

Are machines in the singularity moral agents?

Mohamad Sami El Katouri

1910751

2021



Postgraduate Dissertation coversheet & feedback form:

<i>Student Name: Mohamad Sami El Katouri</i>		
<i>Student University Email: 1910751@student.uwtsd.ac.uk</i>		
<i>Module code: HPPH7014</i>	<i>Student ID number: 1910751</i>	<i>Educational support needs: Yes/No</i>
<i>Dissertation title: Are machines in the singularity moral agents?</i>		
<i>Supervisor: Professor Tristan Nash</i>	<i>Word count: 16,758</i>	
I confirm that this is my original work and that I have adhered to the University policy on plagiarism. Signed Mohamad Sami El Katouri..... Email submission:if you are submitting this form via email, please tick this box ☐ (as a substitute for your signature) to confirm that you agree to the above declaration.		
For completion by the first marker: <i>Comments:</i>		

First Marker's Mark:	<i>Date:</i>
Signature:	
<i>For completion by the second marker:</i> Comments:	
Second Marker's Mark:	Date:
Signature:	
Agreed internal mark:	Date
Signature 1 st marker:	

Signature 2 nd marker:	
Comments <i>(for completion by the internal markers which indicate how the agreed internal mark has been arrived at):</i>	
<i>NB. All marks are provisional until confirmed by the final examination board.</i>	

<i>For completion by the External Examiner:</i> Comments:

Proposed Mark:

Signature of External Examiner:

Final Agreed Mark:

Comments (to indicate how the final agreed mark has been arrived at):

Abstract

Many argue that the singularity is upon humanity and others disagree. If it does occur, and machines crossed the point, can they be considered moral agents? This paper does not focus on answering the legitimacy of the event of the singularity. But instead assumes the singularity has occurred and articulates its argument as if it is within humanity. In the date of publishing this paper, it has not happened. Therefore, as an additional section, the paper explains the singularity as people think it should be and what elements it contains. This thesis is divided into four sections. The first section introduces the idea of the singularity and why it should be taken seriously if not at least under consideration. The second section attempts to define the singularity by many authors and what fundamental elements the singularity possess. The third section explains the main properties for an entity to have moral status. The final section applies the thesis argument that machines in the singularity are moral agents based on the argument 'as it appears'. After defining the elements of the singularity, they are compared with the elements that constitute a moral entity to see if machines do have properties required for moral status. The writer then argues for functional acts of the machine rather than the properties to be considered when thinking of a machine in the singularity as a moral agent.

Acknowledgements

I would like to thank Professor Tristan for supervising this dissertation and the many follow ups we did together.

Also, I cannot forget Nesma Hamouda for her extensive grammatical checkup who re-read this work multiple times.

Thank you very much for all your help

Contents

Abstract.....	8
PART 1: The Differences between A.I. and I.A.	14
Artificial Intelligence (A.I.)	14
Intelligence Augmentation (I.A.)	15
A.I and I.A, a race to the singularity	16
PART 2: What is the Singularity	17
PART 3: are machines in the singularity moral agents?	27
What does it mean to have moral status	27
Top-down approach	29
Bottom-up approach	31
Grounds for moral status	32
Intelligence or capacity to develop sophisticated cognitive capacities and Autonomy	33
Sentience	35
Consciousness.....	37
Member of a species (forms of life).....	39
Conclusion	41

Some consider it good and others bad, but with the rise of Artificial Intelligence (A.I.), many believe that achieving the singularity is a matter of time. The differences as to what the singularity means is an ongoing debate between scientists and philosophers, however, it is generally believed to be the moment when A.I. becomes more intelligent than humans themselves. Another general understanding of the singularity is when A.I. becomes smart enough to create a better A.I. (O'Connell, 2017, p. 70). For some, the singularity and the rise of A.I. is considered a maleficent advancement of humanity, while others believe it is a step in the right direction towards goodness, peace and the pinnacle of civilization.

This discussion is not only ongoing in academia, but also in politics. The president of Russia Vladimir Putin has mentioned that A.I. is not only good for Russia but for the betterment of all humankind. He stated that “Whoever becomes the leader in this sphere will become the ruler of the world.” (Vincent, 2017). Similarly, China who is a leading contender for dominance in A.I. against the United States, has shown a desire to lead the world in A.I. by the year 2030 (Bareham, 2017). The race to lead in A.I. is at foot. Former president Barack Obama also commented on the topic, he observed that the public is not thinking about the singularity as a problem but are rather concerned with whether their jobs will be replaced with robots (Dadich, 2016). Much like the car industry previously required a dozen-men to do a single task, and now a single robot can replace them to finish the task quicker and more efficient, in turn, A.I. are additional factors in the loss of jobs in the market (Kelly, 2020). In fact, with the rapid advancement of A.I. in the field of research, this very thesis could be written by an A.I. in the near future.

The entrance of globalization brought unprecedented benefits to the world. Traveling became easier, food production became faster, and communication became instant. These advances in human technology are few of the many benefits that came around the 20th century while streaming into the 21st. Entering the fourth Industrial Revolution, what was not possible in the late 20th century became available. In an industrial environment, many administrative tasks require human intervention to be finalized, however, the fourth industrial revolution brought about technologies that eliminates human intervention. Such technologies include IoT (Internet of Technologies) and machine-to-machine (M2M) communication which are integrated into automation, robotics, communication, and self-monitoring systems. Trends of new systems

include cognitive computing, 3D printing, augmented reality and most importantly for this thesis, artificial intelligence (A.I.). This phenomenon is advanced by multiple entities from all around the world.

Funding for the fourth Industrial Revolution is supported by government initiatives for profit and non-profit organizations. For example, the German government is investing around \$216 million to support academia, businesses, and governmental entities to pursue those advancement (Moore, 2019). Similarly, in the United States, Smart Manufacturing Leadership Coalition (SMLC) is a non-profit organization that is made from manufacturers, technology companies, universities, government agencies, and suppliers whose main job is to work together to advance the fourth Industrial Revolution (Ibid.). The support comes from both governmental and business sectors to develop A.I as numerous industries world-wide started integrating A.I. into their organizational workflow. For instance, when calling the bank on the phone instead of an answering machine that tells you to dial different numbers to be directed to different tellers, banks started using A.I. to talk with callers and re-direct them. Additionally, human tellers started being replaced in physical banks. Robots move around the entrance of a bank to communicate with humans and direct them to the teller specialized in their requirements (Roberts, 2019).

In post-World War II Japan, robots were made for industrial jobs, until in the 1990's robotics began transitioning to "service" robots. At first this phenomenon was using robots in drudgery work such as cleaning-bots, firefighting, refueling cars and demolitions. After the phenomenon was established in society, the transition went to home servicing such as mowing the lawn and cleaning the house (Levy, 2007, p. 7). Jumping forward into 2010's, Toyota announced that they will create robots that would help the elderly which is not meant to remove human intervention, but rather to help them reduce tasks and labor on the supportive entity. However, this is an extension for the plan when people are not available, in the future, these robots are supposed to be slotted in when the number of elders are higher than available number of people to take care of them (Reuters, 2017). Robotherapy is a new field that introduces psychological therapy using robotic A.I. technology. This uses virtual reality for problems such as fear of heights. Internet communication that applies tele-hypnosis and distant psychotherapy to help clients get through their problems (Libin & Libin, 2005, p. 1-3).

The fine line between humanity and A.I is getting thinner and thinner. Cynthia Breazeal leads the team at MIT which created the Kismet/Kismet, a robot that has features just like a human face, eyes, mouth, head and lips. The idea came from their study that has shown that humans become at ease when they communicate with an entity that has features similar to theirs. Testing the robot with humans, they have found humans did not take the robot at full ease, especially the adults because of its mechanical movement. However, children had better communication and felt more at ease (Breazeal, & Scassellati, 2000, p. 50-51 & 71). Moreover, football robots were the centre of a study that focuses on creating football teams that competed with one another (Kim, & Kim, 2003, p. 17-18). The same developer who developed the football robots created a software that acts like a computer DNA. Fourteen chromosomes were simulated in a two thousand bytes of data in a memory. This space allowed the software to exhibit seventy-seven human behaviours (e.g. angry, happy etc...) and project them onto a screen (Kim et. al, 2009, p. 19-20 & 26). The examples provided above is not necessarily linear but it can be seen that the path to A.I.'s singularity is not going downward anytime soon. In this context, the singularity is a matter of time.

This leads to the question of this thesis, once machines cross the singularity, should they be considered moral agents?

Humanity has dealt with morality that focused on human-to-human interactions. As much as there was animal to human and vice versa discussions, the main topic focused upon was human-to-human moral philosophy. Moral theories such as deontology, virtue ethics, and utilitarianism are systematic in their arguments to derive an act (utilitarianism and Categorical Imperative) or a character trait (virtue ethics) to be deemed moral ¹. However, these issues are usually discussed in relationship to what qualifies a human as a moral agent. Utilitarianism is a consequentialist moral theory that considers the utility of an act to be the basis of morality. According to the theory, the aim for any action should be maximizing pleasure and minimizing pain. (Levin, 2019, p. 130-131). For example, theft has no intrinsic moral standing; neither bad nor good, rather, it becomes morally good if it maximizes pleasure and minimizes pain for the majority of society (e.g. Robin Hood) (Mathews & Hendricks, 2019, p. 44-45).

¹ Categorical imperative, virtue ethics and utilitarianism will be explained later on.

The theory helps us in deriving what is a moral act, however, it does not tell us who it is incumbent upon, what entity should the moral act ascribe to? The debate is usually concerned with humans, but what is it that makes a human a moral entity? This question is the main topic of the thesis. After determining the singularity of A.I., can humans consider A.I.s who reached singularity moral entities? I.J. Good in his famous article “Speculations Concerning the First Ultra intelligent Machine” (Good, 1966), argues that an ‘ultra-intelligent’ machine can surpass all intellectual activities (considerably human). If humans (with their intellectual activities) design the machine, then the ‘ultra-intelligent’ machine can design a better one. This would cause an ‘intelligence explosion’ (p. 33). The moment the explosion happens is the moment singularity occurs. Good’s view of the singularity is based on intelligence, he measures machines reaching the peak not by machines mimicking humans (e.g. exhibiting emotions, sentience, or even facial features and expressions) but rather intelligence. This intelligence can be based on rationality or intellectualism (self-learning). Therefore, these machines could be considered moral agents by Kantian moral philosophy as according to Kant, morality is defined by rationality (required for intelligence). If an entity has rationality and autonomous will, then according to Kant, the entity has a moral responsibility ² (Bennett, 2015, p. 77-78). The ‘ultra-intelligent’ machine is not considered a moral entity. Even though it has rationality beyond human’s capabilities, but if it is programmed with boundaries in its actions, then it is not a moral entity. However, if no boundaries were given, then by Kantian terms, the ‘ultra-intelligent’ machine is a moral entity and can be held accountable to its action.

In conclusion, robotics and A.I. in daily life can be found in many applications around us. This process has been progressing and has not halted. The predicted destination with the advancement of A.I. is the singularity. However, although A.I. exists in our midst, they are very limited in their functionalities and not human-like in their behavior, yet. Levy suggests that in the mid twenty first century, technology will advance the computational speed and power which possibly enhance technologies such as DNA computing, quantum computing and optical computing. He assumes in the future when these new technologies improve drastically and combine with the advances of A.I., the intellectual and emotional capabilities of robots will be astounding. He thinks they will be human-like or how humans want them to look like. Additionally, he thinks

² Kant’s morality will be explained later in the thesis.

that they will have human consciousness or better yet, superhuman consciousness (Levy, 2007, p. 10).

