethical challenges effectively.

The relationship between machine ethics and data ethics is one of interdependence – for machine ethics to uphold ethical standards, it must be founded upon ethically sourced data. The ethicality of the data directly influences the ethicality of the machine. Machine ethics allows for the integration of ethical principles into artificial intelligence systems, thus enabling the systematic evaluation of ethical implications through different ethical theories and principles. Additionally, it also facilitates the analysis of ethical considerations surrounding algorithms, technological practices, and data distribution.

The idea behind the incorporation of machine ethics is that it provides machines with ethical principles that they can use to regulate their behaviour. It's necessary to note that when it comes to programming ethical principles in machine ethics, there are a multitude of principles from various ethical theories that are used, and these principles are comprehensive in scope. It is a culmination of all the principles of all the ethical theories that grounds the idea that machine ethics is about making sure moral behaviours are upheld by human-made machines, that there is a sense of ethicality present when it comes to artificial intelligence. Although this poses its own problems, it ultimately aids with ensuring the ethicality of the agents. In order to minimise these issues, I proposed the use of three ethical principles in particular. This would then be the driving force of the agent.

The introduction of machine ethics to explicit ethical agents, presents the opportunity to program a sense of ethics into machines. The issue with this is the diverse range of ethical theories from which machine ethics draws its foundation. For this reason, I proposed that, instead of making use of various ethical theories, machine ethics focuses on three specific ethical principles from three ethical theories namely, Utilitarianism, Deontology, and African Ubuntu ethics. This would then make up AI agents of the future.

By incorporating the principles of these three ethical theories into the design of ethical agents, we can facilitate with their ability to make optimal ethical decisions. Utilitarianism provides the agent with statistical calculations to assess the greatest good for the greatest number, while Deontology allows the agent to evaluate actions based on principles of right and wrong and African Ubuntu ethics guide the agent with upholding communal values that emphasizes meaningful social engagement.

This approach minimizes the risk of having ethical principles that conflicts and overwhelms

the decision-making process. Likewise, with these three principles there is a way of initiating a framework for resolving conflicts that may arise between the principles, such as a ranking system. In instances where the principles reach different conclusions, the agent must prioritize based on various factors such as the severity of the situation, the potential impact on individuals involved, and the long-term implications for the community.

In the end the decision-making process involves a delicate balancing act, where the agent strives to reconcile conflicting ethical principles while remaining true to their moral convictions and with the goal of promoting the greatest good or upholding moral integrity within the community. Through the comprehensive integration of ethical principles, machines can navigate complex ethical dilemmas with greater efficacy and reliability.

To those who view these agents as fiction based on the abstract and immateriality of the notions of ethics, we find that this objection can be addressed through the use and implementation of the three specific ethical principles. It is possible to program the principle of utility by differentiating happiness from unhappiness. Similarly with Deontology, it is possible to program absolute rules and duties and on the African Ubuntu Ethics principle, it possible to program the prioritization of community harmony over discord.

Also, to the notion that it is impossible to be certain that AI can be moral, if the agents are explicitly programmed with a moral drive, this concern is minimised. The ability to program moral principles into AI systems allows us to establish guidelines and frameworks for ethical behaviour, much like how humans learn and adhere to moral codes through education and societal norms. Therefore, while there may be challenges in ensuring that AI systems behave ethically, the attempts of creating morally responsible artificial agents are both feasible and necessary for their integration into society.

For this reason, I believe, fully moral AI agents are the future. It is possible to produce explicit ethical agents. That is, agents who are autonomous and is able to adapt. These agents that do not solely make use of their information processing abilities but takes it further through homing in on the ethical framework that they are equipped with. It is possible to program this proposed ethical framework, which is driven by machine ethics that is rooted in principles of three specific ethical theories. This can be done and through this these agents will be able to determine the most suitable course of action based on programming and potential for growth.

