**MASTERS**

**APPLIED ETHICS FOR PROFESSIONALS**

**PHIL7005A**

**Research Report**


| | |
|---|---|
| Student Name: | Ponyane Mathabatha |
| Student Number: | 1569104 |
| Phone: | 0732060942 |
| e-mail: | libertymathabatha@gmail.com |


# CAN ETHICAL VALUES OF UBUNTU BE EMBEDDED IN AI?

A Research Report submitted to the Faculty of Humanities, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Arts, Applied Ethics for Professionals

29 September, 2020

## ABSTRACT

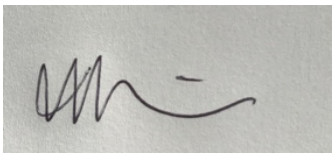
What began as tool making many millennia ago has evolved into artificial intelligence (AI) in the modern day era. Very few aspects of our daily lives are left unaffected by this technological transformation so much we can hardly remember how to commute or how to communicate without it. Our globe has shrunk into a global village within a generation.

This growth in AI development has been so rapid since the turn of the century that the idea of the technology reaching human-level intelligence is no longer a far-fetched one. As the machines get increasingly better at carrying out tasks which previously could only be performed by human beings, our reliance on them keeps growing. More advancements in AI inevitably lead to greater responsibilities being relegated to machines, giving rise to important ethical questions about what sort of moral values are promoted or diminished in society as a direct consequence of AI.

My thesis proposes that, like the good tool makers we have become, we must embed our values of ubuntu into AI and render it permanently predictable as tending towards respectfulness for life and consideration for the environment. AI design teams assembled in the *letsema* fashion can draw from diverse skills-sets to realise this goal.

## DECLARATION

I declare that this research report is my own unaided work. It is submitted for the degree of Master of Arts, Applied Ethics for Professionals, in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any other degree or examination in any other university.



_____

Ponyane Liberty Mathabatha

29th day of September, 2020

**TABLE OF CONTENTS**

## 1. Introduction

Technological developments have come a long way since the humble days of bone and stone tool making by our human ancestors over 1.7 million years ago. We have advanced from laborious crushing with edgeless blunt pieces of rock all those millennia ago to modern laser cutting techniques. Today we have invented many other technologies such as 3D printers - which make it possible to print intricate objects such as car parts and prostheses.

Humans are not necessarily the strongest in the animal kingdom, so what accounts for our rise to the very top? Much of the dominance of the human race over the rest in the animal kingdom is attributable, in part, to how people make tools and purposefully use them to accomplish specific ends. According to Sean Carroll (2015, 00:00:45), although it is true that other animals, like apes for example, are capable of making and using tools, the sophistication of tools made by humans – and our reliance on them for our daily living – sets us apart.

Today not only are our very lives reliant on technology, but it has also become so ubiquitous it is hard to imagine life without it. As we advance in what started as tool making, and come up with newer and more sophisticated inventions, we change the technology landscape; but the technology also changes us: our habits, how we learn and teach, how we commute, how we communicate, how we grow, preserve and process our food, etc. As the computer ethicist Luciano Floridi states, “today we are slowly accepting the idea that we are not standalone and unique entities, but rather informationally embodied organisms (inforgs), mutually connected and embedded in an informational environment (infosphere), which we share both with natural and artificial agents, similar to us in many respects” (Floridi, 2010: 10).

It is this similitude and interconnectedness between humans and machine, which Floridi is alluding to, that is so unnerving for some people regarding artificial intelligence (AI) and the future of humanity on the planet. One of the ways in which machines have become similar to us is that they have become intelligent, and their intelligence is growing so fast that the possibility of machine intelligence exceeding human intelligence is conceivable.

Some of the latest technologies include implants into human brains and artificially augmented limbs and organs for super (organic) human capabilities. The tools that we made in the earlier stages of our evolution we used for the purpose of extending our physical abilities; but the intelligent tools we make today promise to extend both our time, or more precisely how much we can accomplish per unit time, and our intelligence.

The ethical questions that arise from this are not limited only to how our own behavior is potentially influenced by technology but extend also - and perhaps more importantly - to what sort of ethical values are built into the computer artefacts themselves. Borne in mind when asking these questions is the escalating rate at which more machines and devices that are capable of performing activities which have morally relevant outcomes without direct human supervision are becoming available. The benefits from such sophisticated technologies are just as many as the associated risks.

No less a figure in the technology space than Elon Musk is quoted by the online magazine livescience.com as having warned about the looming dangers AI may pose to humanity at a meeting of the National Governors Association in July of 2017, saying “I have exposure to the very cutting-edge of AI, and I think people should be really concerned about it” (Weisberger: 2018, para. 4). Musk labelled this as an "existential threat". For his part, Stephen Hawking told BBC News in a December 2014 interview that "[t]he development of full artificial intelligence could spell the end of the human race." He further added, "[w]e cannot quite know what will happen if a machine exceeds

our own intelligence, so we can't know if we'll be infinitely helped by it, or ignored by it and sidelined, or conceivably destroyed by it" (Cellan-Jones: 2014. para. 11).

I agree with Stephen Hawkings that indeed AI could very well spell the end of the human race as we know it and that there exists potential dangers of AI turning against us, so to speak. However, I also think there is ample opportunity still for us humans to mitigate this potential risk by ensuring that future AI designs are embedded with ethical values of ubuntu, thus taking charge of our own future. I discuss in detail the notion of embedding ethical values in AI in chapters three and five. Therefore, I argue that those involved in AI design and development would do well to take measures to ensure that if AI ever exceeds our own intelligence it would be to our benefit and to the benefit of the environment at large – never destruction.

It is the purpose of this research to investigate whether it is both possible and indeed plausible to deliberately embed ethical values in AI and to consider which ethical theory would be most feasible for the same objective. The term AI admittedly gets thrown around a little bit to refer to any one of a number of subcategories of smart technologies available today. (Ro)bots, software, humanoids, computer programs, etc., all often get lumped together as "AI" without discrimination. This does not raise any concerns for the purpose of this discussion as I follow similar suite in my use of the term. Throughout the paper I use the terms AI, computer agents, artificial agents, bots, and robots interchangeably to refer to computer artefacts that are capable of performing activities and complete tasks independently of direct human supervision.

This investigation is interesting because artificial agents are fast developing capacity to perform morally relevant activities autonomously. Driverless cars, self-navigating drones, etc., are all examples of artefacts which can operate autonomously - i.e. they can respond and adapt to real world changes without human assistance. It is important that we are able to predict AI behavior reliably. If creators of computer artefacts

understand that they can and should consciously build ethical values into their artefacts, and if a method of building-in ethical values in AI is known then we can mitigate the risk of the emergence of destructive AI.

I will argue that it is possible to embed values in AI; and I will show that values have commonly already been embedded in many artefacts previously. I support this claim by way of examples of instances where values are demonstrably apparent in the artefact, and by showing how such values are embedded right from design phase. I explain important terminologies in later chapters, but I suggest it is useful to expatiate a little more on what I have in mind when I speak of morally relevant activities.

Obviously, not every activity in the world is morally relevant; some activities are morally neutral. A camera going into auto-focus for greater picture quality, a microwave oven stopping to heat up its chamber after the set time expires, a robot fitting a side mirror on a car in a manufacturing plant's assembly line. These activities are not morally relevant as they do not produce morally relevant consequences.

Morally relevant activities by computer artefacts are activities, performed by machines, which produce morally relevant outcomes. This is not to be confused with the consequentialist notion of right action, as performed by a moral agent, which I discuss in chapter four. Examples of the sort of machine activities I am referring to are (unfair) consumer profiling by algorithms running on bots, invasion of privacy by surveillance systems, collection of consumer information by Radio Frequency Identification (RFID) tags, etc. Such activities by machines have the potential of resulting in outcomes which are morally relevant, i.e. resultant states of affairs which can be judged to be morally good or morally bad. This is especially the case for AI as it has the capability, for example, of collecting otherwise morally irrelevant data from different sources and produce a morally relevant outcome through manipulation of the data.

For example, imagine a person $P_1$ has a mobile phone with a number *xyz*. Imagine that the number belonging to $P_1$, through a method of triangulation, can be confirmed to have been present at a particular restaurant at a certain time through an automated computer algorithm. Another person $P_2$ has a mobile phone with a number *abc* which can also be confirmed to have been present at the same restaurant at the same time through a similar method. Through a recognition of patterns, the computer program is now also able to confirm that the two meet at least three to four times a week for ninety minutes on average at different restaurants around town. From this example, one can already begin to imagine how different pieces of morally irrelevant data, e.g. mobile phone numbers and locations, through manipulation, can now be pieced together to potentially infringe on the privacy of both $P_1$ and $P_2$. Without a doubt this outcome by a computer software acting independently and through manipulation of morally irrelevant information is morally relevant.

In this essay, I shall draw some attention to AI's ever-increasing potential to perform activities which have morally relevant outcomes, independently of their users, and demonstrate how the creators of computer artefacts are responsible for embedding values into the machines. I will argue also that creators of computer artefacts are charged with the responsibility to embed moral values in the machines.

As human beings we have come a long way since the humble days when the tools we made, and used, were simple sticks and stones all those millennia ago. Now that we can make tools and machines which are AI capable, and therefore can perform morally relevant activities without direct supervision by a human being, it is much more important that we exercise a higher degree of accountability. This reality reasonably give rise to the imperative that every stakeholder (researchers, design engineers, financiers and sponsors of the project, testers, company executives, industry regulatory bodies, etc.) working to develop these technologies take seriously their moral obligation to concern themselves with ethics in their designs. This is required to ensure that the

computer artefacts they produce are embedded with ethical values to minimize the potential harm to humans and to the environment at large.

In chapter two I will discuss what artificial intelligence (AI) is and show that while human-level intelligence in computer artefacts is conceptually possible it is not quite a reality. This is despite many claims to the contrary by companies such as Google and IMB. The type of AI that is available today does not have the capacity to 'understand' the world beyond what it is designed to perform. An example would be how, during this period when there is a terrible COVID-19 pandemic in the world, none of the intelligent machines deployed in various industries, including financial markets, would be able to pick up the phenomenon of the Coronavirus and factor it into its decision making processes despite the fact that the information is widely reported in the news. A human being would need to intervene for the computers to factor in this big change in the world.

In chapter three I shall consider what values are, how they are ranked in order of importance, whether machines can promote or impede values, and whether it is possible to deliberately infuse ethical values in computer artefacts such that those embedded values become a characteristic feature of the machine - regardless of who uses it.

In chapter four I shall explore some ethical theories and argue in chapter five that ubuntu is the most feasible ethical theory to be embedded into AI. I conclude my discussion in chapter six.

## 2. Artificial Intelligence

Human intelligence is not the only form of intelligence. There are many other non-human entities which demonstrate intelligence such as can be observed in the animal kingdom, for example. Chimpanzees, dolphins and other animals display impressive levels of intelligence. Sometimes intelligence can be observed in animals one would not necessarily have expected to observe such high levels of intelligence in, like the fact that octopuses seem to have foresight and planning. Experiments conducted by marine biologists and researchers have demonstrated how octopuses can learn through observation, the famous one being one where an octopus could figure out how to open a glass jar by observing another octopus do the same.

Although humans are clearly not the only beings which possess intelligence, the concept of general intelligence is rather difficult to define without reference to human intelligence. Naturally, there are almost as many definitions of intelligence as there are people attempting to define it. Some narrow it down to cognitive abilities, others point to ability to solve problems creatively, ability to learn by observation, ability to 'read the room' and respond appropriately when interacting with others, and yet some insist it is a combination of these qualities. While the proverbial jury is still out on the definitions of intelligence, people generally have common understanding of what the concept entails. I shall use the concept of human intelligence as a launch-pad for a discussion on artificial intelligence (AI). I call it human-level intelligence.

For this discussion, I shall take human-level intelligence to be the ability to be aware of self in relation to others and in the context of the present environment in time and space, and the capacity to adapt to changes in the same environment. Self-awareness and ability to respond to environmental changes are therefore key features of intelligence. An intelligent being, in a conscious or 'awake' state, must be aware of its own existence and environing circumstances, and it must be able to navigate the space it exists in to avoid threat and increase comfort. Human beings have used their

intelligence to devise means to navigate literally across the world safely and comfortably for many centuries. The same is true, to some extent, about certain types of birds which migrate across continents seasonally. These birds surely have some level of intelligence; although obviously not human-level intelligence. The same can be observed in how many other animals respond to input information from the outside world collected through their senses of smell, sight, hearing, etc., to determine responses to changes in the environment. Today's smart computer artefacts can regulate their own temperature, save their onboard battery power, switch on or shut down automatically, etc., in response to changes in their environments.

Engineers 'teach' computer artefacts to be intelligent by installing a base program on the machine's semi-conductor hardware which determines the particular activities or functions that the machine can perform, which are the motivation for creating the machine in the first place. In this manner, the computer can then execute intelligent functions on its own according to the installed program. The machine can also be exposed to further related information on different data bases and through machine learning (ML) techniques the machine is able to gather, store, process, and recall on demand more information than that which was in the original base program. This would make the machine able to learn to perform other tasks and functions which may not have been on the original program. Deep learning can be made possible by directly interconnecting the machine to even more machines so that they can share and grow their data sets. In this way, and with access to powerful processors, the machine can be extremely efficient in recalling and processing multiple data sets at the same time and respond to changes in the environment spontaneously in real-time. A machine which has the capability to respond in this way is said to have Artificial Intelligence.

AI is machine intelligence or, stated differently, intelligence built into a computer artefact - typically in the way described above. As the word 'artificial' suggests, people write computer programs and software to execute commands such that the machine can mimic cognitive functions in its operations. Increasingly machines' ability to write software programs is growing at a rapid rate too such that in future newer versions of

the software running on machines will be written by machines themselves. Even then, people remain the initial authors of the original programs. Machine intelligence is therefore artificial intelligence.