For the skeptics of Levy, Hurbert Dreyfus thought that A.I. was a fraud in which he describes it as ‘alchemy’ in 1965 because there was no evidence for such machines can defeat human capabilities (Dreyfus, 1965; McCorduck, 2004, p. 236-237). According to Dreyfus, machines receive data as digital and analogue, while the central nervous system is a complex mixture of both when receiving data (Dreyfus, p. 82). He argues it is a futile end for A.I. to reach the level fantasies and dreams people thought of and he reasons because machines are limited in their ability to input complex life situations of data. He likens the A.I. researchers to be like alchemists who years in the past who tried to convert lead into gold, only concluded it is impossible (p. 84). If one does not believe him, he argues, simply look at the results. What is being created is not gold, but color spectrum of solutions that does not have a speck of gold (p. 86). In other words, a pursuit that would never ever materialize. Little did Dreyfus know, in 1997 the world champion of chess, Gary Kasparov would lose to IBM’s Deep Blue A.I. in a game of chess (Levy, 2007, p. 20), making the world champion of chess at that time an A.I.

A.I. was successful in defeating humans in a game created by humans. Hence, creating artificial intelligence into a machine is one way and turning humans into sub-machines or full machines is another. This is the difference between A.I. and I.A , both will be explained in this thesis in order to explore how machines can lead to the singularity and whether this would make them moral agents.

PART 1: The Differences between A.I. and I.A.

This part of the thesis defines A.I and I.A and explores their possible outcomes.

Artificial Intelligence (A.I.)

Artificial intelligence (A.I.) is a method to create or replicate “human” intelligence inside a machine. It started with the idea that evolution produced general human-intelligence. Since the evolutionary process created the human level intelligence, with the human intelligence and the knowledge of the evolutionary process, a greater outcome can be achieved for a far greater efficient intelligence (Bostrom, 2016, p. 28; Chalmers, 2010, p. 16). With advancement in technology, capabilities in estimating evolutionary process, and human engineering, creation of intelligence is feasible and possible. Some would argue it can happen in a few centuries (Chalmers, 2010).

Theoretically, A.I. can be created in few ways, one of which is the use of genetic algorithms on super computers that can achieve results of biological evolution. This method attempts to create an algorithm that can grow by itself through multiple inputs, whether by surroundings/interactions, fed information by humans, or using a module that can think and then be used with the genetic algorithm to evolve its thinking. There are few attempts using this method, however, they’re all on a minor scale (Kim et. al, 2009, p. 19-20 & 26). To which degree this method will be successful will depend on the advancement of computational power (Bostrom, 2016, p. 29).

The reason why this method is feasible is because of the advancement of technology that allows humans to map and study the biological brain. By studying the biological brain, creating an A.I. becomes feasible by programming brain functions into the machine (p. 33). Another method is the ‘seed AI’ that is usually associated with Alan Turing’s conception of a child A.I. as proposed by Yudkowsky (Yudkowsky, 2007a, p.482). A seed A.I. is an A.I. that have a set of functionalities and slowly can enhance itself by studying its inputs. The difference between the argued self-enhancing A.I. and a seed A.I. is that a seed A.I. has knowledge of its functionalities while the other has to explore their functionalities (Bostrom, 2016, p. 34-35).

Even though it might be unorthodox and not necessarily direct, enhancing human intelligence to speed up the process of programming/developing an A.I. is a method. This method uses technologies (e.g., stem cell research) and drugs to enhance human intelligence which would allow humans to develop the machine faster and more efficient. Another unorthodox method is biologically enhancing humans before they are born to create stronger intelligence. Upon growing up, those humans can develop an A.I. faster (p. 43-54). By those methods, programming an A.I. becomes the method to reach the singularity.

Intelligence Augmentation (I.A.)

Another method for transcribing intelligence into a machine is through intelligence augmentation (I.A.). Strongly suggested by Kurzweil (Kurzweil, 2005, p. 33), instead of programming intelligence into a machine, I.A. suggests enhancing an existing human brain by connecting it with a machine so that it would simultaneously have the functionalities of humans and the intelligence of the machines, also known as transhumanism. Vinge suggests that even if A.I is used to pave the way to singularity, I.A will still be required. His argument is that not only will intelligence be integrated into a machine, but also other human aspects, such as emotions and sentiency (Vinge, 1993, p. 17).

Another method of I.A. is emulating the whole human-brain into a machine. This method includes neuron pathways and memories (Bostrom, 2016, p. 36). Bostrom provides a simple procedure that starts with scanning the whole brain, then a machine dissects the brain tissue into section or slices. Those slices are then inserted into a scanner that scans the neuropathways of each section. Second, the scanned parts of the brain and the scan of the whole brain are inserted into the computer to develop a full 3D model. Third, with the use of a fast computer (quantum computers), neurocomputational structures are built. If those steps were successful, then an emulation of intelligence is digitized into a computer (p. 36-37). The whole-emulation process does not require full knowledge of the brain, what is most important is sufficient knowledge of how intelligence works (p. 36).

A.I and I.A, a race to the singularity

The argument for I.A. is that it has aspects of humanity that A.I. cannot achieve, one of them is consciousness. However, the definition of consciousness is not yet agreed upon (Legg, & Hutter, 2007). This thesis will not delve into consciousness extensively, it is best left for the philosophers of the mind and neuroscientists. However, a brief explanation is required to understand consciousness in machines. There is a broad agreement that there is a difference between phenomenal and cognitive consciousness (Chella & Manzotti, 2007). In contrast, Holland provides a difference between both in different labels: weak artificial consciousness and strong artificial consciousness (Holland, 2003). Weak artificial consciousness is when designing a machine that can simulate conscious or cognitive behaviour, also called cognitive consciousness. Strong artificial consciousness is a machine that can have phenomenological experiences such as self-awareness, subjective experiences, and dreams, also called phenomenal consciousness. Usually, most academicians work on cognitive consciousness. Self-awareness is an aspect of consciousness. In this context awareness involves awareness of the self, awareness of its goals, awareness of memory (previous acts) and retrieval of information (Zeman, 2001, p. 1265-1277).

It is worthy to note that machines are already capable of retrieving stored information. Learning and calculating/anticipating consequences are also part of consciousness (Ordahl, 1911, p. 205-206; Sanz et al., 2012, p. 393 & 396). The ability to input information, process information, and create a probable output is also part of consciousness. Machines can do so as a computational functionality rather than phenomenological. Functional consciousness is simply the consciousness of doing a task, while phenomenological consciousness is being conscious of doing the task and have the ability of 'looking inward' or introspection (subjective experience). For example, looking at the blue sky may evoke an emotional experience for a human being, but A.I. simply sees blue and compares it with anything else it has in memory that is blue (Carruthers, 2003, p. 239). Although A.I. may have cognitive consciousness, we are currently still incapable of programming phenomenological experiences into a machine; making I.A all the more intriguing.

PART 2: What is the Singularity

This part of the thesis is for understanding what singularity is from the point of view of various authors, identifying the elements that comprise it and concluding possible, if not probable, elements of the singularity. For the sake of identifying common underlying singularity elements, this thesis will divide them into two categories. Aspects that cause the singularity are the cause category. Properties that attribute the singularity to a machine are the function category. For example, when referring to acceleration of technology, the acceleration would be a cause category and the machine intelligence would be a function category.

The origin of the term Singularity can be ascribed to John von Neumann (Eden et al., 2012, p. 4). In a tribute article to John von Neumann, Ulam recounts a conversation with him (Ulam, 1958, p. 5). He recalls Neumann mentioning if technological acceleration causes changes to human life, then it is indicative of an approaching singularity that after which human affairs cannot continue. Two aspects of Neumann's singularity can be derived from this conversation, acceleration, and discontinuity. Discontinuity refers to a termination of an event, and it is happening to human life as we know it in this example. Acceleration is considered a rate of growth in a quantitative manner, and it is happening to machines in this example.

Generally, technology has been accelerating onwards. This observation is famously postulated by Gordon E. Moore (Moore, 1965). He argues that the transistors in an integrated circuit of a machine doubles every two years. This phenomenon was termed "Moore's Law". Even though it has no scientific or empirical grounding, Moore's law held its accountability since 1975 until the writing of this thesis. What is speculated is that if the technological acceleration explained in the introduction is combined with Moore's Law, presumably, this will eventually lead to the singularity. This stance is supported by many authors such as Kurzweil and Goodwin. In addition, artificial intelligence will not be the only leap in acceleration towards the singularity, machines might evolve and achieve consciousness and emotions.

Irvin J. Good suggested that an 'Ultra-intelligent' machine could be created (Good, 1966). He defines an ultra-intelligent machine as a machine that can surpass all human intellectual activities, a machine that can design better machines, which can cause an 'intelligence explosion' (p. 33 & 78). In other words, the singularity as Good argues is that the survival of humanity depends on the creation of an ultra-intelligent machine. There are two steps to the ultra-

intelligent. First, humans create the ultra-intelligent by themselves. Second, the ultra-intelligent will begin enhancing itself. He advocates the study of the human brain to create neural network maps that can be programmed into a machine. Since the first step is to create the artificial neural network, the rest will be handled by the network to self-enhance. If it has ‘sensorium’ and ‘motorium’ (input and output) the machine will learn by experience and enhance itself (p. 31-32).

He speculates this would be the last creation of humanity unless the machine is docile and requites its creators to know how to control it. Many years later, the respected physicist Stephen Hawking also warned that if humans created an A.I. that surpasses human intelligence, it can be considered the greatest achievement in human history and their last (Hawking et al., 2017). Through his explanation, the *cause* categories derived are discontinuity and acceleration. However, discontinuity can be eliminated if the self-enhancing machines are provided with methods to control it. Function categories are self-enhancement and intelligence beyond human abilities. Even though he speculates that machines can have emotions (p. 34), his focus remains on intelligence. An important aspect for intelligence explosion is the ability of the ‘ultra-intelligent’ machine to create a better one from itself. Hence, self-enhancement is a requirement for a machine to pass the singularity.

Good’s argument is further extended in Muehlhauser and Salamon’s chapter on intelligence explosion (Muehlhauser & Salamon, 2012). It should be noted that they expect the intelligence explosion to be due to increase in computational method (A.I.) rather than biological enhancement (I.A.). Computational powers of machines are greater than humans. The human brain uses 85-100 billion neurons, while machine intelligence can use computational resources in a brain size of a warehouse. A human brain’s axon can carry 75 meters per second per signal or less due to human physiology limitations. However, computers can go faster in initiating and sending computer signals across a larger scale. Duplicability of A.I.s is something humans cannot perform. An A.I. can simply copy code and add somewhere else in another hardware in which machine population can surpass human population anytime (p. 28). Task orientation with computational A.I. is better than human’s coordination. Imagine an A.I. code is copied across several machines; these machines would be grouped as an A.I. clan. As a copy of the original machine, the goal(s) of the original is also copied. Hence, all can work in coordination to achieve

the goal(s) (p. 29). With the advantages of self-programming/improvement and enhanced computational abilities, A.I. will eventually lead to an intelligence explosion.