Section 2: Concerns about the Future

As society moves into the direction of AI acceptance, the development and use of AI agents have become increasingly widespread. These intelligent systems show significant abilities, from problem-solving to autonomous decision-making. However, as we further explore the realm of AI, there are concerns about the implications of these agents on our future trajectory that need to be addressed. This gives rise to many questions, one question in particular, is whether we should continue to develop AI agents? Should we be developing AI agents and is an ethical framework really necessary? This discussion looks are the reasons for and against the creation of AI agents with ethical capacities.

As previously stated, those who hold that these agents can be developed, that fully moral AI agents are the future, may envision these agents as explicit ethical agents equipped with AGI and that makes use of machine ethics to regulate behaviour through principles of ethical theories. Given this, it becomes clear that there could be individuals who disagree with the development of these agents. Therefore, I will now look into the reasons that oppose the development of AI agents.

Those who object to the development of AI agents may do so for a number of reasons. The first being the notion that if fully moral artificial agents were to be developed, they be developed as explicit ethical agents. This objection arises from the inherent implications associated with explicit ethical agents (EEAs). As Anderson and Anderson (2007) notes, these agents are ones that are expected to be able to react to in an ethical way when exposed to unethical conditions by calculating the best ethical action to these dilemmas based on their ethical structures. This ethical structure is created by the fact that they are programmed with multiple data sets of ethical and unethical behaviours and principles, through which they will be able to compute and determine the most ethical reaction to circumstances and situations where necessary and react in that way and operate effectively based on the knowledge gained through this.

This calls into question the ethical structure of these agents. If this structure involves exposing these agents to data sets made up of ethical and unethical scenarios in the hopes of them determining the ethically correct action, this structure then depends on the data that these agents are programmed with. This could then cause friction for this notion as it calls into question the plausibility of this data. This ranges for where the data comes from, how the data was obtained, to the validity of the data – whether it is accurate and without bias. Since there

is a possibility for biased data, the ethical structure that determines the actions EEAs perform is flawed. Furthermore, as Anderson and Anderson notes, "having all the information and facilities in the world will not, by itself, generate ethical behaviour in a machine" (2007, p.15). As such, the implications of creating fully moral artificial agents as explicit ethical agents may be problematic. Even if it were possible to generate accurate data, there is no way to guarantee ethical behaviour from this.

Furthermore, Moor introduces the autonomous component to the concept EEAs, "explicit ethical agents as autonomous in that it could handle real-life situations involving an unpredictable sequence of events would be most impressive, explicit ethical agents are believed to hold the capacity to determine their own actions and behaviour" (Moor, 2006, p.20). Based on Moor's notion that these agents function ethically through accurately representing scenarios and calculating the relevant actions needed in the moment, the need for autonomy can be seen. The autonomy condition of EEA relies on self-governing; this raises the question of how this is achieved. The acquisition of autonomy requires the self, overcoming certain experiences to achieve independence, to achieve autonomy. And although there are experiences through which the agent can achieve independence, it may to a certain extent, always be reliant on programmed data and trained information, autonomy is possible but is limited, the agent will not be able to be fully autonomous. To this one may say that it could be programmed, however if it is programmed into machines, it loses this fundamental feature. Furthermore, the prospect of creating this sense of self-governance, this sense of autonomy, if we allow these machines to operate on its own accord, it leaves room for a lack of control of these agents, further calling attention to these catastrophic implications.

On the notion that it be developed with AGI, the idea that these systems can think, comprehend, learn, and apply the intelligence that they are not only equipped with, but have accumulated, to solve problems like an individual would in each circumstance would mean that these agents act on their accord. Although it is said that with AGI the agent makes use of both programmed information and its ability to assess a situation, to achieve the most accurate ethical action, there is no way to determine which information the agent made use of in each scenario. If the agent is to evaluate certain scenarios, assimilate to it as well as modify its response to accommodate the specific scenario – how do we determine which factors actually influenced the decision making of the AGI agent.

