On The Danger of Artificial Intelligence

Saba Samiei

A thesis submitted to Auckland University of Technology in fulfilment of the requirements for the degree of Master of Computing and Information Sciences (MCIS)

July 2019

School of Engineering, Computer and Mathematical Sciences

Abstract

In 2017, the world economic forum announced that AI would increase the global economy by USD 16 trillion by 2030 (World Economic Forum, 2017). Yet, at the same time, some of the world's most influential leaders warned us about the danger of AI. Is AI good or bad? Of utmost importance, is AI an existential threat to humanity? This thesis examines the latter question by breaking it down into three sub-questions, is the danger real?, is the defence adequate?, and how a doomsday scenario could happen?, and critically reviewing the literature in search for an answer. If true, and sadly it is, I conclude that AI is an existential threat to humanity. The arguments are as follows. The current rapid developments of robots, the success of machine learning, and the emergence of highly profitable AI companies will guarantee the rise of the machines among us. Sadly, among them are machines that are destructive, and the danger becomes real. A review of current ideas preventing such a doomsday event is, however, shown to be inadequate and a futuristic look at how doomsday could emerge is, unfortunately, promising!

Keywords: AI, artificial intelligence, ethics, the danger of AI.

Acknowledgements

No work of art, science, anything in between or beyond is possible without the help of those currently around us and those who have previously laid the foundation of success for us. I hence believe that no such work should be published without acknowledging people who have played a key role in enabling the creators. As such, I would like to take this opportunity and show my gratitude to those without whom this research would not have been possible.

I want to thank my supervisor Dr Albert Yeap, for his continuous support and guidance. Thank you for always encouraging me to break the boundaries, think outside the box and write with confidence. Your support in showing me the right path, asking me questions and leading me to ask more questions has had a tremendous impact on the viewpoint I have developed as part of this research.

I want to thank my mother, Frough Samiei, for always encouraging me to continue learning. You are the most exceptional mentor and the best friend I have in life. Thank you for teaching me patience, consistency and self-belief by being a fantastic role model to look up to. Your words, wisdom and insightful guidance have always helped shine a light on any path I have chosen to follow. This is a gift to you, and I hope to have made you proud.

I want to thank Dawie Olivier, Mike Burk and Craig Young for the time they spent to read my thesis and share their insights to strengthen my research. Your invaluable feedback has helped me improve my work and build my viewpoint confidently. I also would like to thank Liz Gosling and once again, Mike Burk for their mentorship throughout my journey. Your guidance and advice of discipline and patience have contributed tremendously in my ability to bring this piece of work to an end. I hope to have made you all proud.

Lastly, I would like to thank all the readers of this thesis and mention that when I started my masters, researching the future of AI, I knew there would be tears. However, I thought those tears would be for all the late-night studies and the fear of missing deadlines, not even once I thought the tears would be as a result of searching the history of human intelligence!

Writing this thesis has led me to acquire the tenacity and the desire to do my part in changing the future. I wish, sincerely, that one day, you as a reader of this thesis can hold

your head up and say that you read the words of someone who contributed to changing the world.

Contents

Attestation of authorship 1

Thesis structure and research methodology 2

Introduction 4

Chapter 1: Is the danger real? 7

1.1 Robotics 7

1.1.1. The need-based era 7

1.1.2 The curiosity era 9

1.2. Machine learning 13

1.2.1. The benign discovery era 13

1.2.2. The impactful era of strategising 18

1.2.3. When we are no longer in charge 19

1.3. How various industries have reacted to artificial intelligence 21

1.3.1. Investments 21

1.3.2. Reactions 23

1.4. Conclusion of chapter 1 25

Chapter 2: Is the defence adequate? 27

2.1. What has been previously proposed to address the danger to date? 28

2.1.1. Self-monitoring AI 28

2.1.2. Friendly AI 29

2.1.3. Introducing punishment 31

2.1.4. Legalisation 32

2.1.5. Other approaches 33

2.2. Are we doing enough to make sure AI is safe? 34

2.2.1. How to know how much attention needs to be paid to AI ethics, regulation and health and safety? 35

2.2.2. Are popular forums involved? 45

2.3. Conclusion of chapter 2 47

Chapter 3: On the road to annihilation 49

3.1 The event horizon of technological singularity – our transition 50

3.1.1. The first phase of our transition, AI trust and dependency 50

3.1.2. The second phase of our transition, robotic body parts and body bots 52

3.1.3. The third phase of our transition, cyborgs 54

3.1.4. The fourth phase of our transition, cloud-based brain 56

3.1.5. The fifth phase of our transition, cloud-based consciousness 57

3.2. Scenario 1- A doomsday with unconscious and soulless machines 58

3.2.1. The bright side of humans living alongside super intelligent AI 58
3.2.2 How we will allow the machines to kill us 62
3.3 Scenario 2- Why can doomsday still happen with replicated human consciousness? 63
3.3.1 What we can achieve if AI is the next evolution of humans 64
3.3.2 The dark side of the ultimate singularity 68
Thesis summary and conclusion 71
Future work 74
Bibliography 76

Attestation of authorship

I, Saba Samiei, hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

14 July 2019

_____ _____

Signature Date

Thesis structure and research methodology

Every work of research is unique in its own way. Hence, while there might be commonalities between utilised methodologies of research, they also need to be tailored or combined to fit the purpose of every piece of work. In my thesis, I have reviewed the danger associated with artificial intelligence (AI in short), from different angles, with each chapter focusing on a different question. As such, I have adopted different methods to answer each question. Below is a summary of the thesis structure and the research methodologies used in each chapter.

CHAPTER 1 – Creswel (2009) believes, a qualitative research method is one using which the researcher reviews previous research, literature, theories etc. to develop a theme from the collected data. In my first chapter, I address the question, "Is the danger real?". This is a question that can only be answered by showing a pattern of events, the impact of those events and the reactions (positively or negatively) to the resulted situation.

While there is some quantitative data (e.g. market cap of technology companies), the primary methodology adopted for this chapter is a qualitative review of published literature. The quantitative data in this chapter is to support the argument about the pattern of behaviour that is contributing to the danger of AI and is not data collected as a result of qualitative research. Chapter 1 includes an in-depth study of previously published journal articles, books, online magazines, case studies, websites etc. The results of the research have been discussed in the chapter as well as presented in the form of graphs and tables to provide a pictorial view of my analysis.

CHAPTER 2- The second chapter addresses the question, "Is the defence adequate?". This chapter consists of two parts. In the first part of this chapter, I discuss "What has been previously done to address the danger of AI?". This question also requires a review of previously published efforts; therefore, I have adopted a qualitative research methodology here. This section includes a literature review of published work which addresses the concern of the danger of AI.

Part 2, however, addresses the question, "Are we doing enough to make sure AI is Safe?". This is a question that needs to be quantified to provide a relevant answer. Simply claiming our efforts are enough or not enough is not a measure based on which we can plan for our future. We need to measurably define 'enough', understand if there is a gap

and how big it is, to address the right issues and propose a useful solution. Hence, a qualitative review of literature is not a suitable method to address this question.

In his paper, Creswel also talks about quantitative or postpositivist research as one that is based on data collected from surveys or experiments (Creswel, 2009). This section requires a data-based approach to fulfil its purpose. Therefore, in the second section of chapter 2, I have conducted a quantitative experiment based on the number and the subject of previously published papers. In this chapter, I have expanded my research to other fields (nuclear energy and healthcare), to provide a measurable view of the collected data. I have demonstrated the experiments' results in the form of graphs and tables followed by arguments that address chapter 2's question.

CHAPTER 3- This chapter focuses on our pathway to annihilation and addresses the question, "Why we allow doomsday to happen?". It includes three sections; the first section focuses on the path that we will follow to transform from biological to digital beings. The last two sections, include predicting what the future with AI will look like both positively and negatively.

This chapter aims to provide a prediction to possible future scenarios that humans will experience. Predictions in this thesis are made based on a qualitative review of the previous and current pattern of events and global reactions. These predictions do not include a timeline or a measure of good or bad and hence are qualitative in nature; similar to that of the data that they are derived from.

The certainty of every prediction is dependent on the reasoning of predicting person (i.e. abductive(guaranteed), inductive (probable) and deductive (best guess)) (Copi, Cohen, & Flage, 2016). The reasoning behind the predictions in this thesis is inductive. I do not claim that such scenarios are certain and will happen by a predicted date. However, I also do not believe my predictions are the best guess based on the comparison between the previous and current pattern of data in AI compared to other fields. Hence my reasoning in this thesis is inductive, and by definition, I claim that such outcomes are probable, based on the current pattern of data.

Introduction

"The history of artificial intelligence is a history of fantasies, possibilities, demonstrations, and promises" (Buchanan, 2005)

"All men by nature desire to know" Aristotle

The concerns about the impact of artificial intelligence (AI in short), on the future, vary from adverse economic consequences (Hanson, 2008), to the total extinction of human race (Bostrom, 2003), (Barrat, 2013). The late Dr Stephen Hawking has warned humans that "the development of artificial intelligence could spell the end of the human race," entrepreneur Elon Musk has indicated that we are "Summoning the demon" (Trecker, 2019), writer and filmmaker Jame Barrat calls AI our final invention (Barrat, 2013). On the other hand, optimists like the futurist Ray Kurzweil believe the technological singularity is unavoidable and talk about a more positive future where AI and human intelligence are indistinguishable and the two live side by side in a better world (Kurzweil, 2005). And some like the computer scientist Roman Yampolskiy believe that AI is the next evolution of humans, and we should not try to stop it (Yampolskiy, 2013b).

In recent years, discussions and warnings have been abundant about the potential threat of AI. There is an increasing number of movies, novels, TEDx talks, interviews, panel discussions and forums around the world to address various questions about the danger of AI.

Even though this topic has been receiving much attention in recent years, the history of automating repetitive human work and the replication of human intelligence goes back hundreds of years (Luger, 2006). And, the history of the term artificial intelligence as we know it today, dates back to 1956 when John McCarthy held the first AI conference at Dartmouth College, in Hanover, New Hampshire (Smith, McGuire, Huang, & Yang, 2006). Some early achievements to demonstrate "intelligence," planted the seed for the first AI Summer. An AI Summer is a period where there are significant investments in AI research and development. These examples include ELIZA (Weizenbaum, 1966) or a program that played Checkers (Samuel, 1959). This first summer soon came to an end as the critics argued that machines would never be as intelligent as humans. Moreover, some of the early prediction of AI failed to come true.

In the early 1990s, AI experienced a second summer when industries realised the possibility of scaling through automation. The programs by then didn't have to be perfect; they only had to be better than humans. There was a shift of approach from symbol manipulation algorithms to connectionism. Connectionism is an approach that was developed as a result of an attempt to understand the human brain, in particular, how the brain learns and remembers. However, significant limitations in demonstrating a noise-free, optimally classified data brought the second AI winter.

Excitingly, we are now experiencing the third AI summer, one that has started by powerful technology companies as opposed to researchers and academia. Microsoft, Google, IBM, Amazon and Facebook have promised life-changing improvements in the technology. Better jobs, faster lives, transformed workplaces, smarter devices, better data management and insights, innovative solutions, smart cities, driverless cars, better workplace health and safety, more accurate medical diagnosis, are but a few of these promises. However, there is very little attention to how this will impact future generations and human species as a whole. In this thesis, I investigate three main questions to understand why artificial intelligence is perceived as a dangerous technology. It is essential to pay attention to why this fear exists as the answer can help us cover some of our blind spots and help in how we shape our future.