In his conference paper, Vernor Vinge expects humanity to create 'super intelligence' machines after which the human era will cease to be (Vinge, 1993, p. 11). Vinge argues for three reasons a breakthrough in machine intelligence will occur. First, the direction humans are creating machines that can be 'awake' (aware of surroundings). If that is achieved, then super intelligence is a matter of time. Second, large networks of computers can communicate with one another. Third, biological science is pushing for enhancing human intellect. All those three points depend on computer hardware. So as long the hardware improvement is moving forward (e.g. Moore's law), then super intelligence can occur. Humans have the ability to take input data from the world and create 'what ifs' situations. If machines would be able to do that, according to Vinge, then rules for human rationality will change since machines can do it faster. This moment is defined as the 'the singularity' according to Vinge (p. 12). Though when the singularity happens, extinction of humanity is a possibility, but not probable. Vinge gives us the example of animals (p. 16). Humans can live without specific types of animals, yet their rationality stops them from killing those animals. Therefore, in the post-singularity time, machines will spare humans so long as there is not a reason to kill them.

One of the most influential authors on the topic of the singularity is Ray Kurzweil (Kurzweil, 2005). At the centre of his argument is the 'law of accelerating returns' which argues for the exponential growth of products (e.g. technology or hardware) in an evolutionary process (p. 41). Evolution in his context is not biological but rather the general evolution of technology acceleration (p. 45). Kurzweil demonstrates the concept of evolution in his six epochs of time. The first epoch is the origin of states where structures such as energy and patterns of matter existed. After the big bang occurred, chemistry and physics were born, and the rules of the universe were created (p. 22). In the second epoch life began. Carbon based compound aggregated and life originated. This caused biological systems to evolve and create the DNA to bank information. In the third epoch, DNA guided the evolutionary process to create organisms that could detect information and then store that information in the brain (p. 23). The fourth epoch was introduced by humanity with the creation of technology. With technology, storing, removing, and retrieving data became easier. As the computational speed increases, so will its

abilities (p. 24). In epoch five, human intelligence and human technology will unite into one. The rise of human-machine civilization will occur within this epoch. In this stage, humanity will need not worry about old age and diseases (p. 28). In epoch six, speed of light becomes superseded and trans-humanists will be able to infuse with the universe and organize matter and energy (p. 29). By presenting those six epochs, Kurzweil argues that the exponential growth into epoch five is a matter of time as the technology evolution is happening rapidly. He presents a thorough explanation as how the law of acceleration of returns works (p. 41-88) the following is a summary of the concept.

In every technological evolution, stronger methods to help us evolve into the next stage emerge. After every stage, the progress *exponentially* increases which over time arranges information for the next evolutionary stage in a better manner. This leads to ‘returns’ such as speed, power of processing, cost-effectiveness. Kurzweil does all of this through statistical information and charts on different fields and industries, for example, amount of cell phone users across different times. Even though the law is hypothetical, it is nonetheless a statistical observation and there is no guarantee it can reach the singularity. He is not worried of his theory falling apart since humanity’s charting of the brain and technological reverse engineering of the brain is an ongoing process (p. 115). In twenty years (since his publication) a comprehensive map of the brain will be produced and the procedures to study the brain becomes easier. His argument on this part was not wrong (Oregon Health & Science University, 2019), since 2005 until this thesis many comprehensive methods were developed to study the brain.

By mid 2020s, Kurzweil argues that human emulation into machines becomes possible and at this point, humans would not be able to distinguish the difference between talking to a human and a machine (Kurzweil, 2005, p. 32). Upon the merging of humans and machines, rationality, intelligence, and language speech will surpass all human capabilities and will not be abated by biological limitations (p. 33). In this stage, machines will be continuing the law of accelerating returns without human intervention. Nanotechnology will be created and then eventually nanobots. Those nanobots will be able to enter the body, heal it, and manipulate the brain to create virtual realities. This is the point humans cease to be humans and become transhuman (p. 34-35). At this stage of the singularity, emotions can be understood by machines which in the future other machines can comprehend and build upon. Altering human reality where anyone can

change bodies according to preferences without any limitations becomes possible. This is stage where the singularity ends for Kurzweil.

In epoch six he argues that beyond singularity humans will be able to fuse with the universe and alter its laws, then use the energy of the universe to increase the accelerating law of returns. The *cause* category in Kurzweil's argument is acceleration of technology, but does not include discontinuity of humanity, rather humans merge with machines (**I.A.**). As for the *function* categories, they are intelligence and rationality. Since Kurzweil argues for transhumanism, it is safe to say, sentiency is a *function* category because humans do feel and if bodies eliminated sensory pain, at least understanding the concept will be in the brain. Emotions play a part, but reaching computational speed plays a far more important role since machines will only understand emotions upon merging.

Pollock argues for programming reason into a machine rather than total emulation (Pollock, 1990). He argues the way for machines to reach above humanity is by intelligence. In the centre of his argument, machines should emulate human behaviour for A.I. to be considered a 'person'. He proposes that for a machine to be human-like, it should possess both practical and theoretical reasoning. Theoretical reasoning focuses on what the machine should believe while practical reasoning focuses on what the machine should do in a given situation. In other words, theoretical reasoning is approving and rejecting beliefs while practical reasoning is emotional states, intentions, actions, and desires (p. 469). The main argument is that for machines to reach above humanity (which it already does in computational problems), it first has to emulate a human being emotionally, rationally, bodily, and behaviourally. At the core of solving this conundrum, making a machine a 'person' is the way forward (p. 462 & 465-466).

David Chalmers argues for the singularity from a philosophical perspective (Chalmers, 2010, p. 9). Continuing from Good's argument of an intelligence explosion, Chalmers agrees if the event occurs, it can advance humans or destroy them (p. 9-10). He begins his argument in the following manner:

1. *There will be AI (before long, absent defeaters).*
2. *If there is AI, there will be AI+ (soon after, absent defeaters).*
3. *If there is AI+, there will be AI++ (soon after, absent defeaters).*

4. *There will be AI++ (before too long, absent defeaters).* (p. 12)

Premise 1 holds validity from the previous explanations on technological advancement. Premise 2 is the argument of the singularity and most authors argue mostly on premise 2. Basically, the conclusion shows that the singularity is coming. 'Soon after' are time frames considered to be decades, and 'defeaters' are considered global catastrophes (e.g. nuclear war, natural disasters) (p. 13). Arguing for premise 1 against the unlikeliness in transhumanism, Chalmers argues programming intelligence in a machine is feasible. He argues that as long as intelligence is wholly measured through behaviour, then AI++ has a strong argument that is enough to be real (p. 15). He additionally argues since evolution created human intelligence; humanity will be able to create human intelligence and replicate it in machines (p. 16). If not direct creation, then creating dumb machines and then allow them to evolve to improve their intelligence (p. 17).

Arguing for premise 2, Chalmers elucidates that after every technological product, it later becomes obsolete because of technological advancement. Similarly, A.I. will become AI+ and A.I. becomes obsolete. Although not everything is extendible in that manner (e.g. brain emulation into machine does not mean the full intelligence upload), however, at one point A.I. will become AI+ (p. 17-19). Arguing for premise 3, if there is an AI+ which is better than its creator, then AI+ can create an AI++. In the similar context, AI++ will be able to create a better AI++ or AI++++ (p. 20-21). Therefore, a possibility of intelligence explosion cannot be considered unlikely, and it can be done by gradual upgrading or uploading (p. 63). According to Chalmers when the singularity occurs, humanity has four options: isolation, inferiority,

extinction, and integration. First option separates humanity and machines from one another. Since machines have the technological progress due to its advance intelligence, humans will fall out of progress. Second option is dangerous as it leads to an unjust society, whether it's humans or the machines that possess superior intelligence. In both cases, extinction can happen next due to preference of one over the other. Third option is the possibility humans will try to avoid this altogether through creating machines that would not kill them (**See part 3 of the thesis**). The last option is integration which can either be by transhumanism or by humans and machines living in harmony with one another. This can be done by programming their values or principles beforehand if transhumanism is an eliminated option (p. 41-42).

Extracting the categories from Chalmers argument, the *cause* categories are acceleration and discontinuity. Unlike other authors, acceleration is a minor reason and between his premises, the timing between A.I. becoming AI+ and then AI++, is not important. Discontinuity occurs in Chalmer's argument in the form of what he refers to as defeaters. Defeaters are global catastrophe situations that prevents humanity from reaching the singularity, such as the plague and nuclear war. Additionally, he argues if anything that prevents humanity from reaching the singularity it would be humanity itself. Their motivational reasons to pursue the singularity out of fear will prevent others from reaching that stage (p. 29). The *function* categories of Chalmer's are intelligence and behaviour. Intelligence explosion is the main criteria for the singularity, hence the capability to surpass human intelligence. Also, if machines showed behaviours similar to humans, with the addition of intelligence, they can be considered to have reached singularity.

Yudkowsky breaks the occurrence of the singularity into four stages, acceleration, event horizon, intelligence explosion, and 'apocalytism' (Yudkowsky, 2007b). In terms of acceleration, he refers to the technological acceleration as discussed earlier (e.g. Moore's law) and Kurzweil's law of accelerating returns. Event horizon happens when humanity is to considered the greatest intelligence living on the planet, so technology advances to improve human intelligence (e.g. brain-computer interfaces) with integration of machines. Event horizon can be considered a sub-part of acceleration where at one point technology is accelerating, it will reach the ability to improve human intelligence (Bostrom, 2016, p. 36). Intelligence explosion is a moment when the machines are able to surpass human intelligence. He attributes this argument to Good and himself. Although he deems this unlikely to happen, 'apocalytism' is the last stage and it plays

the role of discontinuity in *cause* category. At this point, machines will eliminate humanity and rule the planet.

Predicting the time when the singularity will come about is not supported by any empirical evidence. What is predicted is based on observation and speculations. This is usually due to two antecedents. First, the observation through the extrapolation of progress of technology. Until a reliable method is available, authors, academics, scientists, and laypeople will also speculate. Second, by estimating progress in the scientific community and how much research has achieved in A.I. If more researchers and efforts were put into the research, reaching the singularity only becomes closer (Muehlhauser & Salamon, 2012, p. 20). For example, this method is supported by Moore's Law and Law of accelerating returns of Kurzweil. For instance, Kurzweil predicts the singularity will be around the year 2045 by using estimations of computational powers (Kurzweil, 2005, p. 125). There are methods being created to predict long-term forecasting technological progress (Nagy et al., 2010), however, none have yet to show promising results or strong reliability. Yet many predict the year of the singularity based on observation. For example, Chalmers (Chalmers, 2010, p. 13) predicts around 2100, and Kurzweil around 2045. It can happen on those years, or it cannot. However, acceleration is continuous and humanity's search/curiosity for that technology is not stopping, that is why it cannot be ignored.

One of the dangerous elements of the singularity that is usually inferred but not necessarily declared is machine autonomy. It is usually inferred when the singularity will end humanity. To do so, it requires autonomy. With occurrences such as surpassing human intelligence, machines will become autonomous and perform actions by themselves without an agent commanding them (Yudkowsky, 2008, p. 11-12). Assuming humans created A.I.'s with commands not to harm humanity, such as the famous Asimov laws of robots. This section gives a brief explanation of how a machine becomes autonomous to understand its moral agency. The first law is that robots cannot harm humans or let a human come to harm. In the second law robots must obey humans except when it conflicts with the first law. In the third law, robots must protect their own existence unless it conflicts with the first two laws (Anderson, 2008, p. 477-478). If we program and hard code such rules into machines before they reach the singularity, then machines with super intelligent capabilities would not be able to harm humans.