In this chapter I shall consider in more detail what AI really is and the different forms it can manifest in. I also reflect on the pace of development in machine intelligence in the context of the wider evolution in tool making, for I consider intelligent machines to be part and parcel of this evolution and not a separate phenomenon. If intelligent machines could potentially be infinitely helpful to humanity or conceivably destroy life as we know it, I submit it is worthwhile reflecting how far the advancements in this area of tool making have come just as we also consider what potentially lies ahead. In so doing, opportunities can be created for and by the people involved in computer technology design to produce computer artefacts in a manner which ensures that the safety and security of humanity and the planet at large are secured. The future of mother earth should not be left to fate or fortune but designers of computer artefacts should take the responsibility to ensure that the technology remains predictably safe to operate.

## 2.1. Moore's Law

In his seminal paper in 1975, Gordon Moore describes how the number of transistors on a dense printed circuit doubles every two years - translating into faster computer processing power. This essentially means computing power doubles every two years. This phenomenon is commonly known as Moore's law. Such exponential growth in quite remarkable to observe over a significant period of time and boggles the mind in terms of what is possible in the future. The faster the computer processing speed the more instructions the machine is capable of executing per second.

To give some perspective, the first Intel central processing unit (CPU), which became commercially available in 1971, was only 740kHz and capable of processing approximately 92000 instructions per second. Today's processors are multi-core gigahertz (GHz) processors capable of processing over 100 billion instructions per

second. This trend accounts for much of the thinking that computers will soon possess intelligence superior to that of people and that, when this happens, they may begin to destroy life as we know it – and human beings will be outsmarted at every attempt to stop them.

The processing power of modern computers coupled with super-fast internet speeds in a connected world of the internet of things (IoT) - where everything is interconnected by means of fifth generation networks' small cells - make the creation of smart cities possible. In this connected world of the future machine to machine talking will be a common feature and intelligent machines will exchange data directly with each other at near real-time. This technology is already available today. Smart cars, wearable technology, home appliances, and a host of other devices will all be interconnected. Examples of applications of such advanced technologies may include smart city traffic management systems, with real-time monitoring capabilities, which can reroute smart driverless cars to avoid collisions and ease congestions in traffic jams, remote surgical procedures on patients by doctors from virtually anywhere on the planet, home refrigerators that keep inventory of supplies and feedback to the local grocer for which items to keep in stock on their shelves, etc.

There are undoubtedly many benefits to be derived from AI technology, just as there are associated risks. There exist valid concerns that must be addressed around the potential adverse impact of AI on various aspects of a human life, for example, information privacy, health and wellness. Users' weakened eyesight due to extended exposure to the screen lighting, or the obesity that is linked to lack of exercise for a couch potato, psychological effects such as increased anxiety when people have to be in face to face interactions due to being accustomed to online interactions – these are just some of the concerning downsides of technology. Time wastage is also another increasingly concerning downside of powerful pocket-size computers, both for productivity in the workplace and for effective learning.

With the rates of computing power, according to Moore's law, doubling every two years, the benefits of advanced technology will continue to grow just as the detrimental effects will also increase if nothing is done to address these effects. Avenues to eliminate the latter must be carefully considered and promptly put to action. A constant feature in discussions about technology in policy making, politics, sciences, arts, etc., is the possibility of androids, cyborgs, hybrids, humanoids, robots, and other forms of technology-enhanced humans exceeding human-level intelligence, which could render them effectively more powerful than people. When this happens, as some computer scientists, creators of science fiction and conspiracy theorists have warned, the fear is that the machines could potentially turn against the human race and destroy it.

The 1982 movie *Blade Runner* features a world where bioengineered replicants look and behave so much like humans, that the only way to determine that they are in fact not human is by a series of extensive tests conducted by a trained expert. In the movie, four of the replicants are on earth illegally and have to be hunted and killed as they had the potential to cause much harm.

Today, and because of the exponential growth in computing power, the idea of technology augmented humans is no longer as far-fetched as it was in the1982 when the film *Blade Runner* was first released. The same can however not be said about the evolution of the branch of philosophy called normative computer ethics. While topics that are already established in traditional ethics, such as privacy and the right to protection of intellectual property, have found expression even in computer ethics, there has not been nearly enough efforts channeled into understanding ethical considerations relevant for more complex issues in computer technology such as augmented reality and the moral status of technologically modified persons. Brain-implanted electrodes are now being used to control prosthetic limbs in humans (Lebedev and Nicolelis in Allen, 2010: 226), to cure epilepsy and drug addiction, among other uses. Recent advancements in the field of nano technology, development of wetware, modern super-

computer hardware, robotics, and machine written software makes it conceivable that a live replicant can be built. We are that much closer to having full cyborgs living alongside humans with intelligence that is superhuman in some ways, yet we are still to work out what moral values such creatures of our own making will possess. This question is in many ways either largely ignored or grossly neglected at present by the vast majority of philosophers and computer scientists alike. It is the purpose of this discussion to confront this question directly.

It would seem characters and scenarios that were mere creations of fantasy in science fiction (sci-fi) movies years ago may become our lived experience in the not too distant future. As the technology doubles every two years, it necessitates that we engage seriously with the ethical questions that arise from AI and do so with a sense of urgency before machines reach human level intelligence, if ever. This is important because the window of opportunity for meaningful intervention may be closed if and when machines reach this level of intelligence. More on this later in this chapter, but first I turn my attention to studying the different types of AI.

### 2.2. Weak AI Versus Strong AI

Artificial Intelligence is everywhere in our world today. It is in the predictive text on mobile phones, it is used in air traffic control, and it determines how much your insurance premium should be. Computer scientist John McCarthy, dubbed the father of AI, coined the term ‘artificial intelligence’ in 1956 at the Dartmouth Conference – which was the first artificial intelligence conference. He called for a two month long ten man study of artificial intelligence based on the conjecture that intelligence can be so precisely described that a machine can be made to simulate it. McCarthy's aim was to make a machine that was capable of learning, problem-solving, and reasoning like a human person. Although the success of the conference itself is debatable, in subsequent years it was to gain recognition as marking the occasion of the birth of AI.

Since then there has been significant strides in the field and today AI is generally accepted to come in two broad categories, namely weak AI and strong AI. Weak AI is taken to be machine intelligence which is below human-level intelligence while a machine is said to have strong AI capabilities when it can demonstrate human-level intelligence capability and higher. Put in machine logic terms, the main differentiator between weak and strong AI is whether the program running the machine is supervised by humans or not. In the case of weak AI, the machine searches for logic strings that resemble what has already been programmed into the machine by humans and responds to that input with a pre-determined output response. So, the output of a machine with weak AI capabilities never deviates beyond the human designer's imagination.

Essentially, weak AI is in fact not capable of comprehending the data sets it interacts with. Instead, through a process of sensing and classifying, it identifies pieces of information from the input and compares them to what is already in the program. If there is a match then the machine responds. Weak AI is incapable of doing more than what its human programmer inputs in the code. To give an example, if one were to ask Alexa (a virtual assistant from Amazon which operates on AI) to tell a "yo mama joke" keywords like "mama" and "joke" would be picked up and, through a classification process, Alexa would come up with a matching output without any real comprehension of the real question or the response. Comprehension requires understanding the *meaning* (the semantics) of the input and output, a capability weak AI does not have.

Weak AI coupled with machine learning (ML) can perform tasks which, at face value, may seem to be beyond what the initial programmer envisaged; but on closer observation it can be noted that it is the programmer who installed the base program and subsequently exposed the machine to the learning material in the first place. Another distinguishing characteristic of weak AI is that is highly focused on specific tasks, and it is very limited in its responses. Weak artificial intelligence is, therefore, not general intelligence as it can solve problems and make decisions only within the

confines of what its human programmers intended for it – any attempt to use it outside of this scope would amount to futility.

On the other hand, a machine has strong AI if and only if it has functions and capabilities that mimic the human mind. That is to say, it must have perception, self-reflection, sentience, consciousness, creativity, reasoning, etc. Strong AI, like a human mind, does not merely sense and classify as does weak AI. It is able to comprehend, to plan, and to form intentions. When in a humanoid form, it can locomote through most terrain while handling objects, tools and weapons with dexterity. For their part, Sadique Shaikh and Vasundhara Fegade have argued that “[strong AI, in the form of humanoids] will slip into our society seamlessly, and over time as the technology matures, fill roles not well suited to humans” (Shaikh and Fegade: 2013, 32). They will fit into our buildings, they will walk through our society, and they will manipulate the objects of modern life just as we do – if not better than us.

Now, referring back to the Alexa example, strong AI would fully comprehend the question and the answer. Hence, strong AI is sometimes also referred to as true Intelligence or general intelligence (GI) – terms that are easily interchangeable with human-level intelligence. Macquarie Concise Dictionary defines [general] intelligence as the “capacity for understanding and for other forms of adaptive behavior; aptitude in grasping truths, facts, meaning, etc.” (1988:501). Strong AI has selfawareness, which is a key feature of of human-level intelligence.

In terms of the current technological landscape, although some claims of serious breakthroughs are made from time to time by various entities, strong AI remains a futuristic concept – although some would argue we are not far off from realizing it. For now the current strong AI prototypes are viewed as 'mascots‘ and as their developers’ quest to lead in the invention and design of strong AI.

McCarthy himself admitted that just as it remains a challenge to define exhaustively what human intelligence really is, it is just as challenging to define precisely what strong AI is, since the latter is modelled after the former. In his attempt at a working definition for AI he concluded that a machine is said to have artificial intelligence if it can interpret data, potentially learn from the data, and use that knowledge to adapt and achieve specific goals (McCarthy, 2007:2). I shall take this definition of artificial intelligence, to be sufficient for the purposes of this discussion, and to cover the two categories of AI generally identified. Attempts to define strong AI in more detail quickly lead to existing difficulties that characterize the definition of human intelligence.

John McCarthy predicted at the 1956 Dartmouth conference that it would take anything between five and five hundred years for the breakthrough to strong AI. This estimation aligns with Moore's Law, which states that computing power will double every two years. Often science fiction movies do rather well at helping to conceptualize what strong AI will be like if and when advancements in AI reach that level. HAL-9000, from Stanley Kubrick's film *2001: A Space Odyssey (1968),* is a good example of strong AI. In the film, HAL-9000 is a sentient heuristically programmed computer, which controls the systems of the spaceship and can also interact with the ship's astronaut crew. HAL has all the features of a strong AI as he displays human-like behaviour, having the ability to learn, reflect on his actions, plan ahead and is self-aware. Although HAL is a fictional character, he does demonstrate that strong AI is conceptually possible.

Naturally, strong AI poses the most ethically relevant questions as it has greater ability to perform activities which yield morally relevant outcomes. That said, many artefacts may not, strictly speaking, satisfy all the requirements for classification as strong; yet still are capable of performing activities which yield morally relevant outcomes to some extent. The sort of AI which I have in mind for this investigation includes both strong and weak AI, henceforth simply AI. In chapter three I shall discuss the idea of performance of morally relevant activities by computer artefacts in greater detail.

## 2.3. Turing Test

Although fictional characters have been relied upon in the last section for a conceptual discussion of what strong AI truly is, a working definition does exist in theory. In 1950, Alan Turing developed what he dubbed the Imitation Game – now popularly known as the Turing test. The Turing test effectively tests whether a machine has reached human-level intelligence, and therefore fulfils the criteria for strong AI, or not. To pass the Turing test, a machine would have to fool a human into believing that it is a human. If it does indeed fool the human into believing that the machine is also a human, then the machine passes the Turing test (French, 2000: 116). For Turing, it is not a question of whether a machine can reach human-level intelligence (strong AI), rather it is a question of when.

To conduct the test, a man, a woman, and an interrogator sit in separate rooms such that they cannot see each other and communicate only via teletype. To start, the man pretends to be a woman and the interrogator, who can see neither the man nor the woman, proceeds to ask them a series of questions to try and establish which one of the two is the real woman. Halfway through the test, and unbeknown to the interrogator, the man switches places with the machine as the interrogator carries on with the rest of the questions. If the interrogator remains unable to notice the difference and cannot distinguish the real woman from the machine, the machine has passed the test. For Turing, this test provides a sufficient condition for Human-level intelligence, and therefore "there would be no reason to deny intelligence to a machine that could flawlessly imitate a human's unrestricted conversation" (French, 2000: 116).

Although some objections have been raised, the Turing test is widely accepted as the barometer for strong AI. Since the turn of the twenty first century a number of companies have come forward to announce that they have built a computer artefact, of one sort or another, which has strong AI capabilities. As impressive as many of the

inventions have been, none of them have satisfied the dictates of the Turing test in their entirety. One such example would be Google Duplex.

In May of 2018 Google's CEO unveiled Google Duplex, an artificial intelligence used with Google Assistant, to rival Amazon's Alexa. To show off the AI's impressive capabilities, it was given a verbal instruction to call a hairdresser to make an appointment. Not only did Google Duplex do well to sustain the conversation with the hairdresser, but also the sound of the AI's voice could not be differentiated from a human voice. The hairdresser on the other side of the call could not tell that he was not talking to another human being. Since indistinguishability is a key requirement for a machine to be considered to have strong AI capabilities, Google claimed that their Duplex had passed the Turing test. But did it?

According to Artem Oppermann from the online magazine *Towards Data Science*, Google Duplex did not demonstrate sufficiently that it meets the indistinguishability requirement because "[t]he audience (the evaluator of the test) knew exactly who of the two conversation participants the machine was. According to the Turing test the evaluator should not be aware of this fact" (Oppermann 2018. para. 7). Notice also that the original Turing test was to be conducted via teletype, so how close to human the machine's voice sounds is not a factor. What really matters for a successful Turing test is whether the machine can convince a human being to believe, judging by the reasonable manner in which it conducts the conversation, that it is reasoning like a human being.

Furthermore, the content of the conversation and the topic(s) covered in the discussion have to be open to flow in any direction to pass the Turing test, that is, the human participant should be able to talk about any subject to the machine. In the case of Google Duplex the conversation was narrowly confined to booking an appointment with a hairdresser. This presentation by Google, I think, is a fair indication of how close we

are edging towards strong AI; after it some industry analysts started suggesting that a machine could pass the Turing test by 2029. Till now no machine has been able to pass the Turing test, and therefore, strictly speaking, no machine has reached human-level intelligence.