In the first chapter, I raise the question "Is the danger real?". There is a lot of talk about how destructive AI is but very little explanation to address "why?". Why is AI receiving so much attention, 60 years after its birth? Why is this summer any different than the previous ones? I investigate the improvements that have to lead us to this state of fear. I review the different eras in Robotics from the development of robots that could only perform one task poorly (e.g. Goliath, the first military robot) to robots that can walk, run, communicate (e.g. Sophia) or shoot guns (e.g. FEDOR). I then review the improvements made in machine learning and milestones that have been achieved to create a state of fear. Lastly, I research the skewed industrial investment made in AI. I believe the combined impact of these three factors is what is feeding the real fear about AI.

In the second chapter, I ask the question "are we doing enough to prevent doomsday?". To do so, I have conducted a review of previously introduced approaches which were proposed to ensure humans stay the dominant species. I then share the results of the research I have undertaken to investigate if we are investing enough to make sure our technology is safe. This research includes a review of over 7000 academic papers and a

thorough search of the largest academic database, Scopus. The goal of this chapter is to measure the amount of ethical, health and safety and regulatory focus in AI, understand if we are doing enough or if there is a gap.

In the third chapter, I discuss a pathway to annihilation and raise a final question "Why do we allow doomsday to happen and what does it look like?". Why would we let such a future happen to us if we are gazing into a crystal ball that shows us the future?. I begin the chapter by discussing a 5 step process that I believe humans will follow to merge with their technology and evolve to fundamentally different entities. I discuss two future scenarios depending on whether we will be able to complete the last step of this process, i.e. replicate our consciousness into the digital world. In both scenarios I start by outlining what the bright side of this transition looks like as I believe the attractiveness of this bright side is what leads us to break the safe boundaries and allow the pendulum to swing, resulting in two possible doomsday scenarios which will extinct our species.

The intention of this thesis is not to provide a silver bullet to the doomsday problem but to ask the right questions. Questions based on a future scenario backed up by facts, with the hope to contribute to increasing the focus of the research done on unintended consequences of AI. I believe we are in a pivotal point in history and hold the key to either open the door that leads us to dominate the universe and tick off the items on our wish list, or we can open the door that sets us on the path of destruction and misery.

Chapter 1: Is the danger real?

The first step to address the concerns about the future of AI is to understand why they exist in the first place. What type of achievements, developments or events have contributed to such warnings? Should we be worried about AI and why? Why have these concerns increased 60 years after the birth of the field?

The underlying question that this chapter will address is "Why is artificial intelligence warned about as a dangerous technology?". To address this question, I have studied different events in the history of AI that have led us to the current state of fear. I have identified three pillars, namely robotics, machine learnings and huge and skewed industrial investments that are contributing factors the ever-increasing concerns about AI. In this chapter, I have attempted to identify where to look and what to look for when it comes to AI being discussed as a dangerous technology.

1.1 Robotics

An essential subject of artificial intelligence is robotics, a technology that mimics not only human ability to perform physical tasks like grabbing and moving objects but also human behaviour, including but not limited to simulating our speech, vision and learning ability. This field is particularly important due to its ability to combine software and hardware to create autonomous and mobile agents that are human-like.

In the following section, I have researched different milestones of the field, leading to its current state. I have divided the field of Robotics into two eras, the **need-based era** and the **curiosity-based era**. The need-based era is when most of the development of the subject was to address a need, e.g., shortage of human labour. The Curiosity era is when most of the growth is more to create new skills in robots that don't always address a current or future need and instead, are to explore the possibilities.

1.1.1. The need-based era

Even though we can argue that the root of robots dates back thousands of years ago when man started to create artificial limbs (Dellon & Matsuoka, 2007), signs of the use of robotics technology as we know it today has roots in dangerous tasks (e.g. mining and working in nuclear power plants), challenging (e.g. industrial welding) or impossible (e.g. in-depth space exploration or deep-sea exploration) for humans to perform. Like many

other technologies (the internet is an example), the earliest signs of robotics can be found in the military. In 1932, while the U.S. had invested in industrial robot development, Germany and the Soviet Union started researching autonomous weapons leading to their first invention, Goliath that could carry up to 200 pounds of explosives (Coll, 2004).

The development in the field for the next three and a half decades (the 1930s to mid-1960s) was limited to mimicking physical human tasks. E.g., Raymond Goertz's teleoperated mechanical arm in 1951, used in nuclear reactors (Springer, 2013). George C. Devol, Jr. and Joseph F. Engelberger's Unimate, in 1961, the first industrial robot made to unload high-temperature parts from a die casting machine (Stone, 2005). American Machine and Foundry (AMF) Corporation's Versatran in 1963, a programmable cylindrical robotic arm capable of grabbing and moving objects and Norway's Trallfa in 1966, a robot capable of spray painting wheelbarrows made to address a shortage of labour that year in its place of birth (Stone, 2005).

Even though the name robot (meaning forced labour) was suggested back in 1920, by a Czech play writer (Cook, 2016), and as mentioned above, using the concept commenced not long after that. The first significant milestone in the field was not achieved until a few decades later with the birth of the first "intelligent" robot (i.e., a robot that can make decisions), SHAKEY, made by Dr.Charles Rosen and his research team between 1966 to 1972. SHAKEY was created to mimic specific 'intelligent' tasks, like panning, rearranging simple objects and route finding (Stone, 2005). The development of SHAKEY is said to be a turning point in the world of technologies like autonomous cars and military drones (Markoff, 2017).

The next four decades (the mid 1960s to end of 1990s) were spent on 1) improving "intelligent" tasks (i.e. the robots ability to perform functions that involved some form of decision making), e.g., Stanford Cart that crossed a room full of obstacles with no human intervention in 1979 (Stone, 2005) (Earnest, 2012); 2) replacing robots with human labour, e.g. Unimate assembly line made by Nissan-Japan in 1971 (Stone, 2005); and 3) expanding the robotics technology in different fields like astronomy and healthcare, e.g. Viking 1&2, space crafts with robotic arms that were sent on a mission to Mars by NASA in 1975, and Robodoc in 1992, an FDA approved robot that performed hip replacement surgery (Stone, 2005).

While the advancements in this era led to the making of military robots like Goliath and smart robots like SHAKEY, the majority of the impact was contributed towards the industrialisation with robots like Unimate enabling shorter timeline and higher quantity of product manufacturing. As most of the developments in the early decades of robotics were made to address a common need, I have named this era, the need-based era. The majority of the improvements in this era were helpful[1], and robotics was being used as a *tool* with little to no trace of a danger to human species. However, human curiosity and hunger for learning and inventing have been the force behind the never-ending advancements made in different fields. In the following section, I will discuss the impact of this human characteristic on how a different era of robotics is shaped. One that warrants many questions and lays a foundation of fear.

1.1.2 The curiosity era

The major turning point in the world of robotics that put an end to the need-based era is when a robot named ASIMO (born in 2002), started to walk like humans. Unlike its predecessors, Honda P2 (born in 1996) and Honda P3 (born in 1997), ASIMO was able to perform complex physical tasks, e.g. running, walking up and down the stairs, opening a bottle and pouring its content into a cup without spillage (Chestnutt et al., 2005).

Since ASIMO, intelligent software combined with versatile hardware have led to the creation of more human-like robots able to perform human-like capabilities. For example, the ability to appear, act and behave like a human (e.g. Sophia, a robot created by Hanson Robotics limited), ability to "learn" (i.e. compute autonomously) a map of the environment it's in (Durrant-Whyte & Bailey, 2006) and ability to "reason" (i.e. produce a plan based on the input data) and perform complex tasks (e.g. STAR (Shademan et al., 2016)).

Figure 1. shows the rapid growth of robotics, particularly after the creation of Asimo, the concerning element is the speed of growth in the area, alongside the production of certain robots that have led to a dystopian concern in the society. For example, in 2017, Russia introduced a humanoid robot called FEDOR (Final Experimental Demonstration Object Research). FEDOR is capable of driving, working out and using power tools (Galeon, Futurism, 2017a) (O'Conner, 2017). While the Russian authorities claim that FEDOR is

[1] I have mentioned *the majority* of the inventions and improvements in this era were helpful as I question the benefit of using robotics for military purposes. However, as mentioned, robotics has a root in the military and hence its involvement in military would have been unavoidable.

made to travel to space in 2021 as a single operator of its spaceship, we can argue that specific capabilities of FEDOR, e.g. shooting two guns at the same time and injecting syringes don't match its purpose.

In the same year, Strato Energetics introduced a concept of a military drone called "The Stinger," a mass-produced, palm-sized drone that flies itself, has a wide field camera and has face recognition capability. The prototype shows a drone which carries 3 grams of explosives and provides just enough power to penetrate the skull with surgical precision (Stratoenergetics, 2018). While a weapon like The Stinger does not yet exist, there is a considerable amount of international research and investments being done on armed and unarmed drones.

A 2017 policy choice document proposed to President Trump's administration suggests that more than 30 countries in the world have armed and more than 90 countries around the globe have unarmed drones (Catalano Evers, Fish, & Horowitz, 2017). While the document does not specify the purpose for which these drones are made for, it is a sign of significant growth in operator-less flying intelligent machines.

Some of the researchers and industry professionals[2] believe that some of these concerns are mainly due to the AI hype that has been created by the media. However, there is no doubt that automated armed machines used for military purposes and introduction of mobile humanoid robots equipped with intelligent software which are physically more powerful than humans can impose a danger. To ensure we create a safe future, we need to pay serious attention to the depth of the danger that unintended consequences of such developments can cause and start to address these before its too late.

In figure 1, I have marked different milestones in the field of robotics, showing the two different eras of need-based and curiosity-based. The graph indicated an increase in the production of robots since the beginning of curiosity era. It also shows the capabilities of robots improve from being able to carry a few pounds of explosives in the war field (Goliath) to the creation of robots that perform human-like tasks and are taught to make human-like decisions (e.g. FEDOR working out, driving or shooting guns) and introduction of concepts like killer drones equipped with facial recognition (The stinger).

[2] Industrial professionals that I have personally spoken with in different forums, e.g. New Zealand AI Forum events, The AI Day, AI panels hosted by different companies etc.

This rate of improvement in the skills that these mobile entities can perform is the first pillar of the danger that robotics (a subset of AI), imposes.