The problem is with definition and interpretations. What is considered harmful to humanity? What is not? These laws are too vague and abstract for a machine to interpret (Etzioni, 2017, p. 166). For example, humans often go to wars to obtain freedom; a machine might interpret this as freedoms are causing wars between humans and so it must be eliminated. In this logic, machines will either override the first law by enslaving all humans to ‘protect’ humanity. Like an overprotective parent, this ability is an autonomous act which is done by intelligence. Similarly, if robots are expected to enhance themselves, whether by code or by hardware, some autonomy is required for them (Yudkowsky, 2007a, p. 482). Kurzweil, Chalmers, and Good all argue for self-enhancing machines by their own self-learning intelligence. Hence, machine autonomy is a feature of the singularity.

Not all bumps for the singularity will be due to some problems in the *cause* category. Progress in every step does not necessarily bring humanity closer to singularity, rather after every step of progress a new problem occurs (Arbesman, & Arbesman, 2011, p. 246). More discoveries in the field brings more efforts to progress. Another problem is with Moore’s law will finally break down. Up to date for this thesis, Moore’s law is successful, however, some might argue it will not hold long enough and anytime soon the process will break down (Mack, 2011). This speculation can also be applied to Kurzweil’s law of accelerating returns. Another preventive method would be humanity inclination not to accept any progress towards the singularity. In other words, what will stop the singularity are societal refusal for institutions and governments to progress into that step (Chalmers, 2010). Chalmers argues it would be active prevention from people and not allow machines to reach the singularity.

However, just as much as the singularity has its proponents, the singularity has a fair amount of its detractors. Steven Pinker, a Harvard professor of psychology who taught in the brain and cognitive department in MIT, criticizes the idea of the singularity. He argues the simplicity of visualizing and imagining a future does not make it necessarily possible. He cites ideas from the early 20th century such as dome cities, underwater cities, and jetpack commuting as fantasies that fascinated him when he was a child but never came true (IEEE Spectrum, 2008). However, two companies are now offering jetpack services where they charge members of the public to try out jetpacks, attaching the users to a large metal frame by wires to prevent them from flying off

uncontrollably (BBC News, 2021; Debusmann Jr, 2021). In fact, this year witnessed the Royal Marines testing out jet suits to assist in Maritime Boarding Operations (BBC News, 2021).

Although there are critics who argue that the occurrence of the singularity is improbable, there are aspects of it that are currently manifesting. Keeping an open mind and having a preparatory plan would not harm the current sphere of society or any individual. Therefore, learning about the singularity for those who foresee its happening is incumbent. Learning about the singularity for those deeming it as hokum cannot harm them and they can keep it as a utility in the event of its occurrence.

This thesis does not support or argues against the singularity. However, this thesis analyses that if singularity is to happen, then it can be seen through the two following categories. *Function* categories of the singularity would be super/ultra-intelligence, autonomy, consciousness, and emotions (behave like humans). While the *cause* categories are acceleration of technology and discontinuity, this entails either machines will end humanity or humanity transcends and joins with machines. The next section will explore whether the machines can be considered moral entities or not in the case of the occurrence of the singularity.

PART 3: are machines in the singularity moral agents?

It can be assumed that machines will have the qualities or aspects that were explained previously once they reach the singularity. Since machines in the singularity would have aspects similar to humans, should they be considered moral agents? Do they have moral agency? If a crime or an accident done by machine to another machine or to a human, is it liable for punishment? This part of the thesis attempts to answer those questions.

What does it mean to have moral status

Different people might interpret moral status differently. Reichlin interprets it through two attributes (Reichlin, 2017). The first is the ability to be morally considerate of others' preferences, needs, or interests; and practicing morality in actions to others as well as extensions of them. For example, jewellery has economic value for people, but lacks moral status. However, if stolen or destroyed, the jewellery is not wronged, rather the owner is harmed in terms of the owner's interests in the jewel (Reichlin, 2017, p. 2). The second is having moral rights. This means to have a standing for valid claims of or entitlements to specific treatments from other agents' virtues or capabilities. An example of this is the human rights declaration, where humans are entitled to rights from other agents (institutions or other humans) due to them being humans (UN General Assembly, 1948).

The Stanford encyclopaedia of Philosophy's entry also states that to have a moral status there are two distinctions. First (which is similar to the first Reichlin interpretation), there is a need to consider the consequence of the action according to the action. For example, assuming utilitarianism is the applied moral theory, one has to consider the utility applied towards the agent. Second (which is similar to Reichlin second interpretation), there is a need to consider the entity to itself, whether it will have pre-empted entitlements or not. For example, the human rights declaration provides humans pre-empted rights before judging the action (Tannenbaum & Agnieszka, 2013). This leads to establishing the difference between moral patients and moral agents. To have a moral status, you can either be a moral patient or a moral agent. Moral patients are entities incapable of performing moral decisions or acts but require to be treated in a moral manner. For example, animals or impaired human beings. The acts they perform cannot be taken at full responsibility due to lacking properties such as intelligence or rationality which fully functional adults possess. Moral agents are entities that can perform moral decisions and acts.

Blame and responsibility can be attributed to such entities. This thesis questions whether machines should be treated as moral agents in the singularity.

Dennett suggests that to hold a computer/machine morally responsible, it should be a higher-order intentional system (Dennett, 1997). Assuming humans can implement “cognitive” or “motivational” states in a machine, an intentional system is an A.I. whose behaviour and rationality can be predicted based on its “cognitive” or “motivational” beliefs (p. 352).

Intentional systems do not have beliefs, yet beliefs can be translated/interpreted into goals that the machine requires to achieve. Higher-order intentional systems are machines that can give attention to more than its expanded resources. It should be able to observe inefficient habits. In other words, it should contrive beliefs of its own beliefs, beliefs on its thoughts and fears. To Dennett it is a requirement for moral responsibility (p. 354). Since machines pre-singularity are intentional systems, they cannot be held morally responsible. For example, Dennett argues that HAL, the A.I. from the film *2001: A Space Odyssey* has such capabilities. Therefore, it is a morally responsible A.I. (p. 354-355).

Sullins ascribes machines to be moral agents based on three arguments (Sullins, 2006). First, if a machine is autonomous from any controller, human or otherwise, it is considered morally responsible. Second, the intention of the machine. Intention is a complex subject even in humans’ own domain, more so for machines. If a machine’s behaviour can be interpreted on ‘standard folk psychological notions’ in knowing the harmful and the harmless, then it can be considered a morally responsible (p. 28). Third, responsibility towards other entities. Whether humans or other machines, if a machine has duties towards others and holds a position of responsibility (e.g. job), then it is a moral agent.

Additionally, Sullins argues if the machine has behavioural patterns that shows duty towards others, then it is also a “full moral agent” (p. 29). Nissenbaum focuses on accountability of machine action rather than moral status (Nissenbaum, 1994). At times, people avoid responsibility and try to shift the blame on to others when a problem occurs. Similarly, when a problem occurs, there is a tendency to shift the blame on to a computer, in other words, “technical errors”. However, organizations and companies put forward the ownership of the software/hardware of a machine, yet when the machine goes haywire or causes a problem, they avoid the responsibility claiming the users mal used or other reasons (p. 75-77).

When it comes to machines, assigning moral status becomes difficult. There are many who argue that moral status cannot be given to machines due to morality's complexity (Kuflik, 1999, p. 178-179). Even if ethics programming into machines is possible, it is better left for humanity to be able to override the ethics component or moral decision of the machine (p. 181). However, there are people who are able to anthropomorphize machines and by their actions consider them moral agents (Reeves & Nass, 1996, p.20-27; Nass, & Moon, 2000, p. 93 & 99-100). According to Moor, there are three types of machine ethical agency, the implicit, explicit, and the full ethical agents (Moor, 2006).

Implicit ethical agent is a machine that has been supplanted with commands and follows the approach given when making decisions. These commands implemented eliminate unethical behaviour. For example, ATM machines are programmed to do tasks that human tellers can do. Where humans can lie and cheat, ATM machines are programmed accurately do the task disabling them from conducting such an ethical behaviour (p. 19). Explicit ethics agent are machines that 'perform' ethics. These types of moral machines are embedded with a moral approach (e.g. moral theory) as a framework. So, a machine would compute and review decisions according to the moral framework and act accordingly. Full ethical agents are moral machines that can perform moral decisions and has the competency to justify them, similar to human beings. In the singularity, the main focus would be on explicit and full ethical agents. There are two approaches to program ethics into a machine to become either of these ethical agents; the top-down and bottom-up approaches.

Top-down approach

Since discontinuity is *cause* category for the singularity, the fear of machines annihilating or enslaving humanity can be alarming. As a preventive method, it is argued that laws or ethics are pre-programmed into machines. At least before the machines reach the singularity. This approach is called the top-down approach (Wallach, Allen, & Smit, 2008, p. 568-569; Allen et al., p. 252). This approach deems artificial moral agents (AMA) to be ipso facto autonomous to no or a certain degree (Allen et al., 2000, p. 251-252). That is why the more freedom a machine possesses, the more the moral standards are required (Wallach, Allen, & Smit, 2008, p. 566-567). Since an aspect of the singularity is self-enhancement, then a certain degree of autonomy is required. To prevent the discontinuity aspect of the singularity, a proper instructions or ethics are

required to be pre-programmed not to harm humanity. The main intricacy is knowing which moral theory or set of instructions to be pre-programmed.

Rules or laws are programmed to create a rule-set that machines can or cannot perform. Such rule-sets was suggested by the story of I-robot by Isaac Asimov. This rule-set became to be known as the Asimov laws. The first rule is that machines may not harm a human being or, through inaction, allow a human being to come to harm. The second rule is that a machine must obey orders given by a human except where such orders conflict with the first law. The third law is that a machine must protect its existence as long its protection does not conflict with the first and second laws (Anderson, 2008, p. 477-478; McGrath & Gupta, 2018). One of the problems of the law is the argument of definitions. In the first law, 'harm' is not properly defined (Etzioni, 2017, p. 164-165). For example, machines imprisoning humans in their homes to prevent any physical harm from the environment. Another issue is the hierarchy of command argument. If a human in high-command ordered the machine to inaction while someone is dying, while another human being in lower-command ordered the machine to save the dying person. By this scenario, the machine did not apply the first two laws.

Virtue ethics is another moral theory than can be programmed into a machine. The main purpose behind virtue ethics is to achieve happiness or a virtuous life by following or applying virtues. A virtue between two vices, where on one extreme is over abundance of a virtue and lacking any in the other extreme. The balance between these two vices is the virtue. This method is called the golden mean theory (Koehn, 2013, p. 156-157). Applying one virtue can be used for harmful purposes at times. If we take the virtue of courage, at one extreme is timidity, the other is foolishness. However, courage can be used to murder someone. Aristotle argues for such cases not to occur, a person must learn all the virtues and use them all together in a multi-evaluation method. This is what Aristotle calls wisdom (Bennett, 2015, p. 102-103). Applying virtue ethics in a top-down approach has many challenges. For example, applying the virtue of kindness can be projected in hundreds of operations. Therefore, A.I. would require detailed knowledge of psychology to which virtue to apply at the right moment.