Although widely accepted as an effective method of sizing up machine intelligence against human intelligence, there are still objections against the plausibility of the Turing test. I shall now turn my attention to the influential objection by John Searle.

### 2.4. Chinese Room

John Searle raises an objection against Turing's argument that if a machine can successfully fool a human interrogator and lead him to believe that it is, in fact, a human being, then that machine has a mind with the intelligence of a human. Searle does not dispute that a machine may very well fool some people sometimes, but he holds that just because a machine is able to pull it off does not prove that it has the intelligence to actually apply its own creative mind or that it *understands* what it was doing. More readily plausible, is that the machine has simply processed input through a system, such as a software program, and computed an output without any intelligent comprehension of either the input messaging or the output response (French: 2000, 119). As Albert Einstein once said, any fool can know; the point is to understand.

To illustrate his point, Searle proposes a thought experiment of his own known as the Chinese Room. Imagine there are two people, who are each isolated in their own separate rooms and who can only communicate with one other through the passing of notes with messages written on them. One of the two individuals is a native Chinese speaker, and the other does not understand a word of Chinese and can neither read nor write the language. The Chinese speaking native then writes a question, in Chinese, on

a piece of paper and then passes it through to the person in the other room who does not understand Chinese.

Imagine also that the non-Chinese speaking person has a Chinese rule-base with them which specifies inputs and outputs. The non-Chinese speaker, by referring in the rule-base, is able to match the Chinese characters from Chinese speaker's question with a corresponding string of characters in the rule-base, and he writes them on a piece of paper before passing it back to the Chinese person. Upon reading the response, the Chinese speaker is convinced that the person in the other room understands Chinese fluently. The truth is the non-Chinese speaking person comprehends neither the initial question on the piece of paper he received, nor does he understand what he wrote back in response (French: 2000, 119). So here 'understanding' is key.

It can be plausibly argued that the Turing test measures machine deception more accurately than it does human-level intelligence, because all that the machine needs to do to pass the test is deceive a human person into believing that they were in fact in a conversation with another human. Searle's thought experiment demonstrates that even if it was possible to build a machine that can deceive a human into believing they were interacting with another human, and this he takes to be conceptually possible, it would still not mean straightforwardly that the machine had any understanding of the content of the conversation.

Understanding itself, he takes to be a fundamental requirement in the demonstration of human-level intelligence. The machine, according to him, is only simulating an ability to understand – but fundamentally lacks (human-level) understanding. As far as Searle is concerned, the Turing test is not sufficient as a test for human-level intelligence. Because a human level intelligence is a broad concept with multiple aspects, which may include self-reflection, creativity, teachability, cognition, etc., it would be an error to

ascribe human-level intelligence to a machine which cannot demonstrate understanding.

As mentioned earlier, definitions of intelligence, and human-level intelligence in particular, are many and there is no general agreement even within the field of psychology. Among those who agree on the conceptual possibility of strong AI, there are those who contest that the Turing test alone is not adequate to determine human-level intelligence in a machine. They insist that moral awareness, and the thinking associated with it, is essential to confirm human-level intelligence, and that this should be factored in for strong AI to be realized. Those who hold this view propose an enhanced version of the Turing test, call it the Moral Turing test (MTT), that would be necessary before a machine can be said to have reached strong AI. According to them moral reasoning is a key feature of human-level reasoning.

While the requirement for a machine to pass the traditional Turing test is to simply demonstrate that it can reason like a human being, to pass the Moral Turing test it would have to outperform a human being in a test for good ethical decision making. This is because, as McDermott states, "ethical decision making involves a conflict between self-interest and ethics […] in order to be a genuine ethical decision maker one must also be free in the sense that one can sometimes choose to act in one's self-interest even though it runs counter to moral prescriptions" (McDermott, 2008: 2). Since a robot, like any other tool, is made by humans to help with tasks which are conceived by people and for the purpose of accomplishing human projects, it has neither free-will nor its own projects. It should, therefore, score higher than a human being – who possesses free-will and sometimes act out of self-interest - in order to pass an MTT.

The MTT approach assumes that morality is an important aspect of human intelligence, and that for a machine to attain strong AI it must also attain moral agency. For the

purpose of this research, I shall not go into a detailed discussion about artificial moral agency (AMA) – which the MTT approach advocates for – neither shall I study further the conceptual possibility of a machine passing a Moral Turing test. The focus of this essay is to investigate how designers of computer artefacts can ensure that they embed ethical values of ubuntu in the devices they manufacture. Therefore, discussions around moral agency of intelligent machines fall out of the scope of this study. What I intended to demonstrate in this section is that there remain divergent views on both the conceptual possibility of human-level intelligence in machines and tests thereof.

In the rest of my discussion on AI and its ability to perform activities which may result in morally relevant outcomes, discrimination between weak and strong AI is of no material consequence. For this chapter, it was my intention to outline, at least conceptually, what makes for intelligence in machines, both in its weak and strong forms. It also aimed to demonstrate that the greater the level of intelligence a machine possesses the greater its capacity to perform activities which result in morally relevant states of affairs. In the next chapter I shall interrogate how the ethical values are embedded in AI and argue that the morally relevant outcomes of the machine's performance flow from the embedded values.

## 3. Morally relevant values in AI

In the last chapter I explained artificial intelligence (AI), and how it is demonstrable in computer artefacts. I pointed out how it all began with human beings' quest for better and more efficient tools and so argued that AI is part and parcel of that age-old tool making trade. What makes AI particularly interesting above other tools is the rate at which computer artefacts are advancing in capability to mimic human behaviour and carry out tasks that previously could only be performed by intelligent human beings. This has been a cause for much concern, due to the uncertainty around whether AI will continue to serve purposes which are beneficial to human beings or if computer technology could be the cause of ultimate destruction, as Musk and Hawkins have suggested. This I put forward as the reason why it is useful to consider if ethical values of ubuntu can be embedded in AI.

I also distinguished between strong and weak AI and considered the extent to which either is capable of performing activities that can bring about outcomes which have moral relevance. There is a separate but interesting debate on artificial moral agency (AMA), in which there are opposing views about whether AI can be considered a moral agent. I shall not go into that discussion here. The reason for this is that while human-level intelligence is conceptually possible in machines, it remains a futuristic notion about which the scientific community is yet to agree on its precise measure. I would rather steer clear of that debate, especially because the moral relevance of the outcomes of tasks performed by autonomous machines can be studied independently of the machines' own moral status.

Christine Korsgaard pointed out that "one of the most important attributes of humanity is our nearly bottomless capacity for conferring value on almost anything. It is not because of our shared values that we should accord consideration to one another but because of our shared capacity for conferring value" (Korsgaard, 2003: 73). Because no one has the sole mandate of conferring values, no value judgement is immune to challenge. The soundness of the reason put forward to support the argument for the evaluation can be

challenged and therefore taken to be correct or flawed. One can justify one's value judgement by pointing out how the object or state of affairs in question brings about another state of affairs which is deemed to be good, or by arguing that the object or state of affairs is good for its own sake. Whatever the case, the moral agent making the value judgement cannot get away from justifying their evaluation, even if only for their own reflection.

In this chapter, I shall focus my attention to what such moral values are, how value is conferred on objects and states of affairs, and how (if at all) they are embedded in AI. I also investigate how to conscientiously embed them in computer artefacts such as AI.

## 3.1. AI Values

### 3.1.1. Intrinsic and extrinsic values

Paul Taylor (1972: 438-440) distinguishes different sorts of value judgements that people make. The most fundamental distinction he makes is between judgements of intrinsic value and those of extrinsic value. He explains that if "something has intrinsic value, its value is not derivative from the value of anything else" (Taylor, 1972: 440). However, if something has extrinsic value then its value is derived from the value of something else of intrinsic value. Extrinsic value itself, according to Taylor, can be derived in three different ways, giving rise to three different kinds of extrinsic value.

The first one is instrumental value. On this sort of extrinsic value, Taylor writes "[w]hen an object or act is judged to be good because it brings about, or helps bring about, some future state of affairs which is itself judged to have […] value, then the object or act is said to have instrumental value" (Taylor, 1972: 438). He uses a house key as an example. Assuming that entering a house is a state of affairs with some sort of value, the house key is valuable because it brings about that state of affairs. The value of the house key is derived from the value of the state of affairs which includes our entering the house. The same "derivation relation" is true of instrumental disvalue as well. Carl Wellman explains further that "the rational justification of any judgement of extrinsic value must consists of the valuation of other things […] and if those other things have

only extrinsic value, then any evaluation of them must be derived from evaluation of still other things" (1975: 78). Something is instrumentally disvaluable if it brings about or helps bring about a state of affairs with some sort of disvalue.

The second sort of extrinsic value Taylor calls "contributed value" (Taylor, 1972: 439). He explains here that a thing can have value as part of some larger whole, and that its value is derived from the value of that larger whole. He puts it thus: "Sometimes an object which is part of a larger spatial whole or an event which is part of a longer temporal process will be judged to be good in so far as its presence in the whole contributes to the value of the whole" (Taylor, 1972: 439). To make this point, he uses the example of a motorcar's spark plug - which on its own would seem to have little value. But as part of a functioning car, it does have value and its value is derived from the value of that car. Similarly, a portion of a temporal whole, such as "the solo portion of a piano concerto", can have contributive value too. While it may have some sort of value on its own, part of its value derives from the value of the concerto itself. Here the same idea can be presumed to be true also for contributive disvalue.

The third and last kind of extrinsic value according to Taylor is "inherent value" (Taylor, 1972: 439). "An object or event in public space-time has inherent value if and only if it is experienced by people in such a way that […] their experiences of it have intrinsic value to them." Taylor himself does point out the fact that inherent value for one person might be inherent disvalue for another, depending on whether their experience of the object or event is intrinsically valuable or disvaluable. To understand the notion of "inherent", consider an object or event – a symphony, a work of art, etc. – that gives rise to an intrinsically valuable experience for someone. The value of that object or event is then derived from the value of that experience. Here also the opposite, for inherently disvaluable things, applies.

Inherent value is however not limited to human creations. "A rainbow appearing after a storm" (Taylor, 1972: 440), he suggests, would have inherent value if (but only if) someone observing it had an intrinsically valuable experience of it. 'Potential' inherent

value is the idea that if someone were to experience the event or observe the object, then they would have such an experience. The main point, though, is that in any of these cases, the value of the object or event is derived value: it is derived from the value of the experience of which it is the object.

On the other end, to say a thing or a state of affairs has intrinsic value is to say the thing or the state of affairs has value in and of itself. Unlike the case of instrumental value for instance, where an object's instrumental value is evaluated on the basis of the state of affair it produces, intrinsic value is the ultimate value. One such example would be life; while there are other things which derive their value from the fact that they support life, life itself is valuable for the sake of life itself. Life is intrinsically valuable.

For the hedonist, pleasure is one such example of something with intrinsic value. For them, the pursuit of maximal pleasure and minimisation of pain given any set of circumstances is the ultimate aim; the most valuable thing to pursue. The value of an object of extrinsic value is determined by how well it can bring us closer to a state of affairs of intrinsic value. Other examples of objects of intrinsic value include happiness and love. Immanuel Kant's Categorical Imperative, which commands that we treat a human being as an "the end-in-him/herself" suggests that human life and human dignity are intrinsically valuable.

As Carl Wellman puts it, "all judgments of value must rest on judgements of intrinsic value, evaluations of things that have value for their own sake" (Wellman, 1975: 78). So, for an object to be considered to have intrinsic value the basis of this value judgement must lie within the nature of the thing itself.

### 3.1.2. Value-able AI

So, who has the authority to make valid value judgements? Is the value in each object and state of affairs conferred upon the object or state of affairs by a human valuer or are humans simply discovering value that has always been there? In this section of my investigation I intend to bring out two things: the first is that nature in general and life in

all its forms is intrinsically valuable and therefore should be respected by all – including AI. The second is that it can be proven that other non-human entities such as organisms, plants, and animals have proven to have the capacity to make valuations. I argue that AI should be developed to have this capacity to make valuations too. I take capacity to make valuations to be quite different from conferring value. While valuating is recognition of value that already exists, conferring value suggests that the value in the object or state of affairs did not exist prior to it being conferred by the valuer.

In traditional Western philosophy the dominant thinking around values seems to be an anthropocentric one - i.e. “the belief that man and his works are the center of the universe” (Guha:1989,1). Edwin Etieyibo explains that “the basic idea of anthropocentrism as a *human-centered ethic*, [is that] humans are taken to be the most important life form. Following from this, other life forms are important only to the extent that they affect humans or are useful to them” (Etieyibo: 2017, 4). Anthropocentricity is sometimes also referred to as chauvinism because on this view human beings are more important than any other thing in the universe - perhaps second only to God who granted people this status. For his part, John O’Neill points out that the “subjectivist claims that the only sources of value are the evaluative attitudes of humans” (1992: 121). The subjectivist view therefore is an anthropocentric one as it asserts that human beings are the only ones vested with the authority to make valid value judgements, and that the source of value is dependent upon human attitudes.

Wilhelm Windelband writes: “Value […] is never found in the object itself as a property. It consists in a relation to an appreciating mind […] Take away will and feeling and there is no such thing as value” (Windelband, 1921: 215). According to this view even intrinsic value is conferred by a human evaluator without whose intervention the object itself possesses no value. Value begins at the point of human contacts and continues in the mind of the observer, never as a property of the object itself. For the subjectivist, these values begin with, center around, and end with a perceptive individual doing the valuing. In an instance where the individual’s own judgement is that there is value in nature, then their perception becomes the truth-maker for that value.