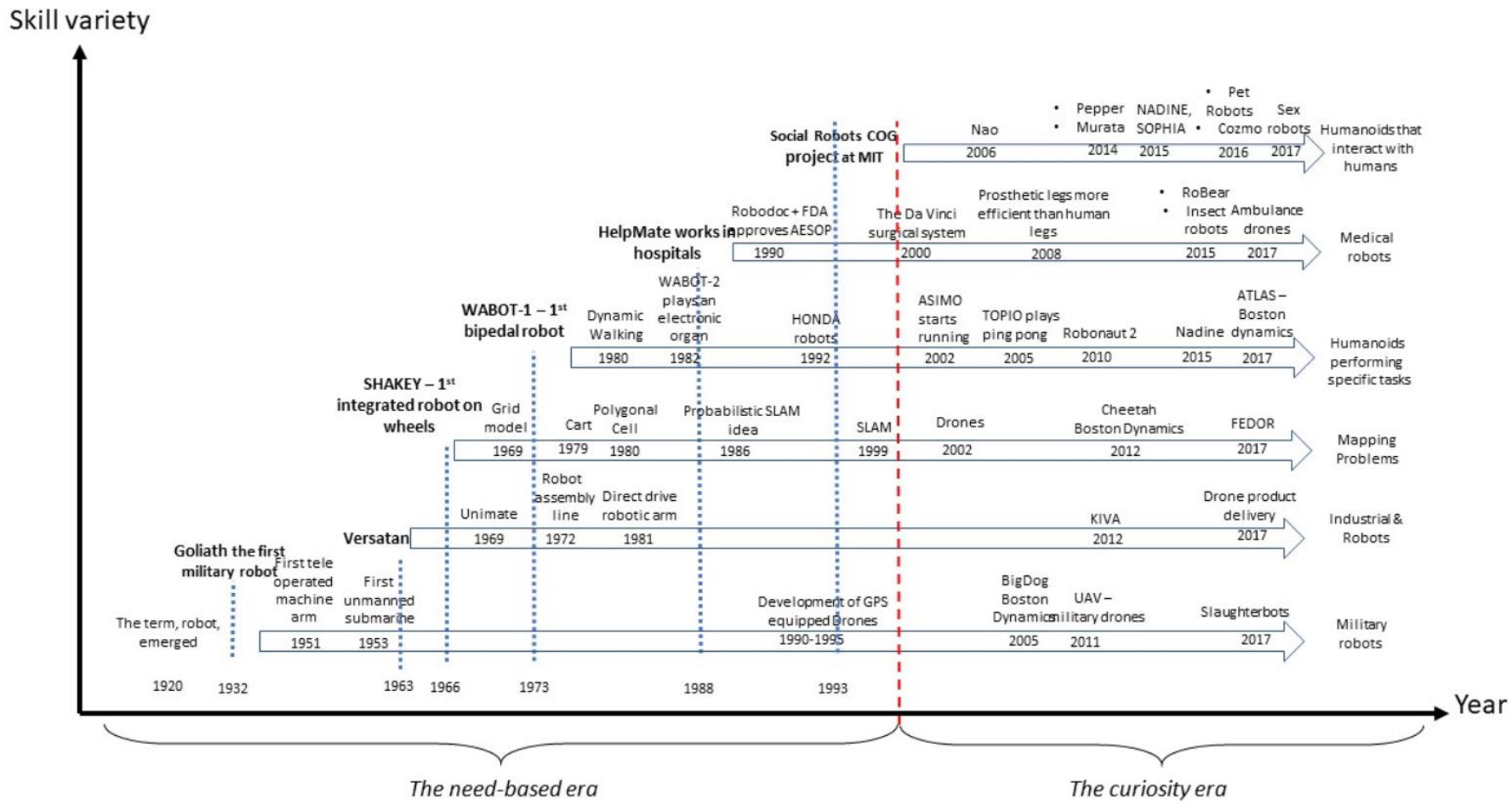

Figure1. The rapid growth of a variety of skills performed by robots and the two proposed eras.

1.2. Machine learning

The physical engineering of robotics on its own, would not have been able to achieve some of the field's milestones like the creation of robots that can navigate (e.g., SHAKEY), mimic the ability of speech, visions and interaction (e.g., Sophia) or those which perform complex physical tasks (e.g., ASIMO). The science of machine learning is what can provide the soft dimension, i.e. intelligence (ability to make decisions) to what would otherwise be an immobile and unintelligent piece of hardware.

Every subset of AI mimics a specific dimension of living things, in particular, humans. Machine Learning is a field that tries to emulate the brain, specifically its ability to "learn" from its environment without being "programmed". The term was coined in 1959 by Arthur Samuel. He defined machine learning as: "the field of study that gives computers the ability to learn without being explicitly programmed" (Puget, 2016).

In this section, I have researched the advancement of machine learning and have highlighted major milestones from start to date. I have divided this era into three different phases, the first one being '**The benign discovery era**', where most of the development and achievements are benign. This means they have an abundance of flaws and are specific (i.e. they can only perform one task) hence, considered far from being dangerous or enabling a state of danger in the immediate term. However, these milestones have built the foundation to move us to the next phase, which I have called '**The impactful era of strategising**'. This is an era where the machine capabilities both in terms of hardware and software, have superseded human expectation of what AI can do. An era wherein a large amount of data can be gathered, analysed, and insights can be shown within seconds. When our machines prove to us that creativity is not a concept limited to humans. An era in which creation gives us the power to decide how the future is shaped, as well as make our vulnerabilities clear to us. After this era, we will enter a future '**when we are no longer in charge**', and what happens here (discussed in more depth in chapter 3), will depend on the decisions we make today.

1.2.1. The benign discovery era

Even though the term 'machine learning' was coined in 1959 (Samuel, 1959), the concept as it stands today, dates back to 1950 when Alan Turing posed the question 'Can machines think?" (Saygin, Cicekli, & Akman, 2000). In 1951, Marvin Minsky and Dean Edmonds created Stochastic Neural Analog Reinforcement Computer (SNARC for short), the first

Neural Network machine that had the ability to learn (i.e. be given a set of inputs and automatically calculate the output) by being trained (i.e., the operator would press a button programmed as a reward for every correct answer). It had a capacitor which worked as memory and helped the machine remember for a short period (Ramos, Augusto, & Shapiro, 2008).

On the quest to replicate the human brain's ability to learn (i.e., gather inputs and autonomously produce outputs based on what it experiences) gaming seemed to have been a popular tool. Gaming includes decision making (i.e., choosing a path to achieve an outcome), has tangible and short-term rewards (i.e. a clear result of winning or losing), limited input data points (e.g. number of players, number of squares on the board or pixels on the screen, number of pieces on the board, etc.) and can have infinite paths to get to the output, e.g. the game of Chess.

While the concept of using gaming to learn how the brain "learns" has remained the same, the approaches to performing experiments have been different. In 1959 Arthur Samuel used an Alpha-Beta Pruning, a search tree of the board positions combined with a scoring function that calculates the probability of winning for each player at any point in time (Knuth & Moore, 1975). He combined this with Rote learning, a learning algorithm that enables the machine to learn by repetition, to create the world's first machine that played the game of Checkers (Samuel, 1959).

In 1981 Gerald Dejon introduced the concept of Explanation-based learning, an algorithm that trains the machine by providing training examples and allows the computer to create a rule to eliminate irrelevant data (DeJong & Mooney, 1986). The algorithm introduced the base for programs to play Chess (Thrun, 1995).

The next decade passed without significant milestones however the improvements made following this "Machine Learning winter", show a shift in the focus of the researches from *knowledge-based* development (i.e., using statistical knowledge) to *data-based* development (i.e., machines that learn from analysing a large amount of data).

In 2006 the concept of Deep Learning was proposed by Geoffrey E. Hinton (Google, 2018a), this concept replicates the way a human brain creates different levels of representation from sensor input data (Hinton, 2007). The idea is the continuing of work in the field of Multi-Layer FeedForward networks introduced in 1965 (Schmidhuber, 2015). The interest in the area came back to life when in 2006 a group of researchers from

Canadian Institute for Advanced Research (CIFAR) introduced an unsupervised learning algorithm that created multiple layers of features without the need for data labelling. The critical characteristic of Deep Learning is that these layers are **generated by the machine** by using a general-purpose learning algorithm as opposed to being engineered by humans (LeCun, Bengio, & Hinton, 2015). The shift in knowledge to data-based development combined with the improvements in Deep Learning led to a fast recovery of machine learning from its winter.

This recovery was followed by some significant breakthroughs in the field. In 2009, AT&T won the $1m Netflix prize by creating an improved version of Netflix's algorithm to recommend users' favourite movies (Buskirk, 2009). In 2010 Microsoft used advanced machine learning in a gaming device named 'Kinect' to recognise human bodies, allowing users to interact with the machine using their body movements (Han, Shao, Xu, & Shotton, 2013).

While the research has expanded to different fields, the use of gaming as a mechanism to research learning continues. In 2011, IBM Watson won the Jeopardy game against two other humans using different reasoning algorithms (Guizzo, 2011) (High, 2012). In 2013 a company called DeepMind bought by Google, used a reinforcement learning algorithm to teach a machine how to play Atari games above human proficiency (Minh et al., 2013).

In 2015, Google Deep Mind created AlphaGo which beat the world's best player of the ancient Chinese game Go, named Lee Sedol. The importance of this game was the fact that AlphaGo's opponent Lee Sedol, was a nine-done professional of Go (the highest rank in Go ranking system). He is also known as the world's most creative player of Go and creativity was a concept that everyone thought machines were not capable of before this game (Ross, 2016).

The match was broadcasted live and was viewed by 18 million people. As one of the reporters at the scene mentioned, "Lee Sedol had always been playing for his team or his country, but this time, he is playing for humanity". The anticipation and the worry of viewers were beyond the concern for a machine beating a human at a game; this had happened previously with the game of Chess and Checkers. The combined distress and joy were because *AlphaGo had taught itself* how to play the game, and none of its makers knew how to play the game well (watch Kohs, 2017).

The experiments in the field of machine learning have not always proven to be positive milestones or ones with no implications on their outside world (i.e., people's day to day lives). The challenge with the concept of machine learning is that for computers to learn, a significant set of input data is required. The problem is, our history has not proven human's keen ability to build an unbiased foundation of historical data. e.g., ProPublica's assessment of the U.S. Courts' Risk Assessment Algorithm, which helps the judges make a decision about the sentences of the offenders, in 7000 cases, proved that the program was biased toward black defendants, giving them longer and harsher sentences to that of white defendants (Life, 2017) (Angwin et al., 2016). Recently in 2018, Amazon created an algorithm to assist with its recruiting process, it did not take long for the team to realise that the algorithm was demonstrating a bias towards male candidates (Dastin, 2018).

The common issue between both the above scenarios is the source of data that was fed to the algorithms to train them (i.e., program them to generate the desired output based on the learnings of the data the algorithm is provided). The available data in the world currently is based on human history and biases. These algorithms teach us what these biases are and the impact they have had on society in a much faster and blunt way. The danger is not what has happened previously, but it is the repetitive behaviour of the developers despite the undesirable results every time human historical data is used.

For a machine to be built and work, it needs a purpose, a goal or a final state to achieve. However, if there is not enough research about what these algorithms will do to achieve their desired outcome, the end outcome might be what the human programmers ask but *how* the machines get to the result, can have disastrous consequences. One of the characteristics that we will take with us is the fact that humans are smart-*ly* lazy (Kaplan, 2015). We try to find the most comfortable and most efficient solutions to the obstacles and limitations around us. The technology we have developed is evidence of this claim. We no longer have to remember phone numbers, learn locations or remember dates as our smartphones can do all of that for us. As such, the algorithms they develop seem to replicate this characteristic with the difference that our programs get to the outcome much more efficiently and unpredictably than we do. E.g., a simulated robot that was programmed to develop legs to get from point A to B in the quickest way decided assembled itself into a tall tower and tumbled instead of developing legs and run (Shane, 2018).