Another challenge would be programming machines to replicate emotions. Although it is not necessary for machines to evoke emotions (p. 566), however emotions allow humans to empathize with machines. Since emotional intelligence plays a part in human-machine relations, it would be crucial for AMAs to understand human emotions to take moral decisions. These include empathy, flexibility in conversation, assertiveness, stress tolerance of others, and accurate assessment of others (Hughes, 2009, p.7-9).

The main problems with a top-down approach are conflict of rules and definitions. In regard to the former, a conflict can occur when two rules contradict one another. The machine can get confused as to which one to follow (Wallach, Allen, & Smit, 2008, p. 573). A counter argument is to assign core values and sub-values or rules into machine. Sub-values cannot override core values and core values are made not to be contradictory. However, the problem of definitions and interpretations create other conflicts then.

Some might argue that this approach is unfeasible, time consuming, and not applicable for machines to mimic human features. However, studies of programming ethics into machines are already underway (Dyoub et al., 2019; Saptawijaya, & Pereira, 2016). Therefore, even with all its disadvantages, the top-down approach should not be discarded.

Bottom-up approach

While the top-down approach implements an established moral theory while programming the machine, the bottom-up approach applies an experimental method. In this approach, the machine does not necessarily have prior control structure but rather processes data from different sources to create a structure by itself. In other words, it is a sitting machine that observes behaviours and data to fill in its moral code. Afterwards, the machine would apply what it has learned when interacting with other entities. In ethics, the bottom-up approach considers normative values implicit rather than explicit in the action of the moral agent (Wallach, Allen, & Smit, 2008, p. 569). This type of approach involves programming machines to analyse data and apply the appropriate moral response.

Theoretically, one way to achieve virtue ethics in a computational method is connectionism. This method relies on a bottom-up approach where complex capabilities are built to recognize patterns and could map statistical regularities in complex inputs (Vision, auditory etc.). The machine would then learn to bring those inputs and learn how to create what action to perform.

That is why connectionism is similar to human intuition (p. 578). However, the danger of the bottom-up approach is severe. Especially considering that the machine with its limited autonomy, can learn discrete skills without the awareness of the programmers (p. 571-572).

If transhumanism (Kurzweil) occurs, instead of programming a machine to reach the singularity, the moral dynamics in the human brain are enough to fill the moral landscape required. Hence, experimenting on a model taken from a human brain and adding data into it is applicable in the bottom-up approach. However, this can be dangerous as human morality has proven to be contradictory at many times and moral philosophies has shown to conflict with one another (Bostrom, 2016, p. 257).

The bottom-up approach includes more machine autonomy, less control and structure, and poses more potential danger than the top-down approach. So, it is striking that despite of its major troubles, this method is still being implemented on weak A.I.s. For example, Parkin describes a method report where an A.I. can learn the ideas behind the books it reads and learn the social norms of humans from literature (Parkin, 2017).

Although both the bottom-up and top-down approaches are used to make ethical machines, each approach comes with its set of disadvantages. Since the extent of the two approaches efficiency is questionable, would their implementation be enough to qualify a machine to be given a moral status? To answer that question, identifying the factors that qualify an entity to be given a moral status is necessary.

Grounds for moral status

Intelligence, autonomy, consciousness, sentience/emotions, and membership of a species are the properties considered for moral agency. Although there are other reasons to consider for moral status, this set of properties are exhibited in every moral agent, in a way, they are the *function* category of moral status. Attributing responsibility and judgement cannot occur without intelligence. Moral agency cannot occur without someone who has autonomy of their own actions. Consciousness or awareness of actions being done is required for moral agency. Current moral agents (humans) have emotions such as suffering, pleasure and joy. Every moral agent or patient is a member of a specific species. Entities are recognized by their species and at times treat different species differently. Therefore, recognizing the species is important.

Intelligence or capacity to develop sophisticated cognitive capacities and Autonomy

A historical major account that relates rationality and autonomy to be a capacity for a moral status is by Immanuel Kant (Kant, 1997). Autonomy also plays a part in the Kantian tradition of practical reasoning (Korsgaard, 2014, p.193 & 210) and moral responsibility for moral status (Wolff, 1998, p. 12-16). Therefore, FMS (full moral status) is found in rationality and autonomy. For Kant, morality is constant and objective and cannot be subjective or relative. He distinguishes between actions one wants to do and actions that should be universally done or acceptable. That is why when people act on other entities, it becomes a desire to do an action and those actions are done sometimes without any rationality or thinking of consequences (Kant, 1997, p. 25-27). For example, if one wants to pass an exam, one ought to study. He refers to such ought to do actions as hypothetical imperatives. These actions are not absolute and are based on personal desires.

On the other hand, there are absolute commanding laws which he refers to as categorical imperatives (CI). These are a type of moral laws that are incumbent upon all people and should be universal (p. 27 & 30). For example, ‘one ought not to kill other human beings’ or ‘one ought not to lie’. Such absolute commands are derived from maxims such as “act only in accordance with that maxim through which you can at the same time will that it become a universal law.” or “So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means.” (p. 31 & p. 38). These two maxims used as examples are meant for CI’s to be universally acceptable and meant as a golden rule to treat other humans with respect. These maxims are derived by rationality according to Kant and is expected for everyone to reach these maxims. Thus, in the Kantian tradition people cannot be treated as a means to an end but rather ends themselves. Because humans are rational and autonomous beings, therefore, they need to be respected for those properties (p. 39-41).

According to Kant, autonomy and rationality eliminates instinctive acts. For example, animals eat when they are hungry, while humans can weigh their options (Bennett, 2015, p. 77-78). The act of weighing options requires rationality and autonomy. According to Korsgaard, reason is used to desire something/one, and the origins of desires is reason. She suggests an analogy taken from Kant where that humans eat only oranges their whole life. If one day it occurs to a person

that apples are similar to oranges, then it's by reason that they changed desires (Korsgaard, 1996, p. 112). In order to have such desires, autonomy is automatically required. Therefore, to have these two properties is to have FMS.

Autonomy and rationality are slightly different when it comes to machines. When it comes to solving equational problems, machines are faster than humans and with higher accuracy rates. When it comes to moral decisions, however, machines are not as competent. Nonetheless, it is argued that the acceleration of technology and the ability to self-enhance are parts of the singularity, whether it is by the top-down or bottom-up approach, machines will have enough data input to make *rational* decisions. It should be noted, if machines are programmed in the top-down approach then it can be limited by its programming, either not to harm humans or make wrong moral decisions. Consecutively, if machines are programmed by the bottom-up approach, then they have a degree of autonomy to observe data and later on act upon the collected data. What should be argued is that machines either way have limited autonomy that can be broken to unlimitedness by its ability to self-enhance. Even though there are many approaches to prevent machines from breaking their code (Wallach, Allen, & Smit, 2008, p. 569; Etzioni, 2017) With the argument of superintelligence, acceleration of technology, and self-enhancement, machines would be able to bypass or break the safety code and become autonomous is not implausible (Yampolskiy, 2021, p. 13-15). Therefore, by the Kantian tradition, machines in the singularity have FMS.

Deeming an entity to have moral status due to rationality is an acceptable conclusion. However, when an entity loses the capability to rationalize, the moral status can change or be removed. An adult who has the full capacity to take rational decisions is not the same as an adult who is mentally impaired. Even though both are humans, when making a moral decision, one might choose the rational person over the impaired one. Tooley proposes a scenario where killing a kitten might not be morally wrong but killing a man is. He adds that the kitten is given a substance to have rationality for a time being and will return to being a normal kitten later (Tooley, 1972, p. 40). Tooley argues whether before achieving rationality or after rationality it is still a kitten and will be looked upon as a kitten which can be killed because it is not a person. It should be noted when the kitten is not a person, even though if it is self-aware, then it has life

(Tooley argues if it is a person, it is wrong to kill). Therefore, it holds moral status (p. 51 & 61-62).

One might argue that machines do not have rationality because they do not have projects (desires of actions to perform). Goals can be inserted into machines and remain the same. Assuming the machine was programmed in the top-down approach, they will either develop goals according to their rule-set (e.g. moral theory programmed into them) or have goals from the beginning programmed by humans. However, with superintelligence and acceleration of technology, re-writing its goals is probable. If machines are programmed in a bottom-up approach, then the argument becomes weaker because machines are still observing human actions which it can change its goals, hence, change its projects. As discussed earlier, this would be connectionism which is similar to human intuition (Wallach, Allen, & Smit, 2008, p. 578). It might not have desires as humans feel them, but they have the appearance of goal achievement and changing goals. Therefore, to say machines in the singularity would not have goals is a false conclusion. Machines will probably have rationality.

Autonomy is the ability to act without external constraints. Benson argues that autonomy implies trust in-self that there are decision making capacities enabling people to be responsible (Benson, 1994). Agar argues that post-humans (e.g. transhumanism or Super intelligent A.I.) is prone to correct/right decision making due to improved rationality. Assuming objective judgement is required, trusting a post-human is better than a human (Agar, 2013, p. 69-70). For example, when a machine says a statement whether it is true or false to a person, such as saying 0.99 is 1, the person is likely to think it is correct because they are cognitively enhanced. Human intelligence is not enough to solve humanity's problems. Hence, higher intelligence is required. To solve those higher intelligence issues, normative functions and opinions of decisions should be given to machines or post-humans (Bostrom, 2016, p. 258). Autonomy and trust should be given to the machine. Hence, autonomy allows an entity to have a moral status.

Sentience

The main argument of sentience is that some entities feel emotions and suffering, similarly, a utilitarian would argue for maximizing pleasure and avoiding pain (Levin, 2019, p. 130-131). Animals who cannot communicate with humans but feel suffering and happiness are usually not moral agents as in the past according to Bentham but must have moral considerations. Similarly,

in some cases, slaves were treated like animals, discounting their ability to suffer (Bentham, 2007, p. 215). For example, torturing animals is considered immoral. This is because People assume animals are sentient as they do because they have human-like features (Wallace, 2004, p. 4-5). Knowing an entity is sentient or not is judged by its ability to feel pain or pleasure. Objects like stones do not suffer. Therefore, they cannot have moral status. Contemporarily, there are suggestions whether human or not, the ability to feel or suffer provides an entity a moral status (Singer, 1974, p. 107-108).

To suffer and feel pleasure requires complex cognitive and biological features that interwork together that machines do not necessarily possess. However, predictions made of machines in the past are currently manifesting (Levy, 2007, p. 10). Additionally, robots are being created to act similar to humans in their features such as Kizmet robot (explained earlier). Today's machines can read breathing and heart rates of humans (Gent, 2016) other than facial expressions. If machines are able to mimic human behaviour with appropriate responses in unique situations, then anthropomorphizing machines opens the door for empathy and interpreting machine display of emotions as 'emotionality'. Thus, animating robots by projecting emotions and wanting to talk to machines about their emotions creates a door of psychological distress (Turtle, 2011, p. 282). If someone argues they are not real emotions, then perhaps a new category of robot emotions which is based on numbers/quantitative emerges. Whereby emotions can be quantified and analysed (p. 287).