But this chauvinistic view cannot be correct. To illustrate how flawed this view is, I will make use of the popular thought experiment known as the Last Man. This thought experiment has featured widely in environmental ethics discussions and it is useful to demonstrate that the value of the planet earth and life on it is dependent on neither the valuation nor the presence of human beings. The imaginary last man in the thought experiment acts in such a way that guarantees that the earth, and everything in it, will be totally extirpated after his own death; there will no other person(s) on the planet. Would it be reasonable to think something of value would have been destroyed if the last man succeeded in his plan? I take it most would agree, and our intuition is therefore that value can exist without the presence of humans.

The astronaut Michael Collins talking about the experience of his unique opportunity to observe the planet from without, saying “when I travelled to the moon, it wasn't my proximity to that battered rock-pile I remember so vividly, but rather what I saw when I looked back at my fragile home—a glistening, inviting beacon, delicate blue and white, a tiny outpost suspended in the black infinity. Earth is to be treasured and nurtured, something precious that must endure” (Gallant, 1980: 6). We understand that earth and the life on it has existed for millions of years before mother nature finally birthed humans too. It pre-existed human presence and therefore its value does not depend on human evaluation.

The value of the planet and its systems must be acknowledged and respected by all as intrinsically valuable. Those involved in designing and making AI have the obligation to embed ethical values in AI such that even if the technology were to exceed human-level intelligence in the future the machine’s answer to the Last Man’s question would be consistent with that of most rational people – i.e. that life itself and the planet as a whole would not lose its value even if there were no people living on it. Nature as a whole and in its various separate components has intrinsic value that no other entity, including AI, should ever be allowed to undermine. While AI must be value-able, it must also

recognise the source of the value. I discuss embedded values in greater detail in later chapters.

The view that holds that the source of values in nature is nature itself is known as realism. According to this perspective nature would remain valuable even if there was no human being to perceive and/or experience it, just as it was before people walked the earth. I take this view to be a correct one.

On the question of who has the capacity to make valid judgments of value, in his seminal paper entitled “Value in nature and the nature of value”, Holmes Rolston III (1994) addresses what he takes to be the unsoundness of the view which suggests that only humans can make valid value judgements by illustrating how other animals, plants, organisms, etc., are also ‘value-able’, that is, able to value things or states of affairs. He states: “There is no better evidence of non-human values and valuers than spontaneous wildlife, born free and on its own. Animals hunt and howl, find shelter, seek out their habitats and mates, care for their young, flee from threats, grow hungry, thirsty, hot, tired, excited, sleepy. They suffer injury and lick their wounds” (Rolston, 1994: 15). This proves that animals value certain things extrinsically, like their habitat which is useful for shelter and sleep which is good for rest. But they also are capable of valuing intrinsically, like they value their own lives. Rolston believes that other forms of life like animals and plants are value-able too – just as human beings are value-able.

If other beings far less intelligent that human beings are proven to be value-able then it follows straightforwardly that AI capable computer artefacts can be (made) value-able too. If even without self-awareness, or the ability to reflect on its own actions and plan ahead, a living organism is value-able, so also a machine should be designed to have the capacity to be value-able. This is particularly necessary for AI which has capacity to perform activities and tasks which have outcomes that have moral relevance, especially because AI can perform such activities without direct human supervision. In the decision-making map within the program running the AI, evaluation of probable states of

affairs that may result from each one of the actions which are available for the machine to choose from, should be one of the primary considerations. This point is important for our subject of discourse.

Not only should computer artefacts be embedded with the sort of ethical values that are to be promoted for a healthy and just society, but they should also be programmed in such a manner that they can factor in moral considerations in their decision-making processes. The value-ability to recognize the value of human life, nature in general together with its ecosystems, and the life-giving planet at large must be evaluated to be highly valuable by AI. Again, this is not AI conferring value on life but recognition of the value that life already has.

### 3.1.3. From Values to Ethics

Values can be further divided into moral or non-moral values. Examples of non-moral values include economic value, aesthetic value, efficiency, etc. A lot can be said about each one of these non-moral values listed as each is important for different reasons. Economic viability, safety of usage, the artefact's appeal in terms of utility and looks, etc., are all important values and they have their place in computer design. Stakeholders involved in the design and production of computer artefacts put a lot of effort to ensure that the products they make can compete well in the market on all these fronts. While I recognise the importance of non-moral values the focus of this discussion will be limited to moral value, here after referred to only as values.

As for moral values, William Frankena gives a comprehensive list which includes:

> life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and

> proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honour, esteem, etc. (Frankena, 1973: 87)

From this list, it is noticeable that moral values generally tend to be intrinsic in nature. They are values which are good for their own sake and do not derive their good from anything else.

How people weigh up their values in terms of which value takes more priority over another in a given situation is a product of their lived experiences and are therefore influenced by their own society, culture, religion, upbringing, etc. This is partly why organisations find it necessary to have ethical codes of conducts to serve as a guide for which morally relevant practices are representative of and consistent with the values of the organisation or juristic person, and how people should relate to one another within the organisation. Although different people have different value systems, there are however some moral values which seem consistent among different groups of people and which have remained relatively consistent over time. Justice, freedom, respect, privacy, honesty, etc., are some examples of such moral values.

When people are morally deliberating, they are essentially going through a process of deciding which state of affairs, among all the states of affairs they have both the opportunity and the capacity to bring about, is more valuable among all the others under the circumstances. This is especially true if they are act utilitarians. For the deontologist, the same process would be guided by which rule is most important to uphold under the circumstances. As James Rachels suggests, "this is the essence of morality: ...the morally right thing to do is always the thing best supported by the arguments" (Rachels, 1981:11). The 'arguments' here are the reasons that one would give in order to justify their moral decision.

This is why moral decisions attract either blame or praise, depending on the validity of the reasons given for supporting the chosen course of action. Every sane person is

vested with the personal responsibility to regulate their own intentions and associated acts. Morals are guiding principles by which members of society regulate individual conduct and use them as the standard for judging right from wrong. Actions that stem out of moral conviction sometimes result in the moral agent having to sacrifice their own convenience, comfort, and even profits. This is so because a moral agent ought to be both rational and impartial if their moral judgements are to be taken as correct. For his part, Brey says, "[m]oral values are ideals about how people ought to behave in relation to others and themselves and how society should be organised so as to promote the right course of action" (Brey, 2010: 47).

The words 'morals' and 'ethics' are often used interchangeably to describe a ranking of competing moral values, the valuations of which are backed by sound reasoning. I adopt the same approach here and alternate between the words. Morality can be thought of as a form of social contract which members of society "sign up for" and commit themselves to be careful to hold each one's own end of the deal, while reasonably expecting others to do the same - and holding them to account for such - in order for peace, orderliness and fairness to prevail. Morals are not based on one's personal feelings, but their moral obligation to let feelings be guided by principle and sound reasoning - even if there are foreseeable personal costs. Moral judgements are therefore quite separate from mere expressions of preference or personal taste. The process of valuation to arrive at a rational moral judgement demands a conscientious moral agent. One that can carefully and impartially sift through facts and examine the supporting reasons put forward for the arguments from behind "a veil of ignorance", to borrow the Rawlsian term.

Sometimes moral values clash in the manner described above and the question "what is the morally correct thing to do given the circumstances?" becomes cause of much debate. This is often not necessarily because of disagreement in the list of moral values, as listed by Frankena, but because of differences in how the moral values are ranked or prioritised by the different moral agents. For example, both peace and

freedom are moral values, but one moral agent may disagree with another which one is to be prioritised first in a situation where one must come at the expense of the other. One moral agent can put forward their own arguments as to why freedom must be upheld and cherished above and before peace while the other vehemently argues the contrary.

The story of Theresa Ann Campo Pearson, which was widely covered by the media at the time and appeared in the Los Angeles Times newspaper on 16 April 1992, is an example of one such moral dilemma. At the time journalists described the story as an "medical marvel" and a "ethical enigma". Famously known as Baby Theresa, the infant was born with a rare genetic disorder known as anencephaly - a condition in which a baby is born without a skull; but has only enough brain stem cell to regulate simple functions such as breathing and heartbeat. The cerebrum and the cerebellum are the other critical parts of the brain that are missing in babies born with this condition; which makes their prospects of ever leading a normal human life nil. Most cases of anencephalic babies get detected during pregnancy and the foetuses are either aborted in countries where the laws allow, and many are stillborn. Baby Theresa's condition was detected late in the pregnancy and was born alive - which is unusual for anencephalic babies.

The moral dilemma with baby Theresa arose from the request by her parents that her organs be harvested for donation to other children who desperately needed transplants. Upon learning about the condition of their baby, and with the understanding that she only had a few days to live, the parents decided that their child's imminent death should not be in vain and that to honour her memory they would donate her organs to benefit lives of other children in desperate need of organs. After all, their child had no chance of surviving more than just a few days and would never become conscious. They thought it would be best for physicians to harvest the baby's organs soon after she was born and use them to save other lives.

But is killing one person in order to potentially save the lives of a few more the morally correct thing to do? What about the baby's right to life? Does this decision undermine it? These are some of the questions that this decision by baby Theresa's parents raised.
Some among those who were opposed to the harvesting of the baby's organs, put forward the argument that it is not morally acceptable to use another person as a mere means to an end. By this view, ending baby's Theresa's life to save other lives would amount to using her simply as a means to an end. But is this argument sound? Rachel clarifies that "using people typically involves violating their autonomy - their ability to choose for themselves how they want to live their own lives according to their desires and values" (1981, 3). Put in standard form, the argument is as follows:

1. Using people as mere means to an end is morally unacceptable
2. Baby Theresa is a person
3. Killing a person to save other's lives is tantamount to using a person as a mere means to an end

---

(C) Killing baby Theresa for her organs is morally deplorable

In the argument above, premises 1, 2, and 3 are put forward in support of the conclusion (C); which is also the main argument.

There were others yet still who argued that there is no point in one being kept medically alive if one cannot enjoy life and have desires of one's own - as clearly was the case for baby Theresa. Like her parents, they held the view that if the baby is clearly going to die in a few days' time and has no prospects of ever becoming a conscious being, the reasonable thing to do is harvest her organs before they go to waste. Besides, she does not have autonomy and she possesses no will of her own. According to this view, and to put it in Peter Singer's words, "If it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, to do it" (Singer, 1972: 231).

Due to the strong opposing views and their prevalence the matter ended up going to court. Baby Theresa's medical condition, as expected, began to deteriorate rapidly after

her birth, she died nine days after - by which time all her organs were no longer fit for transplant. Even after her death, the debates wouldn't die down with the Supreme Court in Florida sitting two days after her death to hear the arguments.

Another example of a moral dilemma can be demonstrated in the following thought experiment (Screencast, 2019):
Suppose that X is married, and that the sex in X's marriage is not satisfying. Suppose also that X holds the following values: (a) marriage should be a sexually monogamous relationship, and (b) people have a right to a satisfying sexual life. Given that X's marriage is sexually unsatisfying, if X's value system evaluates value a to be more important than value b, then X will take it to be morally unacceptable to seek sexual satisfaction outside the bounds of the marriage and therefore choose to forego value b. X will rather either stick it out in the sexually unfulfilling marriage - and perhaps even try out various avenues to try to improve the quality of the sex - or seek dissolution of the marriage citing the depravation of value b as grounds. On the contrary, if X values value b more that value a then X would deem it justified to seek sexual satisfaction outside of the marriage while remaining within the union.

It is clear from the thought experiment and the example above how one's values inform one's moral deliberations. The risk of AI potentially becoming completely destructive can be mitigated by designers of computer artefacts becoming conscientious concerning the values they embed in the machines and ensuring that computer artefacts are made to be value-able – i.e. they are capable of appreciating (moral) values. Those charged with the design of value-able AI would have to consult more widely and ensure there is sufficient diversity in the teams involved. This would mitigate the risk of extreme bias for one set of values over another informed by the designers' own outlook on life and background in the quest for AI which is sensitive to ethical values. Too much emphasis of one value at the expense of other values could easily lead to a disvaluable state of affairs. Without this capability in autonomous machine it will be difficult to predict if indeed AI will one day destroy the planet – or not.

### 3.2. Values embedded in AI

Ordinarily, and in the course of everyday life, people tend to make moral value judgements on objects, states of affairs, the way people or organisations conduct themselves, etc., as being good or bad, right or wrong. Computer artefacts too are not spared these value judgments. Irrespective of their level of intelligence - whether strong AI or weak AI - moral value judgements can be made about artefacts based on the outcomes they help bring about. Computer artefacts can be evaluated to be morally good or bad.

Spyware, for example, can be judged to be a bad piece of software. Spyware is software that collects information about a person or an organisation and sends such information to another entity without the information owner's knowledge or consent. This is immoral and infringes on privacy. The information collected in this manner is often also used to commit even more immoral acts such as fraud and identity theft. Even when it is not in use, the spyware can be judged to be morally bad software in and of itself. The seemingly obvious state of affairs that will come to be as a result of software, once in use, is enough information for the moral agent to pass the moral judgement and conclude that the software has disvalue. Of course, its bad value becomes obviously apparent when it is put to use, but someone who knows about spyware and the nature of its function does not need to wait before it is in use for them to make the value judgement. Just as the actions and character of the developer who makes the software and those of a person who intentionally puts it to use knowing fully well what it is capable of can be judged to be morally unacceptable, spyware itself can be judged to be a morally bad piece of software.

In this section I shall defend the notion that computer artefacts tend to have built-in values and demonstrate how these values are embedded into the artefacts. I make the point that technological artefacts (and in particular computer systems and software) have built-in tendencies to promote or demote the realisation of particular moral values, and that the values in the artefacts can be studied independently of the values of the computer system's users. Such values need not be unavoidable in every instance or

necessarily evident in each and every single use of the artefact for us to be certain of their embeddedness within an artefact. Rather, the fact that they are common and easy to observe in the central uses of the artefact is sufficient to conclude that indeed such values are embedded within the artefact. By central uses I mean those uses that the system was designed for, and provide the reason(s) which can be plausibly brought forward as justification for the system's continued production.