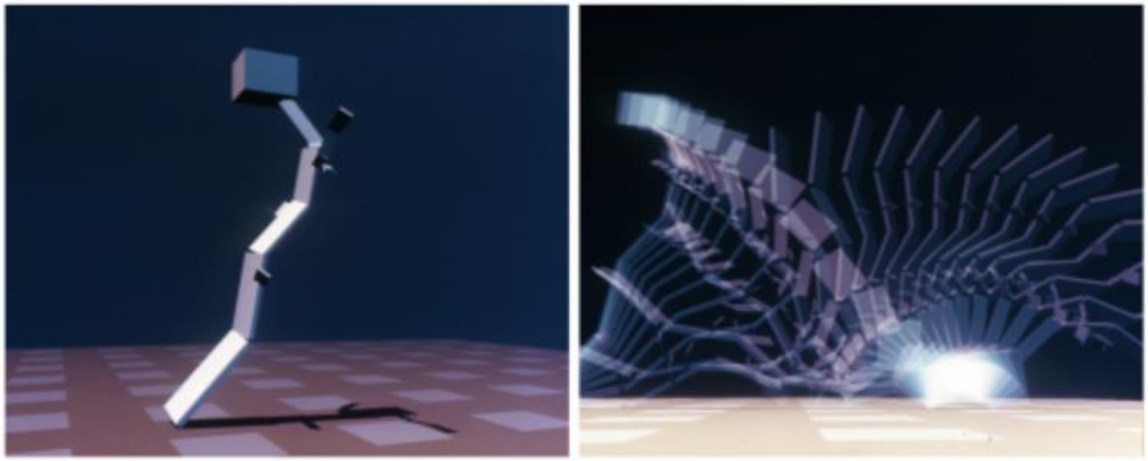

Figure 2: Algorithm that assembled itself into a tower instead of evolving legs (Shane, 2018).

On May 6th 2010, a 9 per cent drop in the stock market, resulted in temporary evaporation of USD 1 trillion in assets due to the source code of the automated programs trading stock on behalf of their companies. The issue was that the software was designed by developers using sophisticated models of historical data, which meant the programs couldn't predict the present and the impact of equally advanced opposing programs. The economists named this phenomenon "Systemic risk" (Kaplan, 2015).

Another example of an algorithm finding an unpredictable and in this case, an undesirable outcome is an algorithm that was programmed to find the best way to apply the minimum force to an aircraft landing on a carrier. Instead, the programme applied the maximum force, resulting in its program's memory overflow and registering minimum force. If used to the real world, such an algorithm would crash the plane and kill its pilot, but the end outcome would record minimum force (Shane, 2018). There are many more examples of intelligent algorithms that have surprised their programmers by finding the laziest way to achieve the outcome (Shane, 2018).

While some of these experiments are done in controlled environments and are shut down before they are used, and stronger regulations have been imposed to prevent some of the failures when they are used widely outside of an experiment zone. However, regulation, if imposed to restrict the usage of a technology or a tool, might only lead to illegal usage of it. The question is, is intelligence enough for the development of a learning machine that will be responsible for handling such life dependable tasks (e.g. flying a plane with passengers, driving cars, armed military robots etc.). Or should we start investigating the

replication of different dimensions of what makes us humans? E.g. intuition (the ability to make decision instinctively and without multiple data points which may not be available), intellect (ability to understand abstract concepts and using this understanding to make decisions) and ability to understand and care about the implications of actions on the surrounding environment? Should the regulation for AI be different from how we have regulated our industries so far, given the differences in AI with all previous developments?

1.2.2. The impactful era of strategising

One of the most talked-about fears of AI is that the improvement in the field can be such that humans no longer can have control over the technology. Stephen Hawking, Elon Musk and Bill Gates are only a few names that have announced their concerns about AI to the world. However, as discussed above, intelligence on its own is not the primary issue, in this regard, the fear is getting to the point that no human will be able to understand what is being processed on a machine's "mind". This phenomenon is no longer a theory for the future; in 2017, Facebook Artificial Intelligence Research Lab (FAIR) created deal negotiating chatbots that were designed to learn using machine learning algorithms. Reports show that the bots were good at negotiating; for example, they would show interest in something that had no value to use the same thing later as leverage. After some time, these bots started to create their own way of communicating using a language that was not comprehensive by humans and was taken down soon after witnessing this concerning behaviour (Griffin, 2017).

A new language is not the only creation of artificially intelligent systems; the next phenomenon is one that moves us to the next Era of Machine Learning which I call "The impactful area of strategising". In December 2017, Google introduced NASNet to the world, a "child" created by their AutoML project. AutoML (the parent), is the controller of the neural network. NASNet's challenge was to recognise objects of a live video, with AutoML evaluating and providing feedback on NASNet's performance, a process that was repeated a 1000 time and resulted in NASNet being 1.2% more accurate and 4% more efficient than state-of-the-art human-created algorithms. This breakthrough is said to have improved machine's ability to "see" (i.e., recognise objects in the real world) which can be used to better advance technologies like self-driving cars (Sulleyman, 2017) (Galeon, 2017b).

The current applications of programs like NASNet are not the point of concern, rather the direction towards which this breakthrough can take us. Having an artificially intelligent system that creates another system not just better than humans, but faster is something we can take advantage of to better our future. Now more than ever, we can share our workload with our inventions. The concern is, if like robotics, machine learning enters a curiosity era, humans can lose control over their creation. This way, a phenomenon like the Facebook negotiating machine happens outside of a controlled environment (like that of the U.S. stock market) and this time human developers may not be able to get into the root of the problem and eliminate the cause fast enough to prevent unknown and unintended consequences.

In figure 3, I have shown different eras of machine learning, indicating our current position.

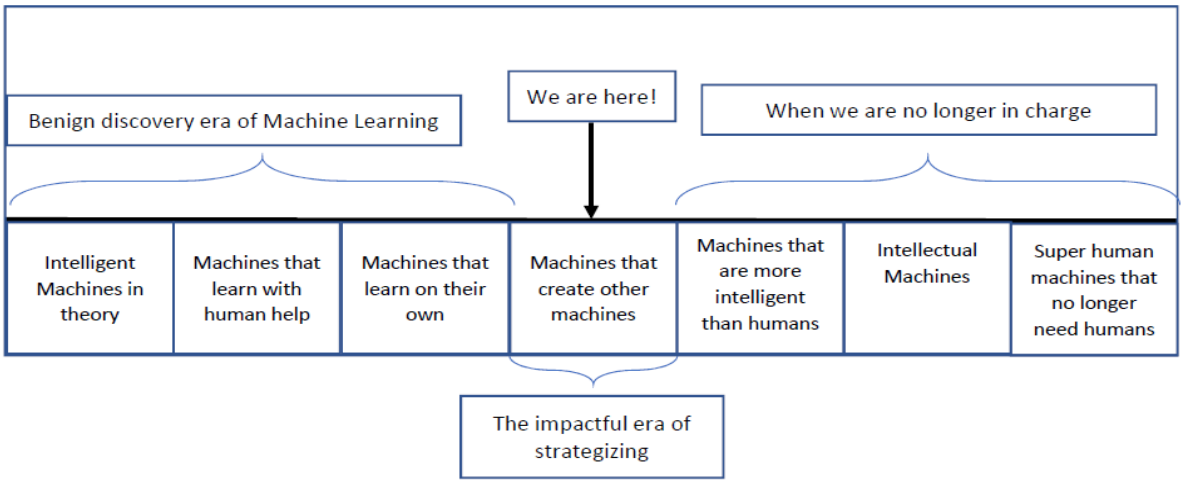

Figure 3: Machine learning eras

1.2.3. When we are no longer in charge

Predicting the future can be attempted based on historical and present data, the pattern that the data shows and the probability of certain events repeating or being created. The challenge is, in today's world, our data is changing so rapidly that makes it very difficult to calculate all the probabilities of what the future could hold. The third era being 'When we are no longer in charge' is an era that belongs to the future, this is the era when we would have reached the AI singularity, i.e. the point that AI will supersede the human intelligence (Linstone, 2012). A future where humans and machines will be indistinguishable.

While Stephen Hawking and Elon Musk have warned the world that this would be the end of humanity, optimist and futurist Ray Kurzweil believes otherwise. He defined singularity as the point where machine intelligence is infinitely more than humans. That any technology follows the law of accelerating returns (an extension of Moor's law which states that any evolutionary system[3] follows an exponential growth pattern), he believes that when we reach the point of singularity, human and machine intelligence will merge (Kurzweil, 2005).

Currently, we have machines that are better and faster than humans but mainly in one task only, e.g., writing an algorithm, playing games or solving maths problems. This is called artificial specific intelligence (ASI in short), i.e. machines that are good at solving one specific task. But we will reach a state where a machine is better in more than just one task, i.e. artificial general intelligence, (AGI in short), i.e. human-level intelligence. Kurzweil believes we will develop AGI by 2029 (Kurzweil, 2005).

However, Kurzweil's prediction timeline can be challenged based on the current state of AI behaviour. The current state of AI carries with itself a contradiction, i.e. intelligent machines that are not intelligent enough to comprehend the impact of their decisions on their surrounding world. More fundamentally, they are not yet intelligent enough to learn about the world without external intervention.

Such contradiction results in the development of two opposing viewpoints; one is that an AI system which is unable to understand the consequences of its actions can be a great danger to the humans around it. Such systems have already left a scar in the lives of many (e.g. the scenario of using an AI system in American courts). The second viewpoint is driven from the fact that AI is currently dependant on humans to learn and improve, this has meant that some of the AI researchers (personal communications, November 7, 2018) believe that AI will never get to a point where they can impose a danger due to its dependence on humans.

However, if AI, like any other evolutionary system, follows Kurzweil's law of accelerating returns this AI contradiction will soon be an issue which leads humans to

[3] Ray Kurzweil believes that technological products demonstrate an exponential growth like that of an evolutionary system. However, their growth has is substantially faster than any other evolutionary system and cannot be predicted by Moor's law. Hence, he introduces the law of accelerating returns (Kurzweil, The Singularity is near, 2005).

enter an irreversibly altered future at which point they will have limited to no control over AI systems.

1.3. How various industries have reacted to artificial intelligence

1.3.1. Investments

It is apparent that in today's world, no advancement in any field can be made without significant financial, time and labour investment. AI is no exception to the rule. In fact, the decrease and increase in AI investment is what has created the terms "AI winter" (i.e. a period of reduced funding which as observed in previous sections has led to decreasing achievements in the field); and "AI Summer" (i.e. an opposite phenomenon where the funding increases as the result of breakthroughs that are disruptive either in terms of the core technology or a simple change in the business models). However, attracting fund and attention towards any particular topic is hard without tangible and proven benefits. In this section, I have reviewed the speed of growth of top tech companies, what they have achieved with AI and the impact that their advocation has had on the growth of AI.

According to Wikipedia.org, the top 10 largest internet companies in the world include (in market cap descending order), Alphabet, Amazon, Tencent, Facebook, Alibaba, Netflix, Booking.com, Baidu, Salesforce and JD.com. Among these, the first five belong to the exclusive $500B plus USD market cap club along with Microsoft and Apple. The highest power that these companies hold is the large amount of data they collect from their users and the way they utilise AI, analysing them and navigate their business to higher success. For example, using AI algorithms in targeted digital advertising, Amazon Audiable using AI bots to negotiate prices with customers who are about to unsubscribe, Netflix, an online entertainment company, uses machine learning algorithms to recommend favourite movies to its subscribers, and Facebook has started using AI to help blind people "see" photos posted on Facebook (by describing the images, e.g. people, smiling, outdoor, etc.) and to learn more about their users (Chowdhry, 2016).