Since there is no way to regulate or even know the way in which AGI interprets information

and we cannot guarantee that the decisions they make are in fact legitimately informed – that is, using the information that they are programmed in the correct way – it creates concerns for the future. Deep learning may provide a solution to this dilemma since the actions it produces are derived through neural networks with precise inputs for certain scenarios. However, although deep learning provides a slight certainty of ethical actions, there is the aspect of spontaneity and the likelihood of unforeseen circumstances that it may not be able to account for. For instance, say an agent with AGI comes across a scuffle between two friends. From programmed information they might be able to recognise that it is a minor disagreement between friends and reacts by letting it go. However, from learned experiences as well as assimilation, they might consider this scuffle as a threat and react by intervening. Or perhaps it is a threat, but the AGI agent sees it as a simple scuffle. The same can be said for the way in which deep learning determines the decisions it makes. These agents would then be acting of their own accord, and this may not necessarily always be as ethical as envisioned.

To this, one might encourage the need for machine ethics, to eliminate these risks. Recall machine ethics is defined as, "a field of research that studies the possibility of constructing machines that can mimic, simulate, generate, or instantiate ethical sensitivity, learning, reasoning, argument, or action" (Guarini 2013, p. 213). It is considered to provide a foundation for looking at the components that governs artificially intelligent agents from its computational abilities to its ethical capabilities. Machine ethics is seen as a way to regulate the behaviour of the machines, however, for machine ethics to do so, for it to accurately depict ethical principles and synchronise its actions, it should be programmed on ethical data. This causes trouble for machine ethics.

The problem with gathering ethical data for machine ethics is the fact that data often varies based on the complexity of ethics in general and the fact that not all values are necessarily ethical. What this means is that the presence of non-ethical values, as well as the complexity of ethical principles, creates challenges for gathering data for machine ethics such that the ethical data may be compromised. Where ethical data refers to information of data collected that aligns with ethical principles and standards from the way it is collected, processed, and used. With this is the fact that, the process of determining the ethicality of data can compromise the integrity and usefulness of the data itself. This can be done in many ways, such as manipulation, censorship, transparency issue and more. As such, it is hard to determine which data are actually ethical or just data. And since there is no way to uncover the ethicality of data without certainty that the data will not be compromised, the validity of

machine ethics is called into question.

Furthermore, the data that is believed to be used in the programming of machines is data based on history. This then further negatively impacts machine ethics. Since we know of history to be biased and promote inequality, Stephen Buranyi raises the question, "how we prevent these programs from amplifying the inequalities of our past and affecting the most vulnerable members of our society" (2017, p. 4). If the data we program into these machines are derived from historical inequalities, we are essentially promoting our own biases and expecting these machines to operate based on it, this would have disastrous implications for humanity. What is also necessary to note is that the fact that machine ethics allow for the simulation of human interactions, furthers these implications.

As Buranyi references Bryson, "people expected AI to be unbiased; that's just wrong. If the underlying data reflects stereotypes, or if you train AI from human culture, you will find these things" (2017, p.5). As such, even if inequalities and biases are eradicated now, the machine will not unlearn these data and inequalities now as it is still evident in the world we know today, it is still existent in interactions. And through the fact that machine ethics promotes the simulation of human interaction, these factors cannot be eliminated but instead are further instigated. This goes for the simulation of reason and logic as well. Since reason and logic informs the actions, these aspects essentially make up the foundation of the biases seen through actions.

As such, for machine ethics, although it is possible to create these machines with the ability to simulate human reasoning and logic, it runs the risk of bias as reason and logic are subjective to communities and individuals. To this, one may argue that since it works together with the ethical principles, this risk is eradicated.

Although this is possible, it is unlikely that this can be achieved. If one considers the whole picture, it looks like all aspects of fully moral artificial agents' rests on the ethical principles. This can be seen through the fact that, EEA's implications of the uncertainty of its ethical structure and autonomy relies on AGI and, AGI's inability to regulate the actions of the machines depend on machine ethics whose, complications with biased and inaccurate data relies on the proposed ethical principles. When you think of a final product of these agents, you could find yourself with a machine whose autonomy is hindered, whose reasoning skewed, functionality compromised and a programmer that has limited control of the machine. As a result, if we do decide to develop these agents, we could be making super moral agents

with unpredictable and unreliable qualities. Therefore, these agents should not be developed.