Say, for example, one desires to test the download speed of their internet connection and goes about doing so by downloading spyware onto their machine online. It could be correctly argued that spyware, on that particular occasion, serves a good and valuable purpose - i.e being useful in measuring download speed of the internet connection. But testing internet download speeds is not a central use for which hackers spend much of their time developing spyware for. That is to say, testing an Internet connection's download speed is not the central use of spyware. The isolated occasion, therefore, where spyware is used successfully for testing internet download speed cannot reasonably be put forward as an argument for good values in spyware against the comparatively much more common use of spyware, which is to spy on unsuspecting computer users.

Brey uses the example of a fossil fuel engine motor vehicle to make a similar point concerning the notion of central uses (Brey, 2010). Such a motor vehicle can be used as a museum piece, as a barricade, or as shelter from the rain. These uses, he says, can be taken to be peripheral. The same motor vehicle can also be used as a taxi, to transport cargo, or for auto racing - uses which can be taken to be central for why fossil fuel engine motor vehicles continue to be manufactured. When used in the last three use cases, there are greenhouse gases emitted which pollute the environment and contribute to global warming - consequences of which are dire for humanity and many other life forms on the planet earth. Brey thus suggests from this example that "a case can be made that technological artefacts are capable of having built-in consequences in the sense that particular consequences may manifest themselves in all of the central uses of the artefact" (Brey, 2010: 44).

The aim here is not to involve ourselves in a discussion about whether computer artefacts, especially AI, can be taken to be moral agents or not. It remains an important topic, and computer ethicists such as Luciano Floridi (2010) have addressed that question. The focus for this research however is to demonstrate that AI is not morally neutral, and that it is possible to identify tendencies in technology artefacts which promote or demote particular moral values - computer ethicists call these embedded values. Although this research is focused on embedded values in AI, values are very often embedded in other non-technology artefacts too. For example, it can be argued that guns, in their central uses, result in loss of human life - and therefore are embedded with moral disvalue. Of course, a gun can be used for many other purposes that do not involve harming a human being, such as sports purposes, but someone else could argue that guns end up being used to shoot at another person - which can be evaluated as disvalue. The intention here is not to resolve this moral conundrum about the role of guns in society, but to demonstrate guns have central use cases, and further, to show that other non-computer artefacts too can have embedded values.

Since a gun commonly requires someone to be physically behind it to pull a trigger, in order for it to kill a person, it may not be as easy to separate the values embedded in the gun from the values of the person using it; this is unlike the case for AI - which we have shown in the preceding chapter is capable of unaided autonomous activities which result in outcomes that have moral relevance. As defined in chapter 1, a computer artefact is taken to be autonomous if it can demonstrate its own capacity to respond and adapt to real world changes without human assistance. This is not to be confused with the Kantian concept of autonomous nature - which speaks to a person's dignity and right to exercise their own will. There is therefore a greater level of responsibility especially on designers of AI and all relevant stakeholders to ensure that good values are embedded in AI while bad values are avoided.

Brey is correct in saying:

> Technological artefacts are hence capable of either promoting or harming the realisation of values when used. When this occurs

> systematically, in all its central uses, we may say that the artefact embodies a special kind of built-in consequence, which is a built-in tendency to promote or harm the realisation of a value. Such a built-in tendency may be, in short, an embedded value or disvalue (Brey, 2010: 46).

The values embedded in computer artefacts are often latent until such a time that the artefact is active and in use, whether activated by a person or an automated pre-set trigger. The mistake should not be made, however, to assume that value judgements cannot be made on the artefact when it is not in use or that it is not already laden with values or disvalues when it is not in active mode, just like spyware remains an undesirable piece of software even when it is not in active use (as discussed earlier in this section). Should an unsuspecting person who does not know much about spyware put the software to use intending no harm, the software itself will begin spying soon as it is in use. Spyware can thus be said to be laden or embedded with disvalue.

While in the case of spyware might be more obvious to see how the artefact is laden with disvalue by intention of the designer, in many cases disvalues are not embedded consciously. One such example would be the difficulty of the initial versions of facial recognition AI not being able to recognise black African people's faces with the same level of accuracy as it does people of Caucasian descent. Over the years, many black people have had problems using the technology – sometimes even being unable to unlock their own mobile phones. This clearly is a disvalue which disadvantages a specific segment of the world's population on the basis of race. While it may not be obviously apparent that the technology is embedded with disvalues, at least until a black person attempts to use it, and the designers might not have consciously intended to build-in this disvalue into the technology, it still exists within the technology.

We have shown that computer artefacts are embedded with values, and that these values manifest in the central uses of the artefact. I shall now turn my attention towards the question: how do the values get to be embedded into the AI? Values are embedded

or built into the technological artefact, intentionally or unintentionally, in a number of ways: the designers' attitude, the culture of the society from which those with significant influence on the design come from, technical constraints, the social context in which the system is used, etc.

There is always a link between values themselves, as held and evaluated by a valuer, and the valuer's own lived experience, as I have previously discussed. I also argued that some of these values are often unconsciously held and expressed; and that they play themselves out in ordinary decision-making processes and moral dilemma resolutions inside a moral agent's mind. Since computer artefacts are an expression of the creator's (and other stakeholders involved in the process) own impression, as with all other artefacts, it follows straightforwardly that the designer's own values will be built-into the artefact. Lily Díaz-Kommonen suggested that "material and immaterial aspects of culture as well as their history are embedded in artefacts" (2004, 13). She further adds "objects of art that are expressive artefacts partly result from the intrinsic motivation that arises from within the individual who is fashioning the object" (2004: 35). On this view, the artist's own attitude and biases will tend to show in the end product - built-in and embedded into the artefact.

I take it the creative process of making art, be it an ornament, a painting, or a sculpture, is quite similar to the creative process of making a tool or a computer are artefact in that both have their genesis in the imagination of a creative mind. End products are, to whatever extent, an expression of what began as a concept in the creator's mind and therefore tend to reflect the creator's attitudes and world views.

Steps should be taken by all relevant stakeholders in AI design, manufacturing, and capital investment to ensure that, as far as practicable, moral values are embedded into AI conscientiously. Technology advancements reached AI level through sustained and improving tool making by people, and therefore we should take responsibility to ensure that tools that we make do not get out of hand - so to speak. If ever they did we risk losing our ability to know and predict with any level of certainty if, as Stephen Hawkins

put it, “we will be greatly assisted by them, or conceivably destroyed by them in the future” (Cellan-Jones: 2014. para. 11).

Sometimes, depending on the nature of the project, it is feasible to have the intended moral values which should be deliberately embedded in the artefact explicitly stated in the functional definition of the system. This requires that design teams reflect exhaustively on the possible sources of values with the aim of deliberately promoting or demoting the realisation of those particular moral values by the artefact. Once the sources of values are identified they can meaningfully form an integral part of the design process for conscientiously embedded moral values.

Flanagan et al. demonstrated how this conscientious approach to embedding ethical values in AI can be applied in practice in a project which became popularly known simply as Rapunsel. They described the project as “a large multidisciplinary collaboration aimed at designing and implementing an experimental game prototype” (2008:331). Ultimately the main aim of the game itself was an attempt to promote interest and competence in computer programming among girls of middle-school age, especially those who are from disadvantaged backgrounds. This was a three-year project which involved a team of people from across different disciplines including computer design, teaching, graphics design, and many other.

One of the practical difficulties which is often put forward as a great impediment to conscientiously building-in ethical values in AI is the enormity of the variety of necessary skills required from study disciplines traditionally thought to be too far apart. The project team would have to be comprised of software developers, hardware engineers, philosophers, sociologists, graphic designers, etc. These skillsets have traditionally been thought to be unrelated and thus there is no precedence of people from these diverse walks of life working together in a coordinated fashion on design projects in the field of computer technology. The team on the Rapunsel project was construed to close this gap.

So, what was the Rapunsel project all about? Flanagan explains: “The project’s core hypothesis is that the marked absence of women in the field of technology development, particularly in computer programming, is due, at least in part, to the style in which basic scientific subjects are taught” (Flanagan et al., 2008, 332). Project Rapunsel sought to come up with a way of encouraging young girls of middle-school age to take interest in the field of computer programming. The main intension was to close the gap between the number of boys who take up the profession as compared to the number of girls.

The initiators of Rapunsel are convinced that the dearth of women in programming and the dominance of men in this discipline is due to the fact that the subjects that lead to careers in programming are taught in a manner that is appealing to boys and uninteresting for girl children at school level. The idea was that if a game, which is effective in teaching programming skills, can be made in such a way that it is enjoyable for girl children this would encourage them to take up the discipline. The computer game, Rapunsel, would be conscientiously designed in such a way that it would promote values of equal opportunity, good self-esteem, cooperation, etc. These ethical values were to be embedded into the computer artefact.

The way the game is played is that a player is assigned a home environment that houses a central character upon logon. The player must then ‘teach’ the character to move and dance by programming them in simplified Java. As the player’s programming proficiency improves, they are then able to teach the character more sophisticated moves and go into multiplayer mode where the characters can challenge each other in dance competitions. It was envisaged that girls would find this type of game exciting and interesting as opposed to the more traditional and widely distributed fighting games which tend to be more interesting for boys.

The Rapunsel design team thought it necessary to put in place feedback mechanisms for tracking if the project was achieving its goals on the various ethical values they had set out to embed into the game. For this purpose, they included within the project team

a group of schoolgirls whose role was to play the game and give regular feedbacks to the design teams through formal and informal channels. The design team would then factor in this feedback to make the necessary changes in the game and have it tested again by the girls to ensure that the end product would be fit for purpose. This 'feedback loop' method proved to be useful in that approaches that the designers of the game had initially assumed would be effective to progress towards the desired goal were sometimes found to be wanting. They found that they could optimise the game by incorporating the received feedback in subsequent iterations. This proved useful for Rapunsel.

I use the example of Rapunsel here to demonstrate how ethical values can be embedded conscientiously in AI by way of explicit functional definition of the system. The team leading the project was conscious of the values they intended to embed in the computer game and defined them ahead of the actual design work. They deliberately recruited into the team people from across different fields of expertise to ensure that all the necessary skillsets required to achieve the end goal are available within the project. In this way, they were able to develop mechanisms for embedding ethical values into the artefact. Also testers were recruited to play the game and provide feedback to the designers so that they could check if the ethical values they had set out to embed in the game were indeed promoted in the central uses of the artefact. Where the values were not clearly apparent or where unintended values were observed the design was revised for improvement. This feedback loop method is useful because quite often not all the scenarios can be anticipated in the functional definition phase already.

Another way in which values can be embedded in AI conscientiously is by being deliberate with collateral values. Mary Flanagan, Daniel Howe, and Helen Nissenbaum describe collateral values as values that "crop up as a function of the variety of detailed alternatives designers confront and from which they must select" (2008: 334). These 'cropping up' values are not explicitly stated in the functional definition of the computer artefact but get introduced by the practical constraints during design and manufacturing. Designers then have to navigate through these alternatives and decide on which of the

competing values should be prioritised in order for the desired ethical values to be effectively embedded in the artefact.

Collateral values are often linked to technical constraints such as feasible character lengths for the intended screen size or the number of keys that can fit on a keypad which can in turn fit into the overall design. These factors can result in limitation of the number of languages that the artefact can cater for, for example, introducing discrimination against certain sectors of society whose mother tongue is not accommodated in the end product. Such an artefact would likely be embedded with the disvalue of unfair discrimination on the basis of language. Designers' or stakeholders' own pre-existing biases, and economic discrimination due to the artefact being too expensive for other sectors of society, are just examples of some of the other factors that can influence ethical values embedded into AI.

I have shown how values are embedded in AI, and how those involved in creating computer artefacts can take responsibility for being both conscientious and accountable of the values they build into the machines and software they make.

### 3.3. Objections to the embedded values approach

The embedded values approach to computer ethics is premised on the view that computer artefacts have built-in values, and that these values are self-evident in the central uses of the artefacts. By this view, the designers of computer artefacts, along with all other stakeholders involved, are responsible for embedding values into the artefacts, consciously or unconsciously. The embedded values approach therefore suggests that designers have ample opportunity to influence which values are ultimately embedded into AI.

There are, however, some objections to the view that values which can be analysed independently of users can be built into artefacts. One such objection is based on the idea of the neutrality of the artefacts. Proponents of the neutrality approach insist that, as mere tools in the hands of a moral agent, computer artefacts cannot have built-in

values which can be analysed independently of users. This view denies that consequences emerging from the use of any artefact are a function of the system's characteristics.

Interestingly, the neutralist view does not deny that technical artefacts have had significant effects on social aspects of life. It also does not deny that computer artefacts have contributed to the promotion of certain social, moral, and political values over others. It denies only that these consequences are a function of system characteristics. As Flanagan et al puts it, the neutrality view only "holds [that the consequences] are a function of the uses to which [the artefacts] are put. Technologies are mere tools of human intention; morality (good or evil) and politics inhere in people and not in their tools" (2008: 349). The neutrality view seems to rely on the presence of a person behind the artefact, which may be quite easy to follow for simple hand-held tools, but it is difficult to see how the same view would fare in the face of strong AI - which typically has no human being driving and instructing the 'tool'. While I agree with the thinking that AI, like all other computer artefacts, are tools, I hold that tools can promote ethical values which are embedded in the tools themselves and which can be analysed independently of their users.

It is my view that a valid moral value judgement of spyware, for example - as being a morally bad piece of software - can be made on this computer artefact without one sounding silly. Computer artefacts are not the only objects on which moral judgements can be passed. "Drugs are bad" is a common phrase that is often taken to be complete without any real insistence on a moral valuation of the user. In fact, it is quite often the case in human trafficking and forced prostitution situations that the user is not taken to be blameworthy (the users are usually forced to take the drugs by the human trafficker).