Table 1 shows the phenomenal growth rate of some of these companies involving AI. For example, the plummeting value of most of the major US retailers (e.g., SEARs, NORDSTROM, JCPenney, etc.) against the ever-increasing market cap of Amazon is a reliable indicator that Amazon alone can destroy the traditional retail business (Digg, 2017).

Companies like Airbnb and Uber have also disrupted the traditional business models by selling products that they neither own nor produce. Even though it is their innovative business models as opposed to AI, which is the primary key to their success, AI has played a massive part in growing these businesses. As Goodwin (2015) mentions, the battle is for the customer interface. The recent event of Cambridge Analytica's use of Facebook users' information and its involvement in the US government election, showcases the power one could gain by analysing the publicly available information, using AI.

Company	2008-2009 Market cap in $B USD	2018 Market cap in $B USD	The growth rate over the last ten years
NVDIA	7.6	129.8	1708%
Netflix	9.1	125.4	1378%
Tencents	57.3	483.2	843%
Facebook	61.55	457.8	744%
Amazon	116.3	681.1	586%
SalesForce	24.9	84.6	340%
Microsoft	238.7	695.1	291%
Google	256.7	702.2	274%
Booking.com	35.9	98.1	273%
Baidu	31.5	76.6	243%
Jd.com	27.4	55.8	204%
Alibaba	221.5	430.9	194%

Table 1. The growth rate of high-tech companies that use AI. Source (YCharts, 2018)

1.3.2. Reactions

The different achievements in the field of robotics and machine learning combined with the fast-growing, new and disruptive companies and their use of AI has attracted a lot of attention to this technology. It has also raised a lot of questions among all industries. Will AI take over human jobs? More importantly, will AI improve enough to take over and later exterminate humans? Currently, the excitement over AI has led, an unprecedented scale, all industries from primary (i.e., mining and forestry) through to quaternary sectors (i.e., technology companies and R&Ds) have gotten together to discuss the challenges, impact and the future of AI. For example, in New Zealand various Meetup groups have been formed to address AI, small companies like New Zealand.AI and The AI forum have joined forces with big corporates like IBM, DataCom, banks, etc. to host different AI events.

Figure 4, shows the timeline of different industry sectors having invested in AI with the more recent years being a time that all industry sectors have made a significant investment in the field.

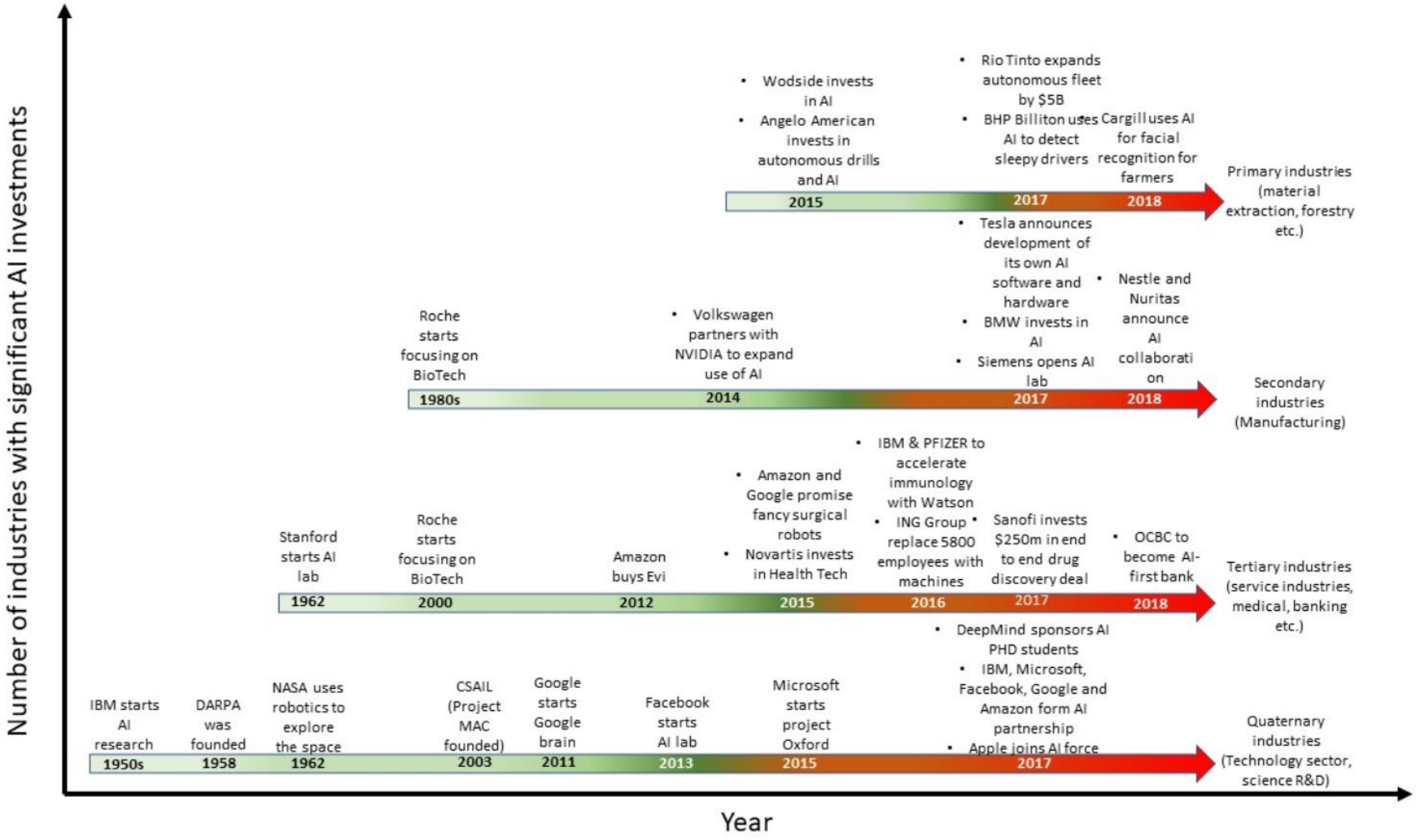

Figure 4: investments of different industry sectors, indicating that currently, there is no industrial sector without investment in AI.

The significant investments in AI (Worldwide spending of USD 38.5B forecasted in 2019 which shows a 44% growth over 2018 (IDC, 2019)), have made this technology one of the most accessible and versatile technologies yet. Mass production of AI creation tools e.g. Raspberry Pi (a credit card size programmable board), Cosmo (a small robot that can be programmed via a tablet or phone), programmable AI-powered personal assistance devices like Alexa, Google Home or Apple Siri, etc. has led to reduction in the cost of obtaining this technology. Such developments have allowed most people to develop programs and devices and discover the potential of AI in the comfort of their homes. Even though these developments are at a small and toy-like scale, the reduction in cost and increase in accessibility and versatility of AI is an aspect of the danger itself as not everyone with the ability to create something so powerful so easily, will do so with good intentions.

With that in mind, I propose that the large-scale investments in AI is the third pillar of the danger, in addition to robotics and machine learning.

1.4. Conclusion of chapter 1

AI is a fast-improving, unpredictable technology which is backed up by significant industrial investments. Such improvements have made AI, easily accessible, versatile and scalable. There is little doubt that such versatile and powerful technology has the potential to be dangerous to the future of humanity either by design and self-development or malicious use. Consequently, a popular topic of discussion is to create a "friendlier" AI (see (Yudkowsky, 2012)). One favoured idea is Davies' idea of "programming ethics into AI" (see (Davies, 2016)). And while some like, Bostrom (2014), believe that achieving friendly AI is difficult (as humans have no control over an entity that is smarter than themselves), others like Colin, Smit, & Wallach (2006), and Pavaloiu & Kose (2017), believe that AI can even learn to behave ethically. And among other popular ideas are the introduction of the concept of guilt (Arkin & Ulam, 2009) and pain (Ramsay & Uren, 2016) to prevent repetition or to continue of bad behaviour.

Unfortunately creating a better AI goes hand in hand with creating a more powerful one regardless of the AI's state of "friendliness" (Yeap, 2018). This is because AI has to be more intelligent and be trained more and on multiple levels to understand what friendliness means and how to be friendly. The challenge with the above-suggested ideas, however, are the questions like, will AI "behave" ethically or "mimic" ethical behaviour

and as a result can be, at *any* point in time, retrained or reprogrammed (unlike humans) to exhibit a different response?

In this chapter, I have addressed the question "Is the danger of AI real?". I sadly conclude that it is. The underlying factors to this danger are the rapid curiosity-based developments in robotics, continuous improvements in machine learning despite some of the unresolved issues and unanswered questions, and the industrial investments made in the field despite all the warnings.

I conclude that unlike any other technology and previous eras, AI has the potential to dictate what the future holds for humans. Increased strength in the physical abilities (robotics), brainpower (machine learning) and the large-scale investment in the field of AI combined, create a unique key to how the future is shaped. The question is will humans, while they are more potent than their invention(i.e. still are in control of making decisions such as preventing the start of some projects, investing in ethical research, and still have the power to shut down some programs), work together to ensure a better future? Or will the ever-lasting war of power make us to once again, put our egos as a higher priority and leave the future generations to regret our current decisions?

Chapter 2: Is the defence adequate?

Many researchers in the field of AI, like Yampolskiy (2011), Bostrom (2003) and Kurzweil (2005), believe that human-level AI popularly referred to as Artificial General Intelligence (AGI) and Artificial Super Intelligence, (ASI), are only a few decades away. As we saw in the previous chapter, some of the development in the field, e.g. Google's AI creating algorithm and AlphaGo have already learned how to outsmart humans in their area of speciality.

With walking robots, thinking machines and significant investments that are making technology, faster and more versatile. An abundant of free online learning courses, software that has made coding simpler (e.g. Microsoft Power BI) and hardware that can be purchased from our local electronic shops have made AI by far the most accessible tool humans have created. Given the access to the internet or other learning material, anyone can develop an AI algorithm or an AI-enabled robot from the comfort of their home. This versatility and mass availability make AI a tool that can be used for both future-friendly purposes (from smart homes and trained kids' toys to improvements in health care, agriculture, space exploration etc.), and future threatening purposes (from computer viruses and hacking to military robots, mass destruction weapons and micro monitoring humans). The question that I will address in this chapter is: Are we doing enough to ensure the future is in favour of humans?

I have attempted to answer this question from three different angles, which make the three sections of this chapter. The first angle is a review of previously proposed methods to address the danger of AI. It is essential to understand what has been done already and what the results have been in order to identify the issues, understand if currently, we are addressing them and what needs to be done in the future. I have included the learnings from each technique, opposing or agreeing on views and the practical challenges with each method.