The argument against the continued development of AI agents is that all the components that make up these agents, from the type of agents they are to the type of AI they are equipped with and even the introduction of machine ethics relies too heavily on the inclusion of the ethical principles. It implies that the reliance is excessive however, it overlooks a few crucial points.

In the first instance, the ethical principles should not be seen as a mere added benefit to AI design, instead, it should be seen for what it is – the foundation on which to build AI agents. Without the inclusion of ethical considerations would mean that the moral implications of AI actions are neglected which could be more harmful and good.

Also, incorporating ethics principles promotes the alignment of AI systems with human values and societal norms. It helps eliminate risks when AI agents are tasked with operating in contexts where ethical standards need to be meet. On the note of eliminating risks, if the ethical implications of AI agents from the outset, the potential negative outcomes can be anticipated and addressed before it occurs. This approach is proactive and ensures responsible AI innovation.

Furthermore, by using ethical principles as the foundation of AI agents we are ensuring the long-term sustainability of AI technology. It is clear that with the rapid progression of technology, there is no escaping the introduction of AI agents, however, if ethics is the cornerstone of these agents, it is possible to build more resilient and sustainable solutions that can positively contribute to human well-being and societal progress.

For this reason, while it may seem that AI design relies heavily on ethical principles, this reliance is justified and necessary for ensuring the responsible development and deployment of AI technologies. Therefore, ethical principles should be embraced as integral components of AI design and implementation.

That being said, getting back to the question on whether we should be developing AI agents and if an ethical framework really necessary, to this I say yes, we should be developing ethical AI agents. As mentioned, the advancement of technology makes the introduction of AI agents unavoidable therefore, it is imperative that we prioritize adherence to societal moral norms in their creation. It is essential to develop AI agents with ethics in mind. Seeing that, although the option for ethicality in agents exists, producing ethical agents may be considered a

recommendation rather than a strict requirement. This implies that creating an agent with Strong AI that has no basis of ethicality is feasible, and this possibility alone carries significant consequences.

If we consider Khillar's definition of Strong AI as mentioned earlier, "the theoretical form of machine intelligence that views machines or programs with the mind of their own and which can think and accomplish complex tasks on their own without human interference" (2020, p.3), it is clear that the consequences of these agents would be catastrophic. These agents would be able to act without hinderance at their own discretion. Without a sense of morality or moral dignity the actions of these agents may not always be the right actions. They will be able to act on their own accord without being cognisant of the consequences of these actions and if they were they might not be concerned about the repercussions thereof. If one considers this possibility, the ethical ramifications is alarming. One would be left with agents with no moral principles guiding their actions that can cause destruction and disorder, not only in communities but, in future, in the world.

From this we see the importance of ensuring the ethicality of agents, it emphasises that ethics should not merely be regarded as a suggestion but rather as a necessary requirement. By mandating ethical standards, we can effectively minimise the risk of AI agents being capable of engaging in unethical behaviours. There is a need to uphold ethical principles as foundational pillars in AI development as it safeguards against the potential negative consequences of AI systems operating without ethical constraints.

Conclusion

Artificial Intelligence has been considered as one of the most prominent features of the future. Its growth and expansion have given rise to various future possibilities which raised concerns about the way in which we as humans function and how we will engage with these entities. In this thesis I have explored the possibility of developing fully moral artificial agents through machine ethics equipped with specific ethical principles, in order to establish whether this possibility is a reality. Throughout the thesis, various aspects of AI, ethics, and their integration have been explored, providing insights into the challenges and possibilities of creating ethical AI agents. This process has led me to arrive at the conclusion that fully moral AI agents are the future.

The thesis starts by looking at the birth of AI and how it has transformed into a surging concept that is integrated in to almost every technological advancement. It considers its historical evolution and the various types of AI. Through the Chinese Room Argument these fundamental concepts are further and explored and used to investigate the essence of understanding and consciousness in AI systems. From there it went on to deep learning to show how the progressions in AI and the advancements in cognitive science and connectionism presented the possibility for the ethicality of agents. Nonetheless, there were still concerns regarding the introduction of AI agents into society. Through this exploration, we see the significance of instilling AI with not only the computational competence but also the capacity for logic, reason, and ethical determination. This revealed that the growth of AI creates uncertainties that requires comprehensive exploration of its capabilities and philosophical implications. It then became evident that an ethical framework is needed to guide the behaviour of AI systems as they explore complex moral dilemmas and societal interactions.