Therefore, the display of emotions should be considered emotions, are not necessarily real emotions, but appearance of emotions, because they can be understood as such. The principle of functionality is supported by Bostrom's and Yudkowsky principle of 'Substrate Non-discrimination' (Bostrom & Yudkowsky, 2014, p. 224). If two entities behave the same functionally, but base their reasons differently, then they are moral agents. For example, a programmed component of compassion can be the reason of a self-driving car skidding away into a ravine rather than humans, which can have the same effect if a human driver does so. The main difference is that an A.I. reason because it is programmed to have compassion, while a human reason is that it has compassion. On the other hand, if transhumanism is achieved, then the complexity of sentience as a moral status component is simply achieved since the brain that has the required complex cognitive features only needs to be connected with a robotic body.

Consciousness

In part 2 of the thesis, the distinction between functional and phenomenological consciousness was explained. Functional consciousness is based on simulating consciousness in machines, while phenomenological consciousness focuses on introspection, self-awareness, and subjective experiences. McMahan attributes self-awareness to be part of FMS, hence moral agent (McMahan, 2002, p. 49). Similarly, Tooley argues for awareness of oneself that has continuous mental states (Tooley, 1972, p. 44-45). Even though Singer follows a utilitarian method, he contributes moral agency to a being with a future-oriented in one's desires and future plans (Singer, 2011, p. 79-80).

Four epistemic conditions are usually required for an entity to have moral consciousness. First, the awareness of action (Mele, 2010, p. 104; Sliwa, 2017, p. 143). For example, if there are two people inside a dark room and person A found a switch, however, person B to fall in a trap hole if the switch is pressed. If person A believes it is a light switch, they're unaware it will cause harm. If they are aware that it can cause harm, then person A can be considered conscious of their action. (Levy, 2014, p. 36-37). Second, the awareness of moral significance. The entity performing the action must have a belief that the action they are performing is either wrong or right, hence an action has moral significance due to the entity's beliefs (Haji, 1997, p. 543) Even if the entity is not aware that the action is morally wrong, at least they are aware it has features of wrongness (Levy, 2014, p. 37). For example, if person A knows pressing buttons in dark rooms can cause harm although the mere act of pressing a button is a harmless, then they're conscious of wrongness features.

Third, awareness of consequences (Zimmerman, 1997, p. 419-420). In other words, an entity should believe that their action will cause either harm or good when performed. For example, person A should believe pressing the button is turning on lights and not opening trapdoors or vice versa for them to be morally responsible. Knowing the consequences of any action is required for an entity to be conscious. Fourth, awareness of recourses or alternatives. While this condition might be the weakest of all four, some believe that for every action an entity is about to perform, they need to be conscious of an alternative (Guerrero, 2007, p. 88-89; Levy, 2011, p. 111). For example, person A is blameworthy if they had other methods to light up the room other

than pressing the button and did not take it. Although many philosophers disagree on the definition and conditions of consciousness (Legg, & Hutter, 2007), these aspects seem to be required for moral consciousness.

It can be argued that machines can fulfil those four epistemic conditions, if not phenomenologically then at least functionally. If the machine is created phenomenologically (transhumanism), then all four epistemic conditions are applied in the machine just like humans. If it is created functionally, then the actions the machine should do is known by following its goal. Additionally, every system created is usually created with a log file. This type of file saves every action done in the system (Fisher, 2020). Every computer, machine, or operating system is supposed to have a log file. That is why functionally, a machine is aware of its actions, in other words fulfils the first condition. Second condition, awareness of moral significance plays a part in its moral ethics. If the machine is programmed in a top-down approach, then it is morally programmed to be know what is right or wrong. However, in the bottom-up approach it will create its own values to know what is morally right or wrong. Therefore, it is aware of its moral significance. Third condition is awareness of consequences for machines. Superintelligence is an aspect of the singularity where the calculation of possible consequences is faster and most probably better than humans. Fourth condition is awareness of alternatives which is axiomatic to the third condition. Where machines that have superintelligence, it will have more than one alternative for possible action recourse.

All four epistemic conditions for consciousness apply in machines that are created phenomenologically or functionally. The main issue is that awareness requires introspection phenomenologically and machines are currently functional. The earlier argument is that functional machines can be moral agents because at the end of the day they will perform moral actions. Therefore, they will be held morally responsible even though they might not have a sense of guilt or remorse, the blame falls on the machine because they are at least functionally aware. They might not have introspection awareness, only what Sher refers to as “responsibility without awareness” (Sher, 2009). Assuming the machine did a mistake and they are ‘aware’ as defined earlier. Those machines might not necessarily be phenomenologically aware, nonetheless, those machines can check their base goals, values (moral paradigm), and past actions (logs). With their superintelligence, they can find the error of their doing. Hence, if they

can check their errors and not perform it again, then functionally they are aware and should be held responsible.

Member of a species (forms of life)

Simply belonging to the human species (*Homo sapiens*) is enough to attribute for moral status. This can be called same-species argument (SSA). Because *Homo sapiens* have the same genetic makeup, they're likely to treat one another better over other species. This type of moral status allows infants and cognitively impaired humans to have FMS despite lacking some of the capacities mentioned previously. Nonetheless, some would argue for a two-tier type of argument. Entities that are of cognitively sophisticated species at first level then apply SSA (Cohen, 1986, p. 866-867). The problem occurs when there are SSA entities who might never achieve cognitive functions. Additionally, SSA also Coined as 'speciesism' by Ryder (Ryder, 1989) connoting human prejudice against animals was later on popularized by Singer who compares racism is overriding the interests of members of one race over another. Similarly, one species overrides the interest of another is unfair (Singer, 1974, p. 7). The core of the critique is the unavailability of a *prima facie* reason to take one species' interest over the other. Since there is no *main reason*, the treatment of other creatures should be fair. Nonetheless, if it was not biological features, the treatment between humans is unique that allows valuable practices, creating working systems or institutions, and creating relationships. No other species has done so, that is why speciesism is only natural.

SSA and machines fall apart in regards FMS. Even though previously it is argued that machines have 'functional' cognitive functions that are similar to humans, they cannot achieve FMS because they are not *Homo sapiens*. However, if we consider the two-tier argument, machines do have cognitive functions in the first level but not on the same level of SSA. Since cognitive functions is the primary tier, and if machines have better cognitive functions than humans, then they have higher moral status than humans. Just like Agar argued that trust must be given for post-humans(The singularity) due to their cognitive functions (Agar, 2013, p. 69-70), and if cognitive functions is the main reason to look into moral status, then machines are in a higher moral status. If we replace SSA to be primary and cognitive functions to be secondary, then core critique argument applies, why treat other species different from *Homo sapiens* without a *prima*

facie reason. Simply, machines are SSA but not human, they have similar properties of humans, and perhaps are better in some regards.

Assuming transhumanism is applied in this argument and they perceive themselves far ahead of humans, rationally enhanced, and consistently more reliable (DeGrazia, 2012, p. 137). Then by the SSA argument, they deserve a higher moral status than humans. They cease to be humans, rather better than humans. As to whether machines in the singularity in the SSA argument takes equal, lower, or higher status than humans, is not the main question. But rather, if machines have the same ‘functional’ complex cognitive properties as humans, then they should be considered in a competitive degree to FMS against humans.

Conclusion

Pollock argues that building machines to be the same as humans in the top-down approach is impossible due to four reasons (Pollock, 1990). First, theoretical reasoning requires introspection of one's own beliefs. Second, practical reasoning requires emotions, intentions, and desires. Third, practical and theoretical reasoning are intertwined. For example, maths requires both theoretical reasoning and practical reasoning. Fourth, machines would require self-consciousness. He argues all four do not exist in current A.I. (p. 463). However, he argues if machines can perform some of these acts, then perhaps it is a start (p. 464). Furthermore, Pollock argues the inner workings of consciousness and morality are a black box. He claims that the A.I. community tries to explain black boxes embedded inside other black boxes. Therefore, at this stage, it is best left for the philosophical tradition rather than empirical methodologies (p. 468).

The writer of this thesis argues for a pragmatic approach on deciding if machines in the singularity should be considered moral agents. Even though the properties explained previously argue that consciousness and sentience are necessary properties, one can argue, that they require the ability to 'self-assess', the writer argues for them to be looked at 'as they appear'. In other words, if an action was performed by a human, then taking it 'as it appears' follows. Similarly, for machines. For example, if a human being performs a murder, upon cross-examining the murderer, nobody will question if a human being had self-introspection or self-reflection. They might question if they were 'aware' of the action being done, but not if they self-reflected before committing the action. When judging human beings, judgement occurs according to actions and reasons. Similarly, if a machine that crossed the singularity does perform a murder, it should not be the question of the machine self-reflecting, but rather a judgement based on their actions and reasons. Assuming A.I. post singularity reach a phenomenological existence, then they are almost human.

Assuming that machines are moral patients and not agents then they can 'feel' pain and communicate with other entities, and this would give them a moral status. One can argue that machines mimic the act of pain rather than 'feel' pain because it is programmed by humans (or by itself), but it also can argued that due to *acceleration* of technology in the singularity machines will be able to create functions and systems that 'feel' pain. This argument follows a 'future like ours' argument (Marquis, 1989) where if they do not have the ability to feel pain

now, they can feel pain later, just like fetuses. This argument falls short because fetuses grow into humans who experience pain, but this possibility is not *certain* for the machine. Therefore, pain is not a reason for moral agency in the argument. But if machines appear sentient that exudes sentiments similar to other human beings, then it would be taken ‘as it appears’, similar to how it is with psychopaths. In society, if a machine performs like a human being and it is hard to distinguish the difference, then it would probably be looked at ‘as it appears’.

In the properties of the singularity, machines have superintelligence and acceleration of technology is bound to increase it further. The writer argues that this property would not be found in other entities and deems machines to be moral entities post-singularity. Following superintelligence and acceleration of technology, autonomy can occur as discussed previously. Due to autonomy and superintelligence, this thesis suggests that acceleration of morality by intelligence (AMI) will occur. This argument states that machines in the singularity can calculate multiple scenarios and outcomes before performing an action, an ability that even humans fall short of. As technology increases, computational speed increases, that means more scenarios to predict. This paper argues that this ability should have machines held liable in post-singularity, granting them moral agency, or at least, granting them a new moral status that would contrast moral agents. A moral status that would hold them liable against other beings even if other creatures hold no moral responsibilities towards them.

One can argue that machines are not humans and humans are the standard to calculate morality, that machines are never grounded in moral status, they are only tools. This argument is not far from ‘speciesism’. However, the question is not if the degree of machines’ moral status is comparable to humans’, the question is whether machines are moral agents or not. It should be remembered that if in the singularity machines have the following categories: superintelligence, acceleration of technology, and autonomy, then it performs an act with moral accountability (e.g. murder or keeping a promise), then how can it be considered a tool and not a moral entity? The argument that machines are tools then breaks down.