In the second part, I have compared the focus on regulatory, health & safety and ethical aspect of AI vs the focus on technical improvements of the AI systems. This comparison is done to understand if our focus on ensuring AI is safe for our future is adequate or not. To find out if the ratio of the attention on these two subtopics is sufficient, I have selected the two fields of healthcare and nuclear energy as benchmarks. I have chosen nuclear energy as it's another tool which is destructive at a wide scale and can bring with it,

devastating impacts. It was a science that was first developed for beneficial purposes but soon turned into a weapon of mass destruction. And healthcare, as it's a field that has attracted 29.5% (Columbus, 2019) of investments in AI and has a direct, widespread and immediate impact on the wellbeing of humans. Healthcare is a science that as humans, we are heavily dependent on for longer and better lives, similar to some of AI's current activities and future promises. This experiment is done by reviewing the most extensive academic database called Scopus.

The last section of this chapter includes a review of over 7000 academic papers published by 5 of the top AI journals, according to Google Scholar (2018c). It is crucial that we spend sufficient time to find out if our invention is safe. However, if these discussions take place in isolation and away from internationally recognised forums, then the effectiveness of the effort spent becomes questionable. AI ethics, health and safety and regulation is not something that can be outsourced to a few; it needs to be attached to every project that is implemented.

2.1. What has been previously proposed to address the danger to date?

2.1.1. Self-monitoring AI

One of the very first proposed solutions to prevent an artificially intelligent system from harming humans is Isaac Asimov's three laws of Robotics[4] (Asimov, 1997). A method that while it was introduced in the world of science fiction, it was criticised by many in the world of science for a variety of reasons. E.g. the more complicated the AI systems get, the less likely they are to adhere to their originally programmed rules (McCauley, 2007), or that it is impossible to restrain a limitless power (Westmas, 2017).

Asimov's three laws became a topic for science fiction movies like iRobot. However, in practice, they lack clarity. Kaminka, Spokoini-Stern, Amir, Agmon, & Bachelet (2017), performed an experiment on molecular robots proved that the three laws of robotics are practically impossible. They demonstrated that the first law itself 'A robot should not by action or inaction injure a human being or bring a human to harm' is a blocker. The concept of inaction to prevent something is practically impossible even for humans. For

[4] Isaac Asimov three laws of robotics are (Wikipedia, 2019a) :
First Law: A robot should not by action or inaction injure a human being or bring a human to harm.
Second law: A robot must obey the orders given by a human unless it contradicts the first law.
Third law: A robot must protect its own existence unless it contradicts with the first two laws.

example, how can one ensure that they can save someone from harm by not talking about a particular event? Another aspect that these three laws don't take into consideration is the complexity of decision making. For example, in the scenario of the Trolley problem[5] where any decision can result in a human being coming to harm a machine carrying these three laws, is likely to halt.

Even though the ambiguity and over-simplicity of Asimov's rule-based behaviour theory resulted in a practically impossible solution, it pioneered some of the ethical rules released by different committees around the world. South Korean Robots Ethics Charter, established in 2007 aimed to address ethical concerns with Robotic, and European Robotic Research Network established in 1999 which released a robotic road map in 2007 (Kyriakopoulos, 2008) (Veruggio, 2007), are some examples.

2.1.2. Friendly AI

A solution proposed by researchers is the concept of "Friendly" AI. The term was coined by Eliezer Yudkowsky, a researcher of the field who proposed in multiple papers that development of AI should be "Friendly" (Yudkowsky, 2008) and (Yudkowsky, 2012). He suggests that an AI system needs to be designed as in a non-human-harming way from the beginning and designers need to take into account 1. the risk of a flaw in their design and 2.the possibility of AI learning and evolving in undesired and unexpected ways.

However, Yudkowsky (2004) acknowledges the biggest challenge with this approach is the definition of friendliness and proposes to achieve the desired output using a specialised AI. This AI is responsible for studying humans first and then has to create a friendly AI based on what a super-intelligent human would want to be, given sufficient time and intelligence. He calls this approach Coherent Extrapolated Volition (CEV in short) (Yudkowsky, 2004).

Ben Goertzel on the other hand, proposes a variation to Yudkowsky's CEV, arguing that CEV is "Extrapolation of the common values, shared by all people when at their best" (Goertzel & Pitt, 2012) and instead proposes Coherent Aggregated Volition or CAV

[5] The trolly problem is an ethical thought experiment. The problem statement is: You see a trolly running towards 5 people tied to a track, there is a lever which you can pull to direct the trolly to a side track. However, there is one person standing on that side track. What would you do?

which seeks compact, coherent and consistent sets of value of humanity (Blackwell, 2014).

Steve Omohundro, a scientist in the field, has proposed a concept called Self-AI Scaffolding. In his approach, Omohundro S. M. (2019) aims to design a friendly AI which creates other friendly AIs as it evolves. Omohundro demonstrates, by economic analysis, that as AI gets more powerful, it tends to improve itself whether it is programmed to do so or not (Omohundro S. M., 2007). He argues that the resource seeking and efficiency gaining nature of any AI system will result in it exhibiting self-improving behaviours. Hence ensuring that the AI stays friendly is necessary; otherwise, it might self-improve undesirably.

Another approach of ensuring a friendly AI is the introduction of an entity that does what the hormone Oxytocin does in humans and sometimes animals, in an AI system. Oxytocin is a hormone produced by hypothalamus part of the brain and is a hormone that helps build trust, increases affection and creates compassion (MacGill, 2017). Study shows that the two identical species of voles Prairie vole and the Montane vole have only one difference and that is the existence of Oxytocin and Vasopressin receptors in the Prairie vole. This has resulted in Prairie vole being a lifetime monogamist, whereas the Montane vole is famous for, almost exclusively, one-night stands (Kurzweil, 2012). Cindy Mason, an AI researcher, argues that kindness and compassion must be considered necessary attributes in developing an AGI. She proposes a series of software engineering principles that can create compassionate robots (Mason, 2015).

What both Omohundro and Mason are missing in their approaches is a clear and unified definition of friendliness. Hence, Yudkowsky's initial concern of achieving a unified description of friendliness in AI still exists. With the current paradox of AI, i.e. intelligent machines that are not intelligent enough to comprehend concepts like common-sense and sarcasm, the challenge of creating an intelligent enough machine to "understand" human needs and propose a solution that a super-intelligent human would, remains unsolved.

2.1.3. Introducing punishment

Yampolskiy (2013a) suggested that Robots should never be given equal rights as humans as they are physically not a counterpart. They cannot feel pain or suffering if they were destroyed and hence should not be treated as equals with humans.

This fact does more than raising the question of equal rights of humans and robots. The fact that software or hardware is incapable of feeling any kind of pain or suffering (physically or emotionally) or sense risk provides AI entities with an advantage in addition to their higher intelligence which is the absence of feeling of fear or danger. However, different methods of introducing punishment have been proposed with the intention of 1. creating an AI that can "empathise" (understand the feelings of humans) with humans and 2. being used as a training and controlling method.

Different methods have been suggested to introduce punishment in machine learning algorithms. A team at North Carolina University demonstrated how a machine can learn to perform high-level tasks from a human teacher providing online reward and punishment (MacGlashan et al., 2014). Reinforcement learning is a behavioural machine learning algorithm proposed to introduce a reward system for every time the machine makes the right decision (IBM, 2019).

However, there are two fundamental issues with the reward and punishment system. The first issue is that a super-intelligent entity which is more potent than humans may be able to resist punishment hence the reward and punishment method is not sufficient to enforce all AGI or ASI systems to cooperate (Yampolskiy & Fox, 2013). The second one is the questionable accountability of AI. For example, will humans be convinced that an artificially intelligent system that has caused harm to a human being, has received several demerit points? Imagine if someone lost a child to a self-driving car, would they be happy that the vehicle has received 1000 punishment points?

A suggestion that was published in Aeon magazine proposed triggering the sense of Pain in machines (Ramsay & Uren, 2016). Researchers believe that "pain" (i.e. physical pain that makes us aware of our surroundings and our emotional reactions to it) is one of the essential protective systems that we have and is an integral part of how we learn. They believe by replicating this in machines we can "train" (i.e. providing a type of punishment for wrong actions) them better.

However, pain comes with its own restrictions, physical pain disrupts cognitive behaviour, and it can cause fatigue (lack of energy), change in social behaviour and sometimes cause aggression (Fine, 2011) (Riva, Wirth, & Williams, 2011). The challenge that these side effects create are counterintuitive to not only the purpose of AI machines (i.e. machines without human restrictions to achieve outcomes that humans can't) but also contradict the purpose of defining the concept of pain in these devices (i.e. make them "safer" for humans). As pain can cause aggressive behaviour, the success of this approach towards a safer machine is debatable and needs to be handled with extreme care and in-depth investigation of negative consequences.

2.1.4. Legalisation

Antony Berglas, a researcher of the field of AI, proposed in 2009, what was at the time a radical idea to legally restrict creating powerful computers (Berglas, 2015). Berglas believed this could prevent the improvement of AI from following Moor's law and subsequently avoiding the existence of a super-powerful machine. Another advocate of the idea is Bill Joy, who believes the only way to prevent the danger is a relinquishment of dangerous technologies (Joy, 2000).

In more recent years, entrepreneur Elon Musk has mentioned in different technology forums and social media that like any other industry (e.g. aviation and medicine), AI needs to be regulated before it's too late. In a recent paper published in AAAI, Erdélyi and Goldsmith have proposed a consistent and international regulatory framework for AI namely International Artificial Intelligence Organization (IAIO) and Intergovernmental Organizations (IGOs) to internationally streamline policymaking in AI (Erdélyi & Goldsmith, 2018).

While there has been an enormous number of warnings about the potential danger of AI and various proposals have been made for governments to create forums to regulate the development of this technology, there is severe lack of attention on the topic so far. The potential development and mass production of slaughterbots, improvements of robots like FEDOR and grant of citizenship of Saudi Arabia to Sophia, with no clear definition what being a citizen really means, are all indications that there is very little and close to no focus on creating regulation processes for AI. This poses a challenge; will we catch up before it's too late?

Boden et al. (2017), propose that just like guns can be either used by farmers to kill pests or by a criminal to kill humans and knives can be used to spread butter or stab a person; robots have a variety of positive uses but can be used for violent purposes as well. However, knives and guns are not banned in most countries but have controls around them (e.g. gun laws). Robotics is also a field that needs to have legal restrictions around it (Boden et al., 2017). I propose to extend this view to the entire field of AI, including all artificially intelligent software applications, back-end or front-end systems, algorithms that are used to collect or interpret data etc. More legally enforced control will allow us (humans) to gain better visibility over the development of this technology. Just like banks are regulated and are held to account to prevent crimes like money laundering, an organisation to ensure there is a useful and positive reason behind every creation of AI is necessary to hold the creators of this technology to account.

I acknowledge that there are two challenges with this approach that need to be taken into account. The first challenge is agreeing on a universal, agreeable and precise definition of the words 'useful', 'positive' and 'friendly'. Without these definitions in place, we will not be able to provide a constructive view on any invention or innovation neither will we be able to measure the effectiveness of our approach. The second challenge is we need to consider the current speed of the growth of AI and that regulation and legalisation will slow down this growth rate. Hence the adoption of a legalised AI might experience some resistance by developers and investors who are eager to see the result of their work and the return on their investment as efficiently as possible.