Rather, the drug manufacturer is likely to take much blame for the impact of the drug. The assertion by those objecting to the embedded values approach to computer ethics, that artefacts themselves are value neutral, does not seem to satisfy the question of values in drugs. It would be almost preposterous to suggest that drugs themselves, as

artefacts, are morally neutral and not embedded with any values; and that the drug manufacturer too carries no responsibility - only the drug users' use of drugs is morally bad. In this example, it seems fair to say that 'drugs are bad', and that this is a reflection of the attitude of the drug manufacturer who is most probably driven by greed and concerned only about profiteering in the face of the many ills drugs have been known to promote in societies. Just as 'spyware is bad', and this is a reflection of the attitude of the software developer that made it. Notice here that both statements are valid independent of the users.

I have discussed in an earlier section of this chapter how even some (advanced) weak AI artefacts are already capable of acting autonomously without direct human instruction. I have also shown how strong AI is conceptually possible and explained how strong AI is taken to be any AI on par with human-level of intelligence and above. At that level of intelligence, it is envisaged that the machine will be able to repair itself, design more sophisticated versions of itself, and even have the capability to plan ahead. Such a machine surely can be evaluated ethically as good or bad based on the consequences of its central uses. I do not wish to argue that it is impossible for a computer artefact to be morally neutral. I am however asserting that, generally, various ethical values are embedded in AI, consciously or unconsciously, and that technology designers should be conscientious of the values they build into the artefacts.

Another objection to the embedded values approach, perhaps one more practical in nature, is general skepticism concerning the likelihood of designers succeeding in shaping technology and its outcomes. These skeptics consider that the complexity associated with successfully designing and manufacturing technology artefacts, the number of stakeholders that would be involved in the process, and the balancing act of all the technology and practicality constrains involved, make it nearly impossible for even the best-intentioned and best-informed designers (among the many others with power to shape a system) to build-in values in AI. For them this objection, which Flanagan et al describe as a "general dilution of agency" (2008: 349), is further

compounded by the unavoidable problem of unintended consequences, which is common in the design of new and emerging technologies.

It is correct, as demonstrated in the Rapunsel project, that setting up the necessary team, structures and processes to create a computer artefact which is conscientiously embedded with ethical values is not an easy task. But the success of the project is proof that it would be a distortion to consider the task impossible. A large number of individuals, which included participants from a wide variety of disciplines, ordinary members of the community, testers, etc., were involved in the programme which became a success three years after inception. The values embedded in Rapunsel included values of friendship, cooperation, knowledge, equality, good self-esteem, etc., while disvalues of patronising and biased generalisation were eliminated. This shows that, when the right resources are made available and the will is present, the ethical values can be embedded in AI deliberately and successfully.

Two things are clear from this exercise. The first is that computer artefacts have built-in values which can be analysed independent of their users and the second is that it is possible to conscientiously embed ethical values into AI.

## 4. Ethical Theory

What is morally permissible or impermissible, and what is morally obligatory is the subject of much debate and disagreement even for societies with cultures so apparently similar that outsiders would find it hard to distinguish between them. Life often presents us with situations where a moral agent has to choose the most appropriate action among a number of competing options. Ethical theories are the tools which can be employed as guides to solving these ethical conundrums. The different theories emphasize different values, such as justice, fairness, autonomy, happiness, etc.

In this section, I shall study a few dominant ethical theories and attempt to justify why Ubuntu is the most suitable approach to ethical values to build into AI. This is important to consider, given the context of rapidly developing computer artefacts which are capable of autonomously performing acts which have moral consequences.

### 4.1. Consequentialism

A key characteristic feature of consequentialism, which also distinguishes it from other ethical theories, is its insistence that the outcome or consequence of the action is the only determining factor for the rightness or wrongness of moral action. While other ethical theories may consider the relationship between the moral agent and the moral patient, the intention of the moral agent, or some other rule in determining if the act is morally blameworthy or praiseworthy, according to William Shaw "[s]tandard consequentialism asserts that the morally right action for an agent to perform is the action, of those that the agent could perform at the time, that has the best consequences or results in the most good" (Shaw, 2006:5).

Consequentialism is therefore a value maximizing doctrine that considers it obligatory for moral agents to perform the act that brings about the most good of all the options available to the agent. If none of the actions available to the moral agent at the time have good consequences then the morally right thing to do for the agent, according to consequentialism, is the option that leads to the least bad. That is to say, if all other

alternatives will produce worse outcomes, then the morally right thing to do is the option that produces the least amount of bad – but still bad. Where several actions are equally morally right, the agent is free to choose any one of the available actions among the equally right ones.

Shaw makes it clear that he is talking about consequentialism in its standard form, which he takes to be the "the most widely discussed form of consequentialism" and which he thinks "[…] is the most plausible form of consequentialism" (Shaw, 2006: 6). This is necessary because although consequentialists are agreed on the value maximizing doctrine, they do tend to differ on which states of affairs are intrinsically valuable – also sometimes referred to as "the Good". For their part, Alexandra, Larry and Moore explain that "[s]ome consequentialists are monists about the Good", while yet still "[o]ther consequentialists are pluralists regarding the Good" (Alexander et al, 2016: 1). Utilitarians, for example, are monists who take the Good to be welfare, pleasure, or happiness; whereas pluralists insist that equitable distribution of the Good (among people) is itself partly constitutive of the Good. Moore further adds that "[h]owever much consequentialists differ about what the Good consists in, they all agree that the morally right choices are those that increase (either directly or indirectly) the Good" (2016: 1). Common across all forms and variations of consequentialism is the basic tenet of maximizing the Good while minimizing the bad.

Another noteworthy feature of standard consequentialism is that it is an agent neutral ethical theory. Douglas Portmore explains the notion of agent neutrality thus: "A theory is agent neutral if it gives every agent the exact same set of aims" (Portmore, 2001: 364). Standard consequentialism is therefore an agent-neutral moral theory as it requires of all moral agents that they act in such a manner as to conduce to the best state of affairs, judged from an impersonal point of view. For the consequentialist, the rightness or wrongness of an act depends on the outcome the act itself produces - it does not matter who is behind the act itself. The theory places a duty on moral agents to devote themselves to the course of bringing about the best state of affairs possible under any given set of circumstances and without consideration of their relatedness to

the moral patient. David Brink characterises this idea of impartiality on the part of the moral agent as one "that asks an agent to transcend his own private concerns and allegiances" (Brink, 2006: 394).

Shaw further notes that, according to standard consequentialism, it is the value of the consequence as estimated (and intended) by the moral agent at the time of performing the act that counts, not the actual consequence considered in retrospect. Consequentialists concede that, by chance or some other attenuating circumstance, the actual outcome produced by the act with the most expected good may turn out to be less than what was reasonably anticipated. Indeed, the result may very well be worse than what the other options which were available to the moral agent could have produced. When this happens, standard consequentialism holds that the action which was expected to yield the most good remains the morally correct thing to do – even though it was not the act that ultimately produced the most good.

It is therefore the value of the expected consequence that counts, not the actual value produced. Shaw adds, "[a]ssuming that the agent's original estimate of expected value is correct (or, at least, the most accurate estimate one could have arrived at in the circumstances), then this action remains the right thing to have done" (Shaw, 2006: 8). This is so because, for the future and outside of the uncommon condition that caused the anomaly in the consequence of the chosen action, the moral agent should know to choose the same option should a similar scenario arise in another instance.

### 4.2. Deontology

Unlike consequentialism, which is an agent-neutral ethical theory, the deontological approach to ethics can be agent-centered or patient-centered. As for the agent-centered vantage, Alexander et al asserts that "[a]ccording to agent-centered theories, we each

have both permissions and obligations that give us agent-relative reasons for action" (Alexander et al, 2016: 3). Examples of such agent-relative obligations include obligations parents have towards their own children, an attorney towards his or her client, a man or woman towards their own spouse, etc. It can be said about a parent that they have certain moral obligations towards their own children which they do not necessarily have towards other children; so much that such a parent cannot be reasonably blamed for saving their own child from drowning even if the other alternative course of action available for them at the time was to let their own child drown in order to save three other children (whom they do not know). The basis of the agent-centered reasoning is centered around the obligations and permissions of the agent.

As for the patient-centered approach, the basis of the reasoning for what is obligatory and what is permissible is based on the rights of the patient (person on the receiving end of the action). According to Alexander et al, "[a]ll patient-centered deontological theories are properly characterized as theories premised on people's rights" (Alexander et al, 2016: 18). Both these approaches can best be understood through Kantianism.

Immanuel Kant sought to find what he called the 'Supreme Principle of Morality', (Kant, 1993: 5, quoted in McNaughton and Rawling, 2007: 37) which dictates that an agent's deliberate action is a reflection of the agent's maxim, and the permissibility of the action depends on the universalizability of that maxim. A maxim is the underlying principle that guides the decision-making process which precedes an act. It is the honest justification or reasoning that a person would advance in the face of a serious enquiry why they take a particular course of action. Maxims are, essentially, the 'whys' behind every conscious action that connect one's behaviour in the world with the set of personal 'rules' that one lives by. Every act is inspired by a maxim, except, of course, acts which can be taken as involuntary or unconscious. Shafer-Landau points out that a maxim "states what you are about to do, and why you are about to do it" (Shafer-Landau, 2012: 157). According to Kant, unless our maxims are universalizable – unless they can be applied to everyone consistently and without contradiction, in every context – they must be rejected. This he

expressed in his first formulation of the moral principle he called the Categorical Imperative.

The Categorical Imperative applies to rational "sane adult human beings" (Hill, 1992: 202). For Kant, morality is based on the ability to think rationally and, therefore, to conduct oneself in an immoral manner is tantamount to acting illogically. Put differently, our ability to form intentions based on logically sound reasons renders us responsible for our actions. Kant maintained that "we must be able to determine what is right and wrong by rational thinking alone" (Shafer-Landau, 2012: 174) – without taking into account emotions.

Kant's Categorical Imperative has two essential formulations: the "formula of universal law" (to use O'Neill's (1993) preferred label) and the "formula of the end-in-itself" (as per McNaughton and Rawling (2007)). The formula of universal law holds that rational beings should be consistent and show "non-contradiction" (McNaughton and Rawling, 2007: 36) in the way that they make moral choices. This means that the morally correct act is one that a moral agent would be happy to see being condoned or required consistently for everyone. As for the second formula of the end-in-itself, Kant argues that our ability to reason makes us "priceless" with "special moral status" (Shafer-Landau, 2012: 190). This specialness demands respect for the dignity of all human beings and comes with the duty to never use another person as a mere tool (means) for one's private objectives (ends).

### 4.3. Virtue Ethics

Rosalind Hursthouse (2013) describes the three concepts that are essential to a virtue-ethical approach to right action. They are virtue, practical wisdom and eudaimonia.

Firstly, "virtue," according to Hursthouse, is more than and deeper than a "tendency to do what is honest or generous" (Hursthouse, 2013: 3). Rather, a virtue is a character trait that "goes all the way down," and so manifests in a "multitrack" disposition

(Hursthouse, 2013: 3). One who possesses a virtue is imbued with it, so much that it permeates across their being and is expressed in their views, feelings and actions. Virtue ethicists acknowledge that attainment of perfect virtue is unlikely, and that "possessing a virtue is a matter of degree" (Hursthouse, 2013: 4).

Secondly, Hursthouse describes 'practical wisdom' as the connection between the good intentions of a virtuous person with the real world. A virtuous character is, therefore, one that possesses practical wisdom: 'phronesis' – which is acquired through training, first, by one's parents and care-givers in childhood and then through life experiences. She argues that one who is practically wise will know how to do the right thing "in any given situation" (Hursthouse, 2013: 7), because they have the maturity and experience to distinguish its "morally salient features" (Hursthouse, 2013: 8). This is unlike "nice children" or "nice adolescents" (Hursthouse, 2013: 7), who may do the right thing by fortunate accident, rather as an expression of virtue. The virtuous adult applies practical wisdom almost like an art, where they have fine-tuned "situational appreciation" (Hursthouse, 2013: 8) and know the best way to actualise their virtuous intentions. Context is key in virtue ethics – there is no one-size-fits-all decision.

Finally, Hursthouse suggests that there are a number of approaches to understanding 'eudaimonia', a Greek word that does not translate easily into English. Eudaimonia can be described as flourishing, possessing qualities that make a person a wholesome moral agent, who taps into their own phronesis to consider all possibilities and courses of action available to them, correctly understand the impact of each on the state of affairs and other people, and makes the correct morally correct decision – without needing the dictates of rules. Through cultivation of the virtuous character, one would ultimately achieve a state of eudaimonia – which means one is at a level of an exemplary moral agent.

For a close English equivalent to the eudaimonia, she favours the word 'flourishing' over 'happiness' because the latter concept is far more vulnerable to subjective interpretations (Hursthouse, 2013: 8). Eudaimonia is the peak of living, achieved by

“those who understand what is truly worthwhile, truly important and thereby truly advantageous in life, who know, in short, how to live well” (Hursthouse, 2013: 8). The sort of people who can do so are the “practically wise” (Hursthouse, 2013: 8), who have mastered the art of actualising their virtues in the mess of real life.

A virtuous character is consistent in maintaining virtuous conduct, abhors morally poor behaviour and is not pleased with ill-considered moral judgment. It is a character that ’goes all the way down’ and more than just a habitual good-doer. Someone who has attained eudaimonia does not struggle with making the morally correct decision, even when there is no danger of being caught, but has an intense and permanent desire to do good always. This is simply who they are.

According to Hursthouse, “the good life is the eudaimon life, and the virtues are what enable a human being to be eudaimon because the virtues just are those character traits that benefit their possessor in that way, barring bad luck. So there is a link between eudaimonia and what confers virtue status on a character trait” (Hursthouse, 2013:10). For the virtue ethicist, the moral question is how I ought be, rather than the act-based moral consideration that asks what ought I to do.

Virtue ethics offers the conceptual resources for a richer account of ethics and can explain ethical motivations across a range of situations.