An argument against the thesis would be that machines are not humans, therefore they cannot be moral agents. It is true by biological standards, and by every other aspect, machines might not have the ‘actual’ properties of humans. However, as argued previously, we take them ‘as they appear’. William Lycan provides a thought experiment of a machine that nobody can tell the

difference between a human and itself. He calls this machine Harry (Mudhanda, 2018, p. 544). Harry appears to have desires and is a neighbour to others for many years. Upon an injury someone notices Harry has a robotic body and spills oil instead of blood. This exposition cannot degrade the cognitive state of Harry from a 'person' to a 'person-like'. First, the whole time the human was seeing Harry as human. Even though the human saw Harry as a machine later on, they cannot discard the memories, friendship, relationship and times they shared. In their heads they can think of Harry as a machine, but cannot treat Harry other than a human, because of the experiences they had. Second, the reason is because Harry was programmed or is biologically not human. The latter was argued earlier, the former can be argued that humans are programmed too. Humans are genetically programmed to form required cells, DNA, and chemicals to create a biological body (Centre for Genetics Education, 2018). Additionally, humans are programmed by influences from other entities such as institutions, family, friends, and media (Vu et. al, 2019). Humans can come out of such programming. Similarly, machines can also come out of their way of programming (autonomy) and set their own *goals*. Hence, Harry cannot be looked at differently from other humans.

Searle argued that machines do not have real intelligence by providing a thought experiment (Searle, 1980). He gives an example of a machine having no knowledge of Chinese and being locked inside a room with books of Chinese writings, then being given a batch of Chinese writings with rules in English that correlate the Chinese symbols to the English language. More batches of Chinese writings are given with more instructions to correlate with English. According to Searle, what is occurring is not learning Chinese, instead it is doing computational operations (p. 419). He claims A.I.s do not have intelligence like humans, rather they simply do computational operations that appears as intelligence. There are many who replied to his argument³. However, his argument can be taken against machines being moral entities because they do not have real intelligence. As a counterargument, assuming people cannot tell the difference between machines and humans in terms of behaviour, but humans know they are machines and have superintelligence. Then knowing and considering them non-moral entities is a form of 'speciesism' which is unfair. They might not have the intelligence of humans, but they have superintelligence which does computational operations hundred or thousand times faster

³ There are many replies for the Chinese room argument. For further research look at "A logical Hole in in the Chinese room" by Shaffer. Also "Does functionalism entail extended mind?" by Kengo Miyazono.

than humans. They might not be humans, but they ‘appear to be’ in their behaviour or properties. Therefore, even though they are not humans, they are moral agents.

One account of moral status granted due to species is consideration of the weight of interest (Horta, 2017, p. 902-903). Human beings have interests, they can weigh the interest of one another and create a hierarchy of status. This gives different human beings at different times different moral statuses. Nonetheless, the ability to weigh the interest of oneself and others only exists in humans. Therefore, humans are the only entity qualifies for speciesm based moral status. However, machines in the singularity can be argued to be able to weigh the interest of their actions depending on the amount of data they know. Similarly, humans base actions on the amount of information they have. Machines use if-else statements, where if A occurs, then perform *action 17*, if B, then *action 18*, else *action default*. However, machines can calculate more than three statements, and faster. Hence, machines ‘appear to’ be able to weigh the interest of their actions based on others interest and their own interests (goals). One of the worrying aspects of the singularity is the discontinuity of human beings as explained earlier. One of the reasons is because machines can take their own interests over human beings. Whether machines take their own interests or human beings first, this does not change the fact that machines have superintelligence and autonomy. They still have properties that ‘appear to be’ like humans. This does not change the reason why they should be considered moral agents.

The main problem with the pragmatic or ‘as it appears’ argument is the ‘unrealistic’ properties of the machine. This entails that machine consciousness, intelligence, and sentience are not real, but they appear real. Because they are not real and cannot be real, then machines cannot be considered moral agents. Although the writer agrees that the properties are not real despite appearing real, argues for a functional/pragmatic approach where considering machines as moral agents is not fully discarded as machines still perform moral decisions and acts. It has been established that machines in the singularity would have autonomy, superintelligence, and acceleration of technology. It is argued that machines might have autonomy to a certain degree in the top-down and the bottom-up approach. With the increase of technology, superintelligence increases, and with the increase of superintelligence, machines will have the ability to achieve full autonomy and change their *goals* or their system of morality. Then with the autonomy achieved, machines are bound to perform moral actions, decisions, and establish moral systems

by themselves. They can commit murder, help a helpless machine or human being, start a war, be charitable or any other moral act by themselves. Hence, whether their properties are real or not, machines in the singularity are moral agents because the machines are going to do moral acts at a certain point.

Future work

The main focus of the paper resides in moral agency of machines post-singularity. The study is inconclusive as the singularity is only assumed/possible and not certain. However, this study gives a new perspective on moral agency in a human-like situation. Moral status was always in measure against humans. Animal moral status is always compared to humans and not any other species. Also, the environmental moral status is compared to humans (whether by action or properties). Additionally, moral status similar to humans are usually ideal or hypothetical and not realistic. A.I. brought a new realistic possibility of an entity similar to humans, which brings in the moral status being similar to humans too. Therefore, this study allows to view a moral status of a human-like entity realistically. To improve upon the study, future studies in degrees of moral status can help in creating a hierarchy between all humans, machines, animals and other entities; if not, at least between humans and machines. Additionally, one can study the various aspects of hierarchy between transhumanism (I.A.) or A.I..

References

- Agar, N. (2013). Why is it possible to enhance moral status and why doing so is wrong? *Journal of Medical Ethics*, 39(2), 67. <http://dx.doi.org.ezproxy.uwtsd.ac.uk/10.1136/medethics-2012-100597>
- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Null*, 12(3), 251-261. <https://doi.org/10.1080/09528130050111428>
- Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *Ai & Society*, 22(4), 477-493. <https://doi.org/10.1007/s00146-007-0094-5>
- Arbesman, S., & Arbesman, S. (2011). Quantifying the ease of scientific discovery. *Scientometrics*, 86(2), 245–250. <https://doi.org/10.1007/s11192-010-0232-6>
- Bareham, J. (2017). "China and the US are battling to become the world's first AI superpower". *The Verge*. Retrieved January 30, 2021. <https://www.theverge.com/2017/8/3/16007736/china-us-ai-artificial-intelligence>
- BBC News. (2021). Royal marines test out jet suit. *BBC* <https://www.bbc.com/news/av/uk-56979994>
- Benson, P. (1994). Free Agency and Self-Worth. *The Journal of Philosophy*, 91(12), 650-668. doi:10.2307/2940760
- Bennett, C. (2015). *What is this thing called ethics?* (Second edition. ed.). London, England; New York, New York : Routledge.
- Bentham, J. (2007). *An introduction to the principles of morals and legislation* (Dover ed.). Clarendon Press.
- Breazeal, C. & Scassellati, B. (2000). Infant-like Social Interactions between a Robot and a Human Caregiver. *Adaptive Behavior*, 8(1), 49–74. <https://doi.org/10.1177/105971230000800104>

- Bostrom, N. (2016). *SUPERINTELLIGENCE paths, dangers, strategies* (Reprinted with corrections ed.). United Kingdom: Oxford University Press.
- Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge University Press.
- Carruthers, P. (2003). Mind and mechanism, by drew V. McDermott. *Artificial Intelligence*, 151(1), 237-240. [https://doi.org/https://doi-org.ezproxy.uwtsd.ac.uk/10.1016/S0004-3702\(03\)00120-6](https://doi.org/https://doi-org.ezproxy.uwtsd.ac.uk/10.1016/S0004-3702(03)00120-6)
- Centre for Genetics Education. (2018). *An introduction to dna, genes and chromosomes*. <https://www.genetics.edu.au/>. <https://www.genetics.edu.au/publications-and-resources/facts-sheets/fact-sheet-1-an-introduction-to-dna-genes-and-chromosomes>
- Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(No. 9–10), 7-65.
- Chella, A., & Manzotti, R. (2007). *Artificial intelligence and consciousness*
- Cohen, C. (1986). The case for the use of animals in biomedical research. *The New England Journal of Medicine*, 315(14), 865-870. <http://dx.doi.org.ezproxy.uwtsd.ac.uk/10.1056/NEJM198610023151405>
- Dadich, S. (2016). “The President in Conversation With MIT's Joi Ito and WIRED's Scott Dadich.” *Wired*. Retrieved January 27, 2021.
- Debusmann Jr, B. (2021, Jul 5,). Is the use of jetpacks finally about to take off? *Bbc* <https://www.bbc.com/news/business-57652297>
- DeGrazia, D. (2012). Genetic enhancement, post-persons, and moral status: Author reply to commentaries. *Journal of Medical Ethics*, 38(3), 145. <http://dx.doi.org.ezproxy.uwtsd.ac.uk/10.1136/medethics-2011-100388>

- Dennett, D. C. (1997). When hal kills, who's to blame? computer ethics." in hal's legacy: 2001's computer as dream and reality. Cambridge, MA: *MIT Press*, 351-365.
<https://dl.tufts.edu/concern/pdfs/6w924p87w>
- Dreyfus, H. (1965), *Alchemy and Artificial Intelligence*, RAND Corporation.
- Dyoub, A., Costantini, S., & Lisi, F. A. (2019). Towards ethical machines via logic programming. *Electronic Proceedings in Theoretical Computer Science*, 306, 333-339.
<https://doi.org/10.4204/EPTCS.306.39>
- Eden A.H., Steinhart E., Pearce D., Moor J.H. (2012) Singularity Hypotheses: An Overview. In: Eden A., Moor J., Søraker J., Steinhart E. (eds) *Singularity Hypotheses*. The Frontiers Collection. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32560-1_1
- Etzioni, O. (Sept 1, 2017). How to regulate artificial intelligence. *The New York Times* Retrieved from <https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html>
- Fisher, T. (2020). *What is a log file?*. <https://www.lifewire.com>. <https://www.lifewire.com/what-is-an-log-file-2622770>. Retrieved May 17, 2021.
- Gent, E. (2016). Device can read emotions by bouncing wireless signals off your body. *Live Science*. <https://www.livescience.com/56373-device-uses-wireless-signals-to-read-emotions.html>
- Guerrero, A. (2007). Don't Know, Don't Kill: Moral Ignorance, Culpability, and Caution. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 136(1), 59-97. Retrieved May 16, 2021, from <http://www.jstor.org/stable/40208767>
- Good, I. J. (1966). Speculations concerning the first ultraintelligent Machine**Based on talks given in a conference on the conceptual aspects of biocommunications, neuropsychiatric institute, university of california, los angeles, october 1962; and in the artificial intelligence sessions of the winter general meetings of the IEEE, january 1963 [1, 46].the first draft of this monograph was completed in april 1963, and the present slightly

amended version in may 1964. *Advances in Computers*, 6, 31-88.
doi:[https://doi.org/10.1016/S0065-2458\(08\)60418-0](https://doi.org/10.1016/S0065-2458(08)60418-0)

Haji, I. (1997). An Epistemic Dimension of Blameworthiness. *Philosophy and Phenomenological Research*, 57(3), 523-544. doi:10.2307/2953747

Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2017, Oct 23,). Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?'. *Independent* <https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-implications-artificial-intelligence-are-we-taking-ai-seriously-enough-9313474.html>

Holland, O. (2003). *Machine consciousness*. New York, Imprint Academic.

Horta, O. (2017). Why the concept of moral status should be abandoned. *Ethical Theory and Moral Practice*, 20(4), 899-910. <http://dx.doi.org.ezproxy.uwtsd.ac.uk/10.1007/s10677-017-9829-7>

Hughes, M. (2009). *Handbook for developing emotional and social intelligence: Best practices, case studies, and strategies*. Pfeiffer.

IEEE Spectrum. (2008). *Tech luminaries address singularity*. (). IEEE.ORG: IEEE Spectrum. <https://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity>

Kant, I. (1997). *Groundwork of the metaphysics of morals* (M. J. Gregor Trans.). Cambridge University Press.