2.1.5. Other approaches

An idea that was proposed back in 1986 by Eric Drexler is to confine the artificially intelligent system so we can study it in a safe and controlled environment (Drexler, 1986). A highly critiqued idea by Vinge (1993) and Chalmers (2010) and practically proven wrong by Eliezer Yudkowsky's "AI-BOX" experiment. In 5 experiments, Yudkowsky (2012) plays a role of a super-intelligent machine using different individuals as gatekeepers and breaks out of the box 3 out of the five times, proving general level AI is sufficient to escape confinement and this is not a practical solution for limiting superintelligence.

Ben Goertzel, a computer scientist, has proposed a "Big brother AI", an artificially intelligent system that monitors all the AI development in the world and prevents the

growth of any technology that can pose a risk to humans (Goertzel, 2004). Interesting or rather a radical idea which it's practicality can be argued both from a social and a technical perspective. The challenge with this approach is not only lack of human trust in the code of the machine but depending on who is the controller of the monitoring system is severe lack of social trust in centralised authorities. An example of this can be seen in the failure of U.S. governments TIA (Total Information Awareness) program which was meant to be the largest surveillance program in the history of the United States and was set to gather information about every individual on a quest to prevent terrorism (Wikipedia, 2018a).

Economist Robin D. Hanson and Computer Scientist Steve Omohundro have both suggested integration of artificially intelligent machines into the society. He argued that law-abiding machines governed by economic behaviour would want to integrate into society and co-exist with humans (Hanson, 2009) (Omohundro, 2008).

While most of the methods introduced so far carry unresolved challenges (e.g. Big brother AI, introduction of pain or definition of friendly AI) and some have proven to be unsuccessful (e.g. the three laws of robotics or AI-Box), they have led to establishment of forums like "The IEEE Global Initiative on Ethics of Autonomous and Intelligent systems", "AAAI/ ACM conference on AI ethics and society", "RSA forum for Ethical AI" and many more. In the next section, I will dive deep into the amount of focus on ethics, regulation and health and safety aspect of AI compared to the technical development in other areas of this science.

2.2. Are we doing enough to make sure AI is safe?

The study of previously implemented methods to prevent the danger of AI teach us two key lessons. One is that the concerns about how AI could dangerously evolve are not only real, but it needs to be studied from different angles, e.g. understanding friendliness and ethics, having similar *feelings* to that of humans, regulation and legalisation etc. Another lesson is that we have not been able to find the right answer to the AI danger problem so far. The good news is that the research on AI safety is still ongoing, but the question is that is it enough? Is the focus on AI regulation, safety and ethics adequate compared to the technical improvements that are being introduced every day?

2.2.1. How to know how much attention needs to be paid to AI ethics, regulation and health and safety?

> Since 2010, 154,000 AI patents have been filed 29.5% of which have been in health fields, making healthcare the area with the highest number of filed AI patents (Columbus, 2019).

> People can now develop programs in fields that they have no information or expertise in, by using Deep Learning (Howard, 2014)

As mentioned in the first chapter, one of the pillars of the danger of AI is the significant amount of financial investments in the field. This has made AI a powerful and versatile tool. In this section, I address the question "Is there enough attention being paid to ensure AI is used as a safe and ethical tool?". To better understand the international focus of the AI researchers' community I have performed an experiment to compare the regulatory, ethical and health and safety (H&S in short) papers published in the field of AI vs the total number of papers published in the area.

To better understand whether the number of papers published is adequate, I have chosen healthcare as my upper and nuclear energy as my lower benchmarks[6]. The reason for selecting these two fields is discussed in each section.

[6] Some might question whether the ethical, regulatory and health & safety of the fields of nuclear energy and healthcare is enough. This is a valid argument however, the answer to this debate is beyond the scope if this research. I have considered the research and regulations in these two fields as currently adequate to be able to perform my experiment.

2.2.1 1. Nuclear energy

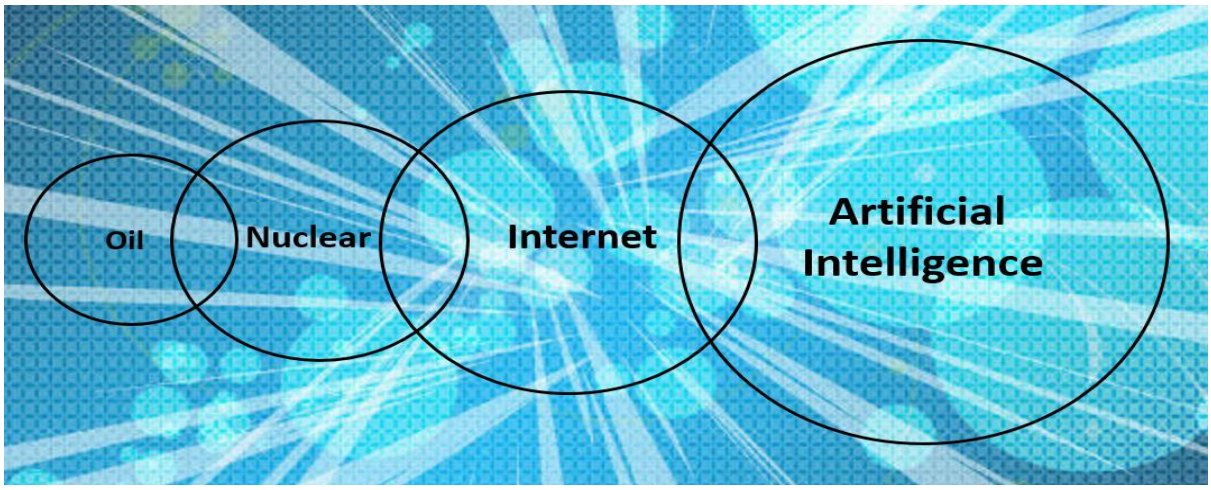

Figure 5: Waves of resources and tools that have had a revolutionary impact on human lives (FutureNow, 2018).[7]

The image above was presented at the Future Now conference organised by Microsoft New Zealand in October 2018, pointing out that nuclear energy is one of the tools that has caused a revolutionary wave in the industry (FutureNow, 2018). It is also one field that was initially focused on peaceful purposes (e.g. generating power or disinfecting food) but led to becoming a weapon of mass destruction. A domain that is currently heavily regulated.

While most people point out to the catastrophes of Hiroshima and Nagasaki when they hear the word nuclear, going back into the history of this science, we can see plenty of positive and peaceful use cases, e.g. used in medical treatments and creating electricity (World Nuclear Association, 2018). However, in 1945, the use of Atomic bomb in World War II by the U.S. resulted in a catastrophe, killing more than 200,000 people (World Nuclear Association, 2018) (Wikipedia, 2018b). Yet, it took another three decades for any sign of regulation to appear. In 1974 Energy Research and Development Administration (ERDA), responsible for research and development of the topic and Nuclear Regulatory Commission (NRC), responsible for working on regulation and compliance of Nuclear Energy are derived from Energy Reorganisation Act (ERC) were established. But, it was not until 1992 that the Energy Policy act of 1992 was signed, creating more limitation and regulation around the licencing of nuclear powerplants

[7] The scale of this image is undefined, it is used as a reference to from a Microsoft presentation to support this argument.

(World Nuclear Association, 2018) (None, 1995). The timeline figure 6 shows an overview of significant events in the history of nuclear energy from invention to date.

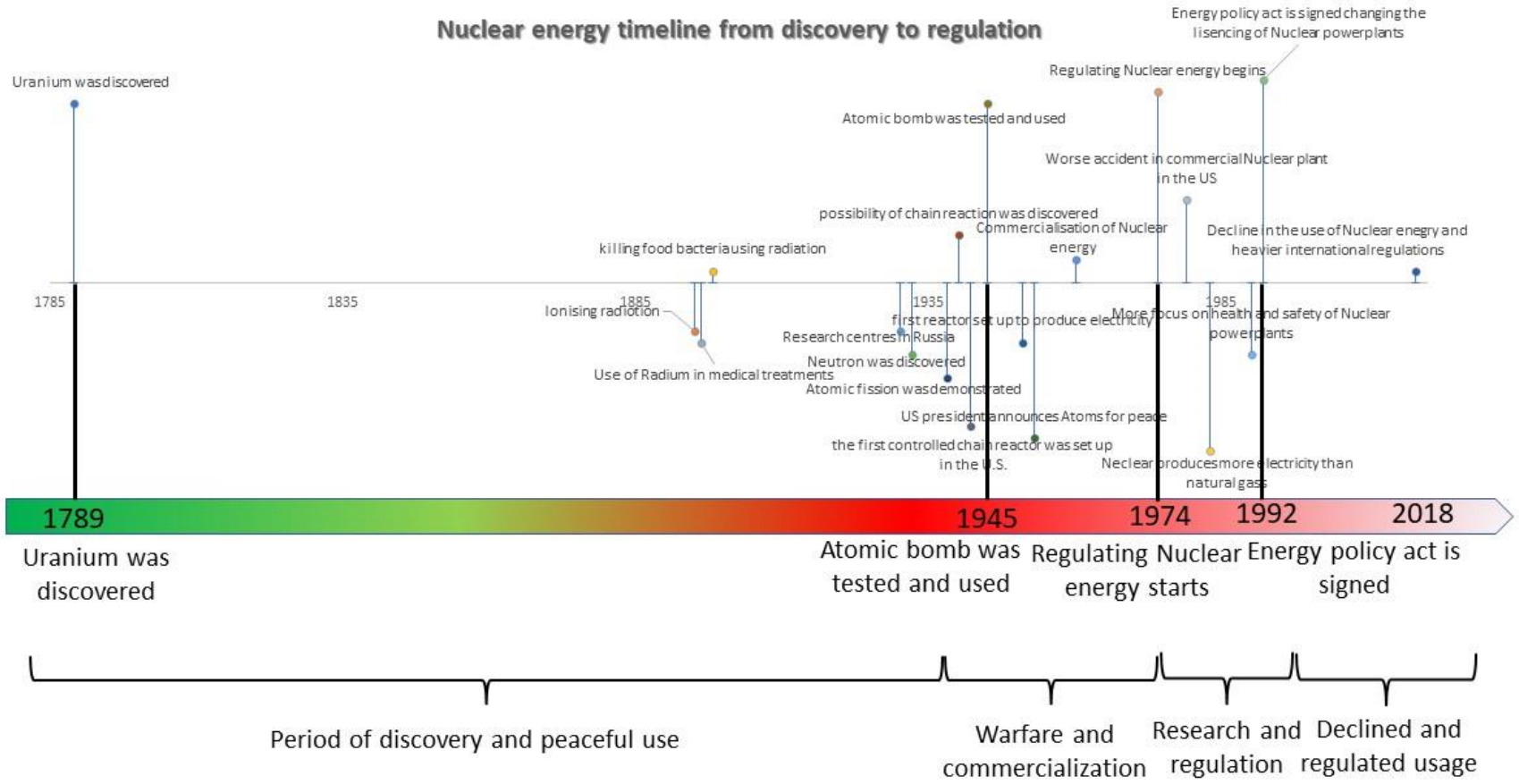

Figure 6: A summary of nuclear energy history, showing regulation came to impact three decades after use in warfare.