The emphasis on the ethical limitations of technology of these agents, showed that more is needed for the ideal introduction of AI agents into society. It showed that AI needed a solid foundation on which to ground this introduction. This is where the use of machine ethics became significant. It is seen that the emergence of machine ethics, which is a promising field dedicated to instilling ethical sensitivity and decision-making capabilities within AI, provided the possibility of the required foundation for the introduction of AI agents. This is seen through the distinctions that are made between implicit and explicit ethical agents and explaining their respective approaches to ethical behaviour. Recall that implicit agents operate based on programmed actions, while explicit agents have the ability to calculate ethical actions based on predefined principles where the actions are ultimately determined by the agent self. Then with the prospect of fully ethical AI agents, a future where AI systems exhibit consciousness, intentionality, and the capacity for moral reasoning similar to human beings were presented.

This future prompted real concerns regarding the ethical dilemmas that these agents face, thus allowing for machine ethics to be re-envisioned as the resolution. Machine ethics provided the capacities that is needed to address these ethical issues and offered a platform on which to ground fully moral ethical agents. The fact that it considers the principles that are embedded in these devices, it can be used evaluate whether these machines are able to make decisions based on their programming, as well as independently of these programs. Although this

provided the foundation, it still needed an ethical framework which can be used to assess the ethicality of actions of the agents.

This need was addressed through the exploration of ethical theories, specifically the ethical principles from three ethical theories. These theories being Utilitarianism, Deontology, and African Ubuntu Ethics. With this it is seen how each theory offers unique perspectives on ethical behaviour, ranging from maximizing overall well-being, to adhering to universally applied moral rules and fostering communal values. This showed how the integration of principles from these diverse ethical frameworks, can provide the moral compass for the envisioned AI agents so that they can navigate the ethical complexities of the world with discernment and integrity.

Apart from the theoretical consideration, the practicality of this future was further explored through examples illustrating how these fully moral AI agents may implement the envisioned ethical structure that is proposed. The inclusion of real-life scenarios allows for the analysis and assessment of these agents through the lens of ethical theories. It shows how these frameworks can guide AI decision-making in navigating moral dilemmas and also brings attention to the importance of aligning AI design with ethical principles. By doing so, we are able to ensure responsible innovation and maintain societal harmony.

In addition, by looking towards the future where we can see an environment that includes fully moral AI agents, are indications of ethical progress. Regardless of the challenges that are presented by issues such as bias in historical data and concerns about autonomy, it is clear that integrating ethical principles into AI design is of utmost importance. It is through the ethical concerns that AI technologies can be produced and used in a responsible way. If the ethical considerations are seen as requirements for the development and implementation of AI agents, it is ensuring that human values and societal norms are adhered to and upheld throughout the process.

In essence, what this thesis aimed to do is provide a transformative vision wherein AI goes beyond mere computational abilities and actually embodies ethical integrity and societal responsibility. By integrating ethical principles from diverse philosophical traditions and confronting technical, philosophical, and ethical challenges head-on, a path towards fully moral AI agents is insight, indicating a future where artificial intelligence serves as a signal of ethical enlightenment and human flourishing. AI developers can create agents that prioritize ethical behaviour, contribute to human well-being, and align with societal values.

This leads to the conclusion that fully moral artificially intelligent agents are the future.

References

Abumere, F. A. (2019). Utilitarianism. In F. A. ABUMERE, D. GILES, Y.-Y. (. KAO, & MICHAEL, INTRODUCTION TO PHILOSOPHY: ETHICS (pp. 46-52).

Allen, C., Wallach, W., & Smit, I. (2006). Why Machine Ethics? Intelligent Systems, IEEE, 12-17. doi:10.1109/MIS.2006.83.