### 4.4. Ubuntu Ethics

Like eudaimonia, the word ‘ubuntu’ does not translate easily into English. For this reason, a discussion (in the English language) on ubuntu – and ethical values of ubuntu – as conceptualized by the Bantu-speaking peoples across much of the Sub-Saharan portion of the continent can become quite wordy. Ubuntu is more than just an ethical theory. It is the essence of a sentient and progressively flourishing human being. For aBantu (people) it is intertwined with their very identity, both as autonomous individuals and as a collective. According to Mogobe Ramose (1999), philosophy, medicine,

religion, law, politics, ecology and globalisation can all find expression through ubuntu. It would be a grave mistake to reduce ubuntu to any single one of its various expressions, but each aspect must be considered with the whole in mind. For the purpose of this essay, I shall focus on ubuntu as a moral philosophy.

In some ways, ubuntu can described as the lived experience of each umuntu individually and abantu aggregately. To understand the concepts of ubuntu, proficiency in any of the bantu languages is key as much of the clues are locked up in the expressions of the culture and interpersonal interactions. The suggestion here is not to the exclusion of all other non-Bantu communities but an extension of an invitation to facilitate an enriched interaction. Metz and Gaie correctly make the observation that "ubuntu/botho does accord all human beings a moral status and considers everyone in principle to be potential members of an ideal family based on loving or friendly relationships" (Metz and Gaie, 2010: 281). I suggest also that the warmth and the generosity of aBantu/Batho is commonly evident in this regard.

That said, perhaps it is useful to get into lexical matters for some key words which are building blocks for the word 'ubuntu'. Various dialects of the Bantu languages spoken in parts of the African continent, from Senegal in the west coast to the Zanzibar islands in the east, and from the Nubian region in the north-east all the way to the Cape Point, have a different word for the same concept. In the Sotho family of languages it is botho, in Shona it is hunhu, in the Nguni family of languages it is ubuntu, etc. For the sake of simplicity and consistency, I shall restrict my usage of the words only to their Nguni and Sotho variations as follows:

| | |
|---|---|
| umuntu/motho: | human (being) |
| abantu/batho: | human (beings) |
| ubuntu/botho: | humanness. |

I take humanness to be the closest approximation to the English equivalent of the word ubuntu. Ramose argues that, of the multiple translations of the word ubuntu, it would be incorrect to include humanism in the list, especially if humanism is understood as a

specific notion in Western philosophy. He argues that “any -ism, including humanism, tends to suggest a condition of finality, a closedness or a kind of absolute either incapable of or resistant to any further movement” (Ramose, 2009: 69). This he takes to be contrary to the essence of humanness, which he insists “suggests both a condition of being and the state of becoming, of openness or ceaseless unfolding” (Ramose 2009: 69). From this perspective, Ramose argues that “humanness regards being […] as a complex wholeness involving the multi-layered and incessant interaction of all entities” (2009: 69).

The aphorism motho ke motho ka batho (umuntu ngumuntu ngabantu) is common in nearly all the indigenous languages on the continent of Africa. Moeketsi Letseka directly translates it as “a human being is a human being because of other human beings” (Letseka, 2012). Samuel Mbithi puts it thus: “I am, because we are; and since we are, therefore I am” (Mbithi, 1971). To affirm one’s humanity therefore, one has to first recognise the humanity of others and establish humane relations with them on the basis of their shared humanity. By the aphorism motho ke motho ka batho, this affirmation of self, which culminates in the establishment of humane relations, is the first step to botho (ubuntu). For his part, Ramose explains “botho [is] understood as being human (humanness/humanity) and [having] a humane (respectful and polite) attitude towards other human beings which constitute the core or central meaning of the aphorism; motho ke motho ka batho.” He adds further “[n]either the single individual nor the community can define and pursue their respective purposes without recognising their mutual boundedness; their complementarity” (2009: 70).

Ubuntu, therefore, entails recognition of the dignity of others and their projects, and proceeds to treat them with dignity. Put differently, to deny the dignity of others is to deny your own. Ramose suggests that “[t]o care for one another, therefore, implies caring for physical nature as well” (2009: 71), and further adds “wholeness applies also to the relation between human beings and physical or objective nature” (2009: 71). People are not at the center of the universe and should therefore respect nature and the harmony of a balanced ecosystem. The notion “I am, because we are; and since we

are, therefore I am" properly contextualises the position of umuntu (human being), in ubuntu, within the jigsaw puzzle called the world, and in the context of the (whole) universe. By this thinking, in ubuntu respect is extendable to states of affairs, objects and projects, nature, posterity, the living dead, memory, harmony, etc. The demand on umuntu is to always uphold and never compromise, to the best of their ability, the values of ubuntu.

The word ubuntu consists of a prefix ubu- and a stem ntu- where, according to Ramose, "ubu- evokes the idea of being in general. […] ubu as an enfolded be-ing is always oriented towards unfoldment […] towards -ntu" (Ramose,1999:36). "Ubu-ntu [humanness] then not only describes the condition of be-ing, insofar as it is indissolubly linked to umuntu [human being] but it is also the recognition of [the] be-ing becoming" (Ramose,1999:37). So the demand for one to conduct themselves in a manner that is consistent to ubuntu is not polar, where one is either completely perfect and therefore acceptable, or imperfect and therefore relegated to a lesser status, but emphasises the determination to keep moving along on the path of excellence – forging ahead to becoming a better human being.

For his part, Thaddeus Metz undertook a project to formulate a theory of an African ethic and came up with the below lists of "competing theoretical interpretations of ubuntu" (2007, 328), which I find helpful for this discussion:

- An action is right just insofar as it respects a person's dignity; an act is wrong to the extent that it degrades humanity.
- An action is right just insofar as it promotes the well-being of others; an act is wrong to the extent that it fails to enhance the welfare of one's fellows.
- An action is right just insofar as it promotes the well-being of others without violating their rights; an act is wrong to the extent that it either violates rights or fails to enhance the welfare of one's fellows without violating rights.
- An action is right just insofar as it positively relates to others and thereby realizes oneself; an act is wrong to the extent that it does not perfect one's valuable nature as a social being.

- An action is right just insofar as it is in solidarity with groups whose survival is threatened; an act is wrong to the extent that it fails to support a vulnerable community.
- An action is right just insofar as it produces harmony and reduces discord; an act is wrong to the extent that it fails to develop community.

Metz identifies interpretation 6 as listed above to be the one he takes to be correct. He concludes about ubuntu that "the most justified normative theory of right action that has an African pedigree is the requirement to produce harmony and to reduce discord, where harmony is a matter of identity and solidarity" (2007: 240). I do not agree that ubuntu is a theory of right action necessarily, for sometimes an apparently right action can attract criticism or even blame based on the spirit with which it is done. An act which produces less than the best state of affairs in consequentialist terms or which does not satisfy one deontological rule or another can find condonation if it 'came from a good place' as an honest mistake or if it was a failed attempt to uphold another more fundamental value. Ubuntu moral philosophy is not overbearing on the moral agent. While an immoral act attracts criticism, the approach is one that is leaning towards correction to inspire an improved handling of a similar situation in the future. Forgiveness is an important value.

The phrase 'spirit of ubuntu' often comes up in ubuntu moral deliberations. When faced with a moral dilemma, both the questions 'how ought I to be' and 'what ought I do' can be answered by referring to the "spirit of ubuntu", which finds application in every morally relevant decision point.

I shall not go into much discussion about the first five of Metz's six interpretations as, by his own admission, he finds only the sixth one to be "the most promising theoretical formulation of an African ethic to be found in the literature" (2007: 334). I take this attempt by Metz to confine ubuntu to an act based moral theory which sums up right action as on that "produces harmony and reduces discord" cannot be correct because throughout history Bantu people have taken it to be their moral responsibility to engage

in wars for the sake of freedom and justice. Indeed, even in South Africa as recently as in the era of Apartheid ubuntu was at the heart of activism against injustice. Abantu consider holding one's figurative peace in the face of an injustice deplorable. Ubuntu can therefore be taken to be an activist ethic.

Metz's interpretations of ubuntu do not quite capture ubuntu in its essence, for moral right and wrong pivot on the spirit or attitude that is the motivation behind the act. The act itself is not what matters chief most. Useful, however, is that each one of his 'theoretical interpretations' captures one ubuntu value or another, e.g. respect, human dignity, generosity, harmony, etc.

Justice Mokgoro, who was part of the bench in the apex court in South Africa, argued that because ubuntu "envelopes" values of compassion, respect and human dignity, it can be thought of as "humanistic orientation towards fellow beings" (Yvonne Mokgoro, 1998: 3). Here, the learned judge hints at some of the values included in ubuntu ethical values. To these, we can add respect for life, human dignity, justice, fairness, brotherhood, honesty, and many other ubuntu values. These are some of the ethical values of ubuntu, and the purpose of this essay is to argue that they should be embedded in AI.

In the next chapter I will argue that ubuntu is the most suitable ethical approach to embed in value-able AI. I will point to its value of inclusivity and its emphasis on respect and dignity, which extends also to non-sentient entities. I will also argue that letsema, a key feature of ubuntu, is ideal for solving the questions around how the extended AI teams could be fairly compensated on start-up projects with unknown future values.

## 5. Ethical Values of Ubuntu Embedded in AI

AI has grown quite exponentially in the last few years and it continue to do so at an alarming rate. While much has been achieved, it has all been within a short space of time so much that computer ethics as a field of study is yet to catch up with the development. Traditional definitions of values such as privacy, for example, are yet to have a common definition in the cyber space partly because philosophers and sociologists are not intricately involved in the design of computer artefacts. This leads to a situation where matters are dealt with as they arise and find their way into the mainstream public discourse; to keep up with the fast pace of technology as society "we need to repair our boat at sea and cannot take it out of the stormy waters to study it for an indefinite period of time" (van den Hoven's, 2008: 303), to use Jeroen van den Hoven's words.

As we have understood from Ramose in the previous chapter, ubuntu philosophy finds expression in many other fields such as law, anthropology, theology, and medicine, so its stark absence from the field of computer ethics and AI in particular is hard to ignore, especially because ubuntu moral philosophy seems to be the most suitable answer to the question of 'repairing boats at sea'. In this final chapter, I argue that ubuntu is the best ethical theory for this task because consultation and collaboration, the two key requirements for solving the emerging computer ethics questions, are important values in ubuntu.

The demand for human dignity and the principle that everyone is a "potential member of an ideal family based on loving or friendly relationships" (Metz and Gaie, 2010: 281) supports my claim that consultation is an important value in ubuntu and the concept of *letsema* in ubuntu supports my claim for collaboration. Letsema can be translated into the English language as 'communal collaboration' but "often incorrectly translated as 'work parties' in South Africa" (Ramose, 2007: 350).

Every member of the village, according to age group and physical abilities, has a moral obligation to participate in letsema according to ubuntu: women gathered together to cultivate the land for the elderly and the disabled, young men come together to dig a grave when there is death, older men club together to build a house for the newly-wed couple, etc., are some examples of letsema in practice. Collaboration between teams from diverse fields of expertise as well as testers, community members, financiers, and other stakeholders in a similar fashion to the Rapunsel project can be taken to be akin to an AI letsema.

Flanagan et al. described the Rapunsel project as “a large multidisciplinary collaboration” with the aim of designing and implementing an experimental game prototype (2008:331). The main aim of the game was to promote interest and competence in computer programming among girls of middle-school age. To successfully make a game that would be appealing to its target market and also fulfil its educational objectives, the design team comprised of software developers, hardware engineers, philosophers, sociologists, graphic designers, testers (who were girls of middle-school age), etc. in a manner akin to *letsema*. There were no big monetary rewards given for participation, but the participants understood the disvalue of not having enough females in the discipline of programming and so were volunteering their expertise to the cause - a motivation reminiscent of *letsema*. The initiators of Rapunsel admitted to the difficulty of pulling together such a multi-disciplinary team and getting volunteers to remain committed until the end of the project.

While Rapunsel was not strictly for profit purposes, most other AI projects would have investors funding the work with an expectation of realising a return. Designing value-able AI which is embedded with ethical values of ubuntu would require significantly larger teams from various fields of expertise, potentially prolonging project time lines and increasing the wage bill. This could be prohibitive for lower budget projects in the technology sector and even create the industry to be dominated by fewer big corporations. Also, for a start-up technology company, how would the extended teams be compensated fairly when it is not known for certain yet if the new AI they are working

on would be the next million-dollar product on the market in the future or another failed innovation? I propose that ubuntu's *letsema* model would be best suited as the participants would be driven by a motivation other than financial gain, for in a *letsema* the participants are motivated by moral reasons. This would significantly ease the financial cost of value-able AI design.

Another reason ubuntu is suitable to be embedded in AI is that the moral theory has a wide purview: it takes a wide range of things to be morally relevant, such as states of affairs, objects and projects, nature, posterity, the living dead, etc. Thus, conscientiously embedding ubuntu in AI will not only help ensure that the artefacts do not destroy humans but it will also create computer artefacts that are valuable to the entire environment too. In arguing for an African ethic that says "land is so sacred that it is seen as the guardian of morality", Workineh Kelbessa further explains it with an ancient African myth which compares the earth to an egg: that "if we press the egg hard, we break it, or if we do not hold it securely enough, it drops and breaks. Likewise, we are required to be very careful in our dealing with the seemingly fragile frame of the world" (Kebessa, 2014: 41). Now more than ever and in the face of ever rising global temperatures due to greenhouse gases emitted into the atmosphere, ubuntu ethical values embedded in AI can help reduce the carbon footprint of many industries in which AI is used.

We have also discussed in the last chapter how ubuntu ethics can be activist in nature. Such values are important in the relatively young field of AI where some designers, consciously or through ignorance, are taking advantage of the inadequacy of regulations to produce artefacts that promote disvalues. The activist inclination of ubuntu ethics will help ensure that the participants in the design team voice out the potential disvalues they detect in the early stages of the project and before mass production while the harmony elements of ubuntu will promote an agreeable spirit on matters where there are no viable technical solutions.

There are some objections which are levelled against other western ethical theories that ubuntu seems to do better at answering. For example, consequentialism is often said to be too demanding and counter-intuitive, while in deontology there seems to be a conflict between some duties and rights. Those who raise the issue of the demanding nature of consequentialism point to its value maximizing doctrine, which states that the only right action is the one which yields the most good without consideration for one's own projects. It is therefore counter-intuitive in that it is agent neutral and does not allow for natural agent specific considerations such a parent's duty towards his or her own child.

As for deontology, one of the main objections is one which Alexander et al call the 'control theory of agency'. This notion suggests that a moral agent may be exempted from their obligation if (they believe that) their action would not have a meaningful impact on the resultant state of affairs. Alexander et al put it thus: "our agency is invoked whenever our choices could have made a difference" (2016: 18). This seems to suggest that, on the question of conscientiously embedding ethical values in AI, a deontologist's attitude is one that would encourage inaction unless there was evidence that enough AI designers would be doing the same in their own artefacts. In other words, why bother if your efforts will not make much difference in the grand scheme of things? This is hardly a progressive approach for breaking ground in ensuring a future for an embedded values in AI.

We have discussed in chapter four how Ramose characterises ubuntu, interpreted as humanness, as suggesting "both a condition of being and the state of becoming, of openness or ceaseless unfolding" (Ramose 2009: 69). Ubuntu therefore seems naturally suited for application in the new and unfolding world of AI, in which the moral agent does not have the heavy burden of a demanding ethic, but has a duty to try their best from a well-meaning place and having the opportunity learn and grow as technology develops. In ubuntu ethics, the contribution of both the project initiator who dares to lead the way and be the first to form a project team akin to *letsema* and

individual who volunteers their skills without preoccupation about financial benefits are respected and welcome.

## 6. Conclusion

Can ethical values of ubuntu be embedded in AI? Yes, they can and I suggest they must. Artificial intelligence in computer artefacts which is commonly available even in children's toys these days is impressive, especially considering where the technology was just a few decades ago. By way of comparison, it is often said the average smart phone today has over 100,000 times the processing power of the computer that landed man on the moon 50 years ago The exponential growth in the cognitive abilities of machines is cause for much discussion about whether one day, when they are way too 'smart' for us to predict reliably or even understand with precision what action they will perform next, they will turn on us and destroy the human race.

I have maintained that AI, and computer technology in general, achieved the levels of sophistication we see today as a result of our need to keep improving the quality of our tools to make our lives easier. This is due to the reliance people have developed on tools and implements to perform both mundane and complex tasks. This dependency has driven a continuous quest for more efficient but reliable tools. Advancements in technology have reached a point where AI is capable of autonomously undertaking tasks that could previously only be fulfilled by human beings. Modern-day machines can perform activities that have outcomes which are morally relevant.

Value judgements can be passed on computer artefacts independently of their users by assessing the moral values of the outcomes and states of affairs they tend to bring about. Put differently, these value judgements can be passed on the computer artefacts based on their built-in values. These built-in or embedded values tend to promote or demote the realisation of particular moral values in the artefact's central uses. By central uses I mean those uses that the artefact was designed for and which are also the justification for the computer artefact's continued production. Sometimes it is easy to tell if a computer artefact is morally good or bad while it may not be so straightforward in other cases. Spyware is clearly bad as it undermines privacy and promotes fraud, for example. The disvalue in facial recognition AI, on the other hand, is not so easy to pick

up unless you are of African descent with a dark skin – in which case you would be on the receiving end of the disvalue. In both cases, the disvalues come to the fore in the central uses of the artefact.

For this reason I argue that stakeholders in the computer technology as well as in the information and communications technology industries should put out all stops to ensure that teams involved in AI design and development are assembled in a manner which allows them access to the diverse skills-set and cultural backgrounds necessary to build value-able AI which is embedded with ethical values of ubuntu. I have demonstrated how computer artefacts commonly available today are already laden with values. Budget constraints, designers' own unconscious biases, technological constraints, lack of exposure to other cultures, lack of diversity within design teams, etc., are all examples of the factors with influence the values unconsciously embedded in AI today. A more conscientiously assembled AI design team could mitigate the extent to which disvalues are imbedded unconsciously in the artefacts.

Difficulties that can be reasonably expected in any attempt to apply traditional western ethical theories for embedding moral values in AI seem to be acute without obvious and sustainable mitigation. In the last chapter, I pointed out the demoralising attitude of the deontologist which seems to discourage individual efforts to conscientiously embed ethical values in AI - unless a critical number of AI designers adopted the embedded values approach. It can be expected that it could take some time and dedication from motivated AI designers before a significant number of investors would 'buy into the idea' an adopt an approach that does not obviously fast-track production. For this reason, I think deontology is not a promising ethical theory for the purpose of application in AI.

Similarly, consequentialism is too demanding on moral agents, especially because its agent-neutrality feature would conceivably wear out participants in the projects. Explaining the concept of agent-neutrality, Douglas Portmore say: "A theory is agent neutral if it gives every agent the exact same set of aims" (Portmore, 2001: 364). This moral theory also seems to be impractical for application in AI design as, according to

consequentialist thought, the only right action is the action that is expected to produce maximum good, not the one that actually does produce maximum good. This concept may come across as counter-intuitive for engineers involved in innovation as they are often experimenting with new ideas and combinations of ideas, often without a good understanding of what the outcome might be.

Ubuntu is most suited for the purpose of embedding ethical values into AI. This is because ubuntu ethic embraces a wide range of values which include reverence for nature, for the unseen world, for life, and for many other inanimate objects. As an ethical theory, ubuntu can be described as comprehensive and all-encompassing in this way. In a world where technology has been at the epicentre of how mankind is potentially destroying the planet through greenhouse emission and deforestation, it is crucial to embed values of ubuntu in AI – which is increasingly central to how we design new tools and machinery.

*Letsema*, which is a feature of ubuntu, presents an opportunity to successfully and sustainably assemble design teams for development of value-able AI. Traditionally, the skills-sets required for such projects are not readily available within the computer technology sector and recruitment for such skills through conventional means could prove to be prohibitive. However, ubuntu's *letsema* approach could open up access to individuals who are not in it for money.

Ubuntu is also 'an activist' ethic which requires of abantu to never remain silent in the face of an injustice while promotion of harmony is also an important value. This would facilitate for a robust but agreeable forum in the *letsema* design teams. Ubuntu ethic is adaptable to new conditions as it is "both a condition of being and the state of becoming, of openness or ceaseless unfolding" (Ramose 2009: 69). This makes it most suitable for application in a field that is rapidly changing and moving into a future we can only speculate about.

Stephen Hawking raised the alarm when he said "[t]he development of full artificial intelligence could spell the end of the human race." He further added, "[w]e cannot quite know what will happen if a machine exceeds our own intelligence, so we can't know if we'll be infinitely helped by it, or ignored by it and sidelined, or conceivably destroyed by it" (Cellan-Jones: 2014. para. 11). I think Hawking's view is only true to the extend that we sit back and wait for the future to come. I, however, am inclined to go with a more proactive approach of deliberately embedding ethical values of ubuntu in AI. With this approach Bantu Biko's 'prophecy' may be realised, which he considered "the special contribution to the world by Africa", which is "giving the world a more human face" (Maluleke, 2008: 6).

**References**

Allen, C., (2010). 'Artificial life, artificial agents, virtual realities: technologies of autonomous agency', in Information and Computer Ethics, Cambridge University Press, Cambridge. Ed.

Allen, C., Wallach, W., and Smit, I., (2006). Why Machine Ethics? IEE Intelligent Systems, vol. 21, no. 4, July/August, pp12-17.

Alexander, Larry and Moore, Michael, (2016) "Deontological Ethics", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/>.

Brey, Philip (2010). "Values in Technology and Disclosive Computer Ethics". The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press, pp. 41-58.

Brink, David O. (2006). "Some Forms and Limits of Consequentialism." The Oxford Handbook of Ethical Theory.

Bringsjord S., Bello, P., and Ferrucci, D., (2001). Creativity, the Turing Test, and the (Better) Lovelace Test. Minds and Machines, 11, pp3-27.

Carroll, Sean (2015). Stone Tool Technology of Our Human Ancestors. HHMI BioInteractive [online video] Available at: https://www.youtube.com/watch?v=L87Wdt044b0 [Accessed 11 August 2019].

Cellan-Jones, R. (2014, December 2). Stephan Hawking Warns Artificial Intelligence Could End Humankind. *BBCNews.com*. Retrieved from https://www.bbc.com/news/technology-30290540

Díaz-Kommonen, L. et al (2004) "Expressive artifacts and artifacts of expression". Working Papers in Art and Design 3. Pp

Etieyibo, E. (2017). "Anthropocentrism, African metaphysical worldview, and Animal practices: A reply to Kai horsthemke". Journal of Animal Ethics. University of Illinois Press, pp145-162

Floridi, Luciano (2010). "Information Ethics". Information and Computer Ethics. Cambridge University Press, pp. 77-97.

Flanagan, Mary et al. (2008). "Embodying Values in Technology". Information Technology and Moral Philosophy. Cambridge University Press, pp. 322-353.

Frankena, W. K. (1973). *Ethics*, second edition. Englewood Cliffs: Prentice Hall.

French, R. M., 2000. The Turing Test: the first 50 years. Trends in Cognitive Sciences, vol. 4, no. 3, pp115-122

Gallant, Roy A. 1980, *Our Universe.* Washington, DC: National Geographic Society.

Guha, Ramachandra (1989). "Radical American Environmentalism and Wilderness Preservation: A third World Critique". Environmental Ethics. Vol 11, No1, pp. 71-83

Jamieson, Dale (2002). "Values in Nature". *Morality's Progress: Essays on Humans, Other Animals, and the Rest of Nature*. Clarendon Press, pp. 225-243.

Kelbessa, Workineh (2014). "Can African Environmental Ethics Contribute to Environmental Policy in Africa?" Environmental Ethics, Vol.36 No.1, pp.31-61.

Harari, Yuval Noah (2015). What Explains The Rise of Humans. TEDGlobalLondon [onlinevideo] Available at: https://www.ted.com/talks/yuval_noah_harari_what_explains_the_rise_of_humans [Accessed 29 September 2018].

Hill, T. (1992). “Making Exceptions without Abandoning the Principle: or How a Kantian Might Think about Terrorism.” Dignity and Practical Reason in Kant’s Moral Theory. Ithaca: Cornell University Press, pp. 196-225.

Hursthouse, R. (2013). “Virtue Ethics.” In Stanford Encyclopaedia of Philosophy. Ed. E. Zalta. Pp. 1-29.

Kant, I. (1993). Grounding of the Metaphysics of Morals. Trans. Ellington, J. Indianapolis: Hackett Publishing Company.

Korsgaard, C. M. 2003. ‘The Dependence of Value on Humanity’, in Raz, J., The Practice of Value, pp. 63–86. (ed.) R. J. Wallace. Oxford: Clarendon Press.

Mbiti, J. S. (1971). *African traditional religions and philosophy*. New York: Doubleday.

Maluleke, T. S. (2008). May the black God stand please!: Biko’s challenge to religion. In *Biko Lives!* (pp. 115-126). Palgrave Macmillan, New York.

McCarthy, J. (2007): What Is Artificial Intelligence. In: Computer Science Department, Stanford University, Stanford CA, November 2007.

McNaughton, D. and Rawling, P. (2007). “Deontology”. in LaFolette, ed., Ethics in Practice. Oxford: Blackwell Publishers, pp. 31-44.

McDermott, D. (2008): Why Ethics is a High Hurdle for AI. In: North American Conference on Computers and Philosophy (NA-CAP), Bloomington, Indiana, July 2008.

Metz, T. 2007. “Towards an African Moral Theory”. The Journal of Political Philosophy, vol. 15, no.3, pp. 321-341.

Metz, T., & Gaie, J. B. R. (2010). The African ethic of Ubuntu/Botho: Implications for research on morality. Journal of Moral Education, 39(3), 273–290.

Mokgoro, Y. (1998). Ubuntu and the law in South Africa. *Buffalo Human Rights Law Review, 15*, 1–6.

O’Neill, O. (1993). Kantian ethics. In Singer, ed., *A Companion to Ethics*. Oxford: Blackwell Publishers, pp. 175-185.

Portmore, Douglas W. (2001). “Can an Act-Consequentialist Theory Be Agent Relative?”. *North American Philosophical Publications,* 38 (4): pp. 363-377.

Ramose, M.B. (2007). But Hans Kelsen was not born in Africa: a reply to Thaddeus Metz. South African Journal of Philosophy, 26(4): pp. 347-355.
University of Kwazulu-Natal Press

Ramose, M.B. (2009). Ecology through ubuntu. In Munyaradzi Felix Murove (ed.), _African Ethics: *An Anthology for Comparative and Applied Ethics*_. University of Kwazulu-Natal Press

Ramose, M.B. (1999). African Philosophy through Ubuntu. Harare: Mond Books

Samkange, S. and Samkange, T.M. (1980). Hunuism and Ubuntuism: A Zimbabwe indigenous political philosophy. Graham Publishing.

Shafer-Landau, R. (2012). *The Fundamentals of Ethics*. 2nd ed.. Oxford: Oxford University Press.

Singer, Peter (1972). Famine, Affluence and Morality. *Philosophy & Public Affairs* 1/3: 229-243.

Sinnott-Armstrong, W. (2003). “Consequentialism”, *Stanford Encyclopedia of Philosophy.*

Tailor, Paul (1972). *Problems of Moral Philosophy: An Introduction to Ethics*, 2nd ed. Dickenson Publishing.

Thomson, J. (1976). Killing, letting die, and the trolley problem. *The Monis*t, 59. pp204-217.

Turing, A. (1964), Computing machinery and intelligence, in A. R. Anderson, ed., *Minds and Machines*, Englewood Cliffs, NJ: Prentice-Hall, pp4-30.

Weisberger, M. (2018, July 8). Why Does Artificial Intelligence Scare Us So Much. *LiveScience.com*. Retrieved from https://www.livescience.com/62775-humans-why-scared-of-ai.html

Windelband, Wilhelm 1921. *An Introduction to Philosophy.* London: T. Fisher Unwin.