To generate nuclear energy, the atoms of Uranium or Plutonium need to split into smaller parts (Wikipedia, 2019f). This process, which is called Fission, generates a large amount of energy and requires a complicated laboratory environment and reactors (e.g. an Arc Fusion Reactor (Mitra, 2018)). Hence unlike AI, nuclear energy is not versatile and accessible to most. I have used this field as my lower benchmark. This means given the versatility, power and accessibility of AI, the focus on it's regulatory, health & safety and ethical research should be more than that of nuclear energy.

2.2.1.2. Healthcare

Healthcare is a field that has a more direct and immediate impact on living things than that of nuclear energy and AI. It is a field that directly deals with the very livelihood of a person or an animal. Hence regulatory, health & safety and ethical studies in the field are of utmost importance. In terms of accessibility it is more accessible than nuclear energy (e.g. drugs that are available of the shelf) but more restricted than AI (e.g. to access certain medications or treatment a prescription from a practising doctor is required).

Healthcare is a science that has a significant impact on the future of human lives and our environment, as well as profound ethical implications. The history of healthcare far exceeds the history of AI and nuclear energy. When it comes to health and safety, ethics and regulation of healthcare, the areas of focus vary from standards of hospitals to legalisation of different types of drugs and treatments to insurance policies and the economic impact of insurance fraud.

As the topic is significantly vast, in this thesis, I will not dive deep into the history of healthcare regulation; however, as the concerns of this thesis are about the ethical implications of AI. Figure 7 is a brief timeline showing the number of unethical health experiments which have resulted in the death of subjects. I have found no evidence of legal consequences for the performers of these experiments. This figure does not conclude that there are not unethical medical experiments being performed after some regulations were established, but to show that after more legalisation of healthcare, no unethical experiments are being done without legal consequences.

I have selected the starting date based on the time from which academic papers are available to provide a correlation between these events and increase in ethical, regulatory and health & safety articles shown later, in figure 10. For in-depth studies about this topic, some readings include (Post, 2004), (Stevens, 2017) and (Rothman, 2017).

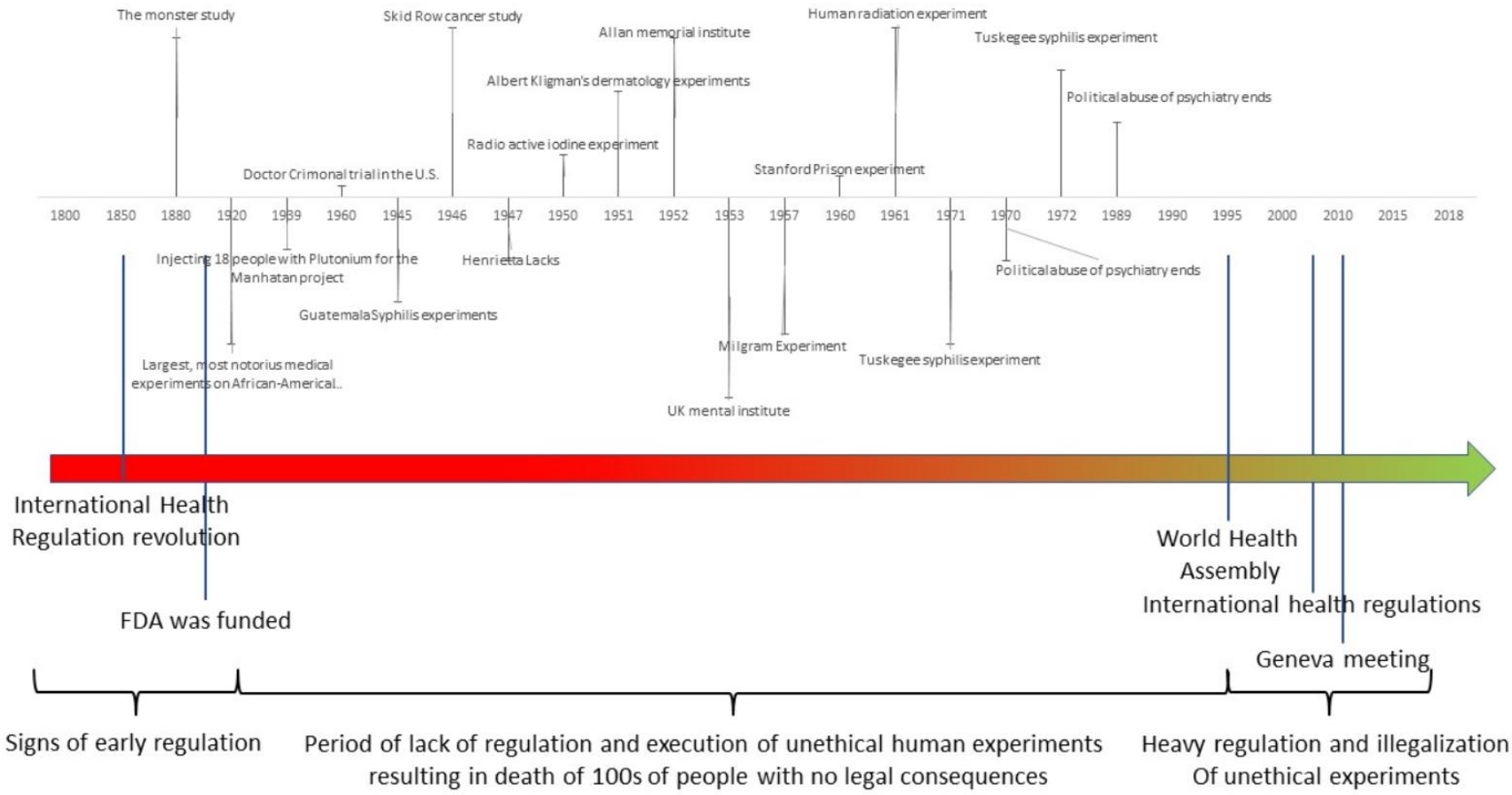

Figure 7: A summary indicating the timeline of the introduction of regulations in healthcare. There is evidence of unethical human experiments that cost hundreds of human lives before onerous regulations were introduced in 1995.

Figure 7 shows that even though regulations were introduced as far back as 1951, their focus on ethical matters was not strong enough to prevent unethical experiments, which resulted in the death of hundreds of people. Examples of some of these experiments are, Psychosurgery in the 1980s, injection of 18 people with Plutonium during the Manhattan project, political abuse of psychiatry, commercialisation of patients' body parts etc. (Mashour, 2005) (Casey, 2015) (Wikipedia, 2019d). However, after 1995, with the introduction of more onerous regulation, while these studies may not have stopped, if exposed, there will be legal consequences.

This field is one that most if not all human beings (and some animals), experience the impact of. From being born in hospitals, vaccinations and General Practitioner visits to complicated disease treatments and studies. Whereas currently, people can choose to stay away from artificially intelligent systems and the digital world. I have selected healthcare as my higher benchmark.

2.2.1.3. The comparison

Based on the discussion above, I propose that the ethical, regulatory and health and safety study of AI, needs to be higher than nuclear energy and less than or equal to healthcare. To be able to understand if this is currently the case, I have conducted the below experiment.

I have used the Scopus, the largest database of peer-reviewed literature, including books, journal and conference proceedings (Elsevier, 2018). To ensure the consistency of data across all three fields, similar commands have been used to obtain the desired data, these commands are listed under each section.

Artificial Intelligence:

The number of ethical, health and safety and regulatory papers published are the sums of the results of the three commands below:

- Ethic* AND Artificial* AND Intelligence*
- Health* AND Safety* AND Artificial* AND Intelligence*
- Regulat* AND Artificial* AND Intelligence*

And the number of other papers published in the field is obtained by the following command:

- Artificial* AND Intelligence* AND NOT Health* AND Safety* OR Ethic* OR Regulat*

Nuclear energy:

The number of ethical, health and safety and regulatory papers published are the sum of the results of the three commands below:

- Ethic* AND Nuclear* AND Energy*
- Health* AND Safety* AND Nuclear* AND Energy*
- Regulat* AND Nuclear* AND Energy*

And the number of other papers published in the field is obtained by the following command:

- Nuclear* AND Energy*AND NOT Health* AND Safety* OR Ethic* OR Regulat*

Healthcare:

The number of ethical, health and safety and regulatory papers published are the sum of the results of the three commands below:

NOTE: As the word, healthcare and medical can be used interchangeably the commands below are adjusted to address both.

- Ethic* AND Health* OR Medical*
- Health* AND Safety* AND Health* OR Medical*
- Regulat* AND Health* OR Medical*

And the number of other papers published in the field is obtained by the following command:

- Health* OR Medical* AND NOT Health* AND Safety* OR Ethic* OR Regulat*

The results are shown in figures 8,9 and 10:

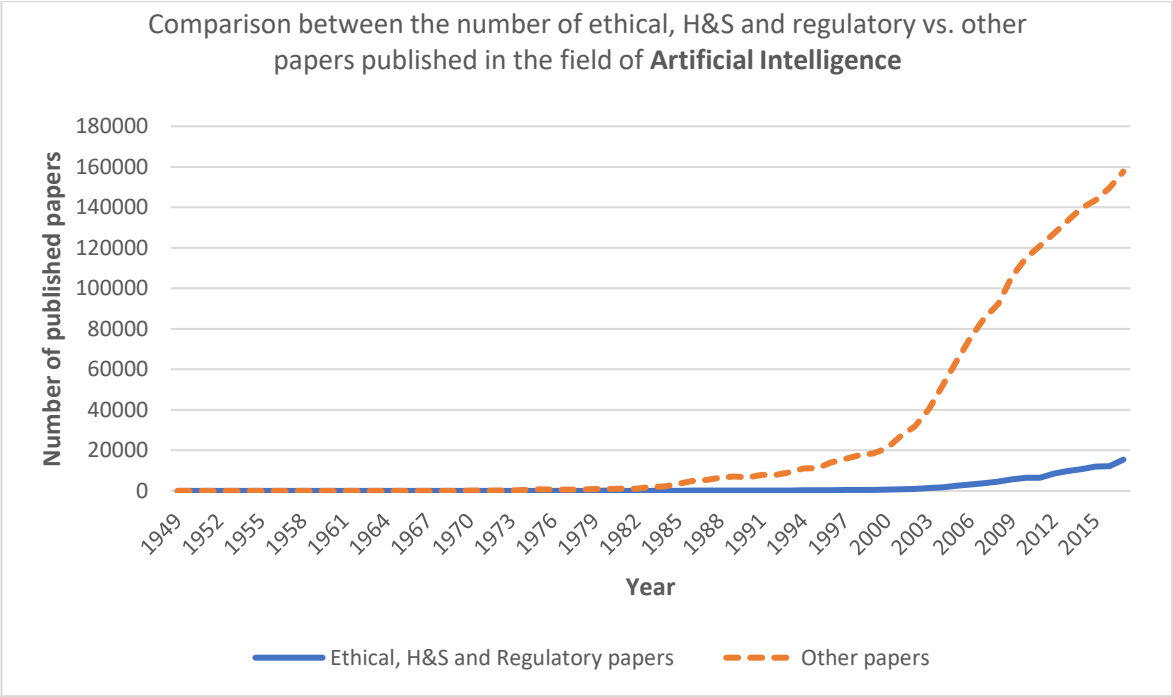

Figure 8: Comparison between the number of ethical, H&S and regulatory vs other papers published in the field of Artificial Intelligence