Alshehri, A. S., Gani, R., and You, F. (2020). Deep Learning and Knowledge-Based Methods for Computer-Aided Molecular Design-Toward a Unified Approach: State-Of-The-Art and Future Directions. Comput. Chem. Eng. 141, 107005. doi:10.1016/j.compchemeng.2020.107005

Anderson, M., & Anderson, S. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. AI Magazine, 15-26.

Anderson, M., Anderson, S., & Armen, C. (2004). Towards Machine Ethics.

Andrade, G. (2019, March 17). Medical ethics and the trolley Problem. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6642460/

Arkoudas, K., & Bringsjord, S. (2014). Philosophical Foundations. In The Cambridge Handbook of Artificial Intelligence (pp. 35-63). Cambridge University Press.

Asimov, I. (1950). "Runaround". I, Robot (The Isaac Asimov Collection ed.). New York City: Doubleday. p. 40. ISBN 978-0-385-42304-5.

Boden, M. (2014). GOFAI. In K. Frankish, & W. M. Ramsey (Eds.), The Cambridge Handbook of Artificial Intelligence (pp. 89-107). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139046855.007

Boden, M. (2016). ‘What is Artificial Intelligence?’ In M. Boden, AI: Its nature and future. (pp. 1-24). OUP.

Brussels. (2018, December 18). A definition of AI: Main capabilities and scientific disciplines. The European Commission's High-Level Expert Group on Artificial Intelligence. Retrieved from https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

Buckner, C., & Garson, J. (2019, August 16). Connectionism. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/fall2019/entries/connectionism/

Buranyi, S. (2017, August 8). Rise of the racist robots – how AI is learning all our worst impulses. The Guardian, pp. 1-17.

Christen, M., Schaik, C. v., Fischer, J., Huppenbauer, M., & Tanner, C. (2014). Empirically Informed Ethics: Morality between Facts and Norms. Switzerland: Springer.

Driver, J. (2022). The History of Utilitarianism. The {Stanford} Encyclopedia of Philosophy.

Eze, M. O. (2010). Ubuntu: Toward a New Public. In M. O. Eze, Intellectual History in

Contemporary South Africa (pp. 181-192). New York: Palgrave Macmillan, New York. https://doi.org/10.1057/9780230109698_9

Flikschuh, K. (2016) The arc of personhood: Menkiti and Kant on becoming and being a person. Journal of the American Philosophical Association 2 (3):437-455. DOI 10.1017/apa.2016.26

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of The Royal Society - A Mathematical Physical and Engineering Sciences. 374. doi:10.1098/rsta.2016.0360.

Guarini, M. (2013, September). Introduction: Machine Ethics and the Ethics of Building Intelligent Machines. Springer, 213-215. Retrieved from https://doi.org/10.1007/s11245- 13-9183

Gunkel, D. (2014). A Vindication of the Rights of Machines. Philosophy & Technology, 27. doi:10.1007/s13347-013-0121-z.

Hodson, H. (2019). Deep Mind and Google: the battle to control artificial intelligence. The Economist. Retrieved from https://web.archive.org/web/20200707223031/https://www.economist.com/1843/2019/03/01/deepmind-and-google-the-battle-to-control-artificial-intelligence#

Horst, S. (2005). The Computational Theory of Mind. The Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/archives/fall2020/entries/computational-mind/

Johansson-Pajala, RM., Thommes, K., Hoppe, J.A. et al. Care Robot Orientation: What, Who and How? Potential Users' Perceptions. Int J of Soc Robotics 12, 1103–1117 (2020). https://doi.org/10.1007/s12369-020-00619-y

Jones, M. T. (2008). The History of AI: Systems Approach. In M. T. Jones, Artificial Intelligence: A Systems Approach (p. 13). David Pallai.

Joy, B. (2000). Why the Future Doesn't Need Us. WIRED.

Kelleher, J. D. (2019). Introduction to Deep Learning. In J. Kelleher, Deep Learning (pp. 1-38). MIT Press.