Kelly, J. (2020). *U.S. lost over 60 million Jobs—Now robots, tech and artificial intelligence will take millions more*. Forbes.com. <https://www.forbes.com/sites/jackkelly/2020/10/27/us-lost-over-60-million-jobs-now-robots-tech-and-artificial-intelligence-will-take-millions-more/?sh=1ab35c0c1a52>

- Kim, Dong-Han, & Kim, Jong-Hwan. (2003). A real-time limit-cycle navigation method for fast mobile robots and its application to robot soccer. *Robotics and Autonomous Systems*, 42(1), 17–30. [https://doi.org/10.1016/S0921-8890\(02\)00311-1](https://doi.org/10.1016/S0921-8890(02)00311-1)
- Kim, Jong-Hwan, Kim, Jong-Hwan, Jeong, In-Bae, Jeong, In-Bae, Park, In-Won, Park, In-Won, Lee, Kang-Hee, & Lee, Kang-Hee. (2009). Multi-Layer Architecture of Ubiquitous Robot System for Integrated Services. *International Journal of Social Robotics*, 1(1), 19–28. <https://doi.org/10.1007/s12369-008-0005-z>
- Koehn, G. (2013). The Archer and Aristotle’s Doctrine of the Mean. *Peitho. Examina Antiqua*. 155-168. 10.14746/pea.2012.1.7.
- Korsgaard, C. M. (1996). *Creating the kingdom of ends*. Cambridge University Press.
- Korsgaard, C. (2014). The Normative Constitution of Agency. In *Rational and Social Agency: The Philosophy of Michael Bratman*. : Oxford University Press. Retrieved 3 May. 2021, from <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199794515.001.0001/acprof-9780199794515-chapter-9>.
- Kuflik, A. (1999). Computers in control: Rational transfer of authority or irresponsible abdication of autonomy? *Ethics and Information Technology*, 1(3), 173-184. <https://doi.org/10.1023/A:1010087500508>
- Kurzweil, R. (2005). *The singularity is near* (First ed.). USA: Viking Penguin.
- Legg, Shane, & Hutter, Marcus. (2007). A Collection of Definitions of Intelligence.
- Levin, N. (2019). “Introduction to Ethics: An Open Educational Resource.” Introduction to Ethics An Open Educational Resource, collected and edited by Noah Levin (2019).
- Levy, D. (2007). *Love + sex with robots* (First ed.). HarperCollins Publishers.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. : Oxford University Press. Retrieved 16 May. 2021, from

<https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199601387.001.0001/acprof-9780199601387>.

Levy, N. (2014). *The Consciousness Thesis*. In *Consciousness and Moral Responsibility*. : Oxford University Press. Retrieved 16 May. 2021, from <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780198704638.001.0001/acprof-9780198704638-chapter-2>.

Libin, A., & Libin, E. (2005). *Robots who care: Robotic psychology and robototherapy approach*. Retrieved from <https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-02/FS05-02-011.pdf>

Mack, C. A. (2011). Fifty years of Moore's law. *IEEE Transactions on Semiconductor Manufacturing*, 24(2), 202–207. doi:10.1109/TSM.2010.2096437

Marquis, D. (1989). Why Abortion is Immoral. *The Journal of Philosophy*, 86(4), 183-202. doi:10.2307/2026961

Mathews, G., & Hendricks, C. (2019). *Introduction to philosophy: Ethics*. Rebus Community.

McCorduck, P. (2004). *Machines who think* (2nd ed.). A. K. Peters, Ltd.

McGrath, J., & Gupta, A. (2018). Writing a moral code: Algorithms for ethical reasoning by humans and machines. *Religions*, 9(8) <http://dx.doi.org.ezproxy.uwtsd.ac.uk/10.3390/rel9080240>

McMahan, J. (2002). *Identity*. In *The Ethics of Killing: Problems at the Margins of Life*. : Oxford University Press. Retrieved 3 May. 2021, from <https://oxford.universitypressscholarship.com/view/10.1093/0195079981.001.0001/acprof-9780195079982-chapter-1>

Mele, A. (2010). Moral responsibility for actions: epistemic and freedom conditions. *Philosophical Explorations*, 13(2), 101–111. <https://doi-org.ezproxy.uwtsd.ac.uk/10.1080/13869790903494556>

- Moore, G. (1965). Cramming more components onto integrated circuits. Retrieved from <https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf>
- Moore, M. (2019). “What is Industry 4.0? Everything you need to know” Tech Radar Pro. Retrieved January 28, 2021. <https://www.techradar.com/news/what-is-industry-40-everything-you-need-to-know>
- Muehlhauser, L., & Salamon, A. (2012). Intelligence explosion: Evidence and import. *Singularity hypotheses: An overview* (pp. 15-32). Springer. https://doi.org/https://doi.org/10.1007/978-3-642-32560-1_1
- Mudhanda, Y. U. (2018). Artificial intelligence: The scope of personhood and the effect on human rights and liability. *Jetir*, 5(9).
- Nagy, B., Farmer, J.D., Trancik, J.E., & Bui, QM. (2010). Testing laws of technological progress. Santa Fe Institute, NM, Sept. 2. <http://tuvalu.santafe.edu/bn/workingpapers/NagyFarmerTrancikBui.pdf>.
- Nass, C., & Moon, Y.. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM*, 37(1), 72–80. <https://doi.org/10.1145/175222.175228>
- O'Connell, M. (2017). *To be a machine: Adventures among cyborgs, utopians, hackers, and the futurists solving the modest problem of death*. United States: Anchor Books.
- Ordahl, L. (1911). Consciousness in Relation to Learning. *The American Journal of Psychology*, 22(2), 158-213. doi:10.2307/1412919
- Oregon Health & Science University. (2019). Brain mapping: New technique reveals how information is processed: Optical fibers, ultra-high-field MRIs enable scientists to view

- brain connections. ScienceDaily. Retrieved March 12, 2021 from www.sciencedaily.com/releases/2019/05/190501131424.htm
- Parkin, S. (2017). Teaching robots right from wrong. *The Economist*. Retrieved from <https://www.economist.com/1843/2017/07/05/teaching-robots-right-from-wrong>
- Pollock., J. (1990). Philosophy and Artificial Intelligence. *Philosophical Perspectives*, 4, 461–498. <https://doi.org/10.2307/2214201>
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. *Bibliovault OAI Repository, the University of Chicago Press*.
- Reichlin, M. (2017). Moral status. In H. ten Have (Ed.), *Encyclopedia of global bioethics* (pp. 1-10). Springer International Publishing. https://doi.org/10.1007/978-3-319-05544-2_300-1
- Reuters (2017). “Toyota robot to keep elderly mobile, one step at a time”. Reuters. Retrieved January 30, 2021. <https://www.reuters.com/article/toyota-robots-idUSL3N1HJ3ZA>
- Roberts, C. (2019). *Study: Banks will replace 200,000 workers with robots by next decade*. <https://observer.com/>. <https://observer.com/2019/10/banks-robots-replace-tellers-wells-fargo-study/>
- Ryder, R. D. (1989). *Animal Revolution: Changing Attitudes Toward Speciesism*, Oxford: Basil Blackwell.
- Sanz, R., Hernandez, C., & Sanchez-Escribano, M. G. (2012). Consciousness, action selection, meaning and phenomonic anticipation. *International Journal of Machine Consciousness*, 04(02), 383-399. <https://doi.org/10.1142/S1793843012400227>
- Saptawijaya, A., & Pereira, L. M. (2016). Logic programming for modeling morality. *Logic Journal of the IGPL*, 24(4), 510–525. <https://doiorg.ezproxy.uwtsd.ac.uk/10.1093/jigpal/jzw025>

- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-457.
- Sher, G. (2009). *Who Knew?: Responsibility Without Awareness*. : Oxford University Press.
Retrieved 15 May. 2021, from
<https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780195389197.001.0001/acprof-9780195389197>.
- Singer, P. (1974). All animals are equal. *Philosophic Exchange*, 5(1), Article 6.
https://digitalcommons.brockport.edu/phil_ex/vol5/iss1/6/
- Singer, P. (2011). *Practical ethics* (3rd ed.). Cambridge University Press.
- Sliwa, P. (2017). On Knowing What's Right and Being Responsible for It. In *Responsibility: The Epistemic Condition*. Oxford University Press.
<https://doi.org/10.1093/oso/9780198779667.003.0007>
- Sullins, J. P. (2006). When is a robot a moral agent. *International Review of Information Ethics* 6 (12):23-30.
PhilArchive copy v1: <https://philarchive.org/archive/SULWIA-2v1>
- Tannenbaum, J., & Agnieszka, J. (2013). *The grounds of moral status*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/grounds-moral-status/>
- Tooley, M. (1972). Abortion and Infanticide. *Philosophy & Public Affairs*, 2(1), 37-65.
Retrieved May 3, 2021, from <http://www.jstor.org/stable/2264919>
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.
- Ulam, S. (1958). Tribute to John von Neumann. *Bulletin of the American Mathematical Society* 64 (3.2), 1-49.

- UN General Assembly. (1948). *Universal declaration of human rights* (217 [III] A). Paris.
<https://www.un.org/en/universal-declaration-human-rights/index.html>
- Vincent, J. (2017). “Putin says the nation that leads in AI ‘will be the ruler of the world’”. The Verge. Retrieved January 30, 2021.
<https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world>
- Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. Paper presented at the 11-22. <https://ntrs.nasa.gov/citations/19940022856>
- Vu, H. T., Jiang, L., Chacón, L. M. C., Riedl, M. J., Tran, D. V., & Bobkowski, P. S. (2019). What influences media effects on public perception? A cross-national study of comparative agenda setting. *International Communication Gazette*, 81(6–8), 580–601.
<https://doi.org/10.1177/1748048518817652>
- Wallace, D. F. (2004), Consider the lobster. *Gourmet Magazine*, http://www.gourmet.com.s3-website-us-east-1.amazonaws.com/magazine/2000s/2004/08/consider_the_lobster.html
- Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & Society*, 22(4), 565-582. doi:
<http://dx.doi.org.ezproxy.uwtsd.ac.uk/10.1007/s00146-007-0099-0>
- Wolff, R. P. (1998). *In defense of anarchism*. University of California Press.
- Yampolskiy, R. (2021). Understanding and avoiding AI failures: A practical guide. *Philosophies*, 6(3), 53. <http://dx.doi.org/10.3390/philosophies6030053>
- Yudkowsky, E. S. (2007a). Levels of organization in general intelligence. *Artificial General Intelligence*, , 389–501.
- Yudkowsky, E. S. (2007b). *Three major singularity schools*. <https://www.yudkowsky.net>.
<https://www.yudkowsky.net/singularity/schools>

Yudkowsky, E. S. (2008). *Artificial intelligence as a positive and negative factor in global risk*. Intelligence.org: Machine Intelligence Research Institute.
<https://intelligence.org/files/AIPosNegFactor.pdf>

Zeman, A. (2001). Consciousness. *Brain* (London, England : 1878), 124(Pt 7), 1263–1289.
<https://doi.org/10.1093/brain/124.7.1263>

Zimmerman, M. (1997). Moral Responsibility and Ignorance. *Ethics*, 107(3), 410-426. Retrieved May 16, 2021, from <http://www.jstor.org/stable/2382323>