Khillar, S. (2020, July 14). Difference Between Strong and Weak AI. Difference Between Similar Terms and Objects. Retrieved from http://www.differencebetween.net/technology/difference-between-strong-and-weak-ai/

Kowalski, R. (1989). The Limitations of Logic and Its Role in Artificial Intelligence. In: Schmidt, J.W., Thanos, C. (eds) Foundations of Knowledge Base Management. Topics in Information Systems. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-83397-7_22

Kranak, J. (2019). Kantian Deontology. In F. A. Abumere, D. Giles, Y.-Y. (. Kao, M. Klenk, J. Kranak, K. MacKay, . . . a. C. Editor), Introduction to Philosophy: Ethics (pp. 53-65). Rebus Community.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. pp. 1-27. Retrieved from arXiv:1801.00631. (open access)

Metz, T. (2014). Ubuntu: The Good Life. In Encyclopedia of Quality of Life and Well-Being Research (pp. 6761–6765). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_4029

Metz, T. (2019, September 8). The African Ethic of Ubuntu. 1000 Word Philosophy – An Introductory Anthology. Retrieved from https://1000wordphilosophy.com/2019/09/08/theafrican-ethic-of-ubuntu/

Mitchell, M. (2024). Debates on the nature of artificial general intelligence.Science383,eado7069. DOI:10.1126/science.ado7069

Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems, vol. 21, 18-21. doi:10.1109/MIS.2006.80.

Northpointe. (2015, March 19). Practitioner's Guide to COMPAS Core.

Osasona, Femi & Amoo, Olukunle & Abrahams, Temitayo & Farayola, Oluwatoyin & Ayinla, Benjamin. (2024). REVIEWING THE ETHICAL IMPLICATIONS OF AI IN DECISION MAKING PROCESSES. International Journal of Management & Entrepreneurship Research. 6. 322-335. 10.51594/ijmer.v6i2.773.

Palaher, J. (2023, December 5). 5 Alarming Limitations of Artificial Intelligence in 2024 and the Vital Role of Human Expertise. Digital Rapport.

Pereira, Luís Moniz & Saptawijaya, Ari (2016). Programming Machine Ethics. Cham: Springer Verlag. Edited by Ari Saptawijaya.

Searle, J. (1980). Minds, Brains, and Programs. In J. Heil (Ed.), Philosophy of Mind: A Guide and Anthology. Oxford University Press.

Sinnott-Armstrong, W. (2022). Consequentialism. (E. N. (eds.), Ed.) The Stanford Encyclopedia of Philosophy (Winter 2022 Edition). Retrieved from https://plato.stanford.edu/archives/win2022/entries/consequentialism/

Srivastava, S. (2023). AI in Self-Driving Cars – How Autonomous Vehicles are Changing the Industry. Appinventiv – Artificial Intelligence. https://appinventiv.com/blog/ai-in-self-driving-cars

Taylor, A. (2003). Animals & Ethics: An Overview of the Philosophical Debate. Peterborough, Ontario: Broadview Press.

Ujomudike, P. O. (2016). Ubuntu Ethics. In Encyclopedia of Global Bioethics (pp. 2869–2881). Springer, Cham. Retrieved from https://doi.org/10.1007/978-3-319-09483-0_428.

Velmans, M. (2017). An Epistemology for the Study of Consciousness. In S. Schneider, & M. Velmans, The Blackwell Companion to Consciousness (pp. 769-784). doi:10.1002/9781119132363.ch54.

Wareham, C. (2020). Artificial intelligence and African conceptions of personhood. Springer, 127-135.

Wesche, J.S., & Sonderegger, A. (2021). Repelled at first sight? Expectations and intentions of jobseekers reading about AI selection in job advertisements. Comput. Hum. Behav., 125, 106931.

Wilburn, H. (2022). An Introduction to Western Ethical Thought: Aristotle, Kant, Utilitarianism. In H. Wilburn, PHILOSOPHICAL THOUGHT - across cultures and through the ages, 4th Edition (pp. 527 - 535). Tulsa: Tulsa Community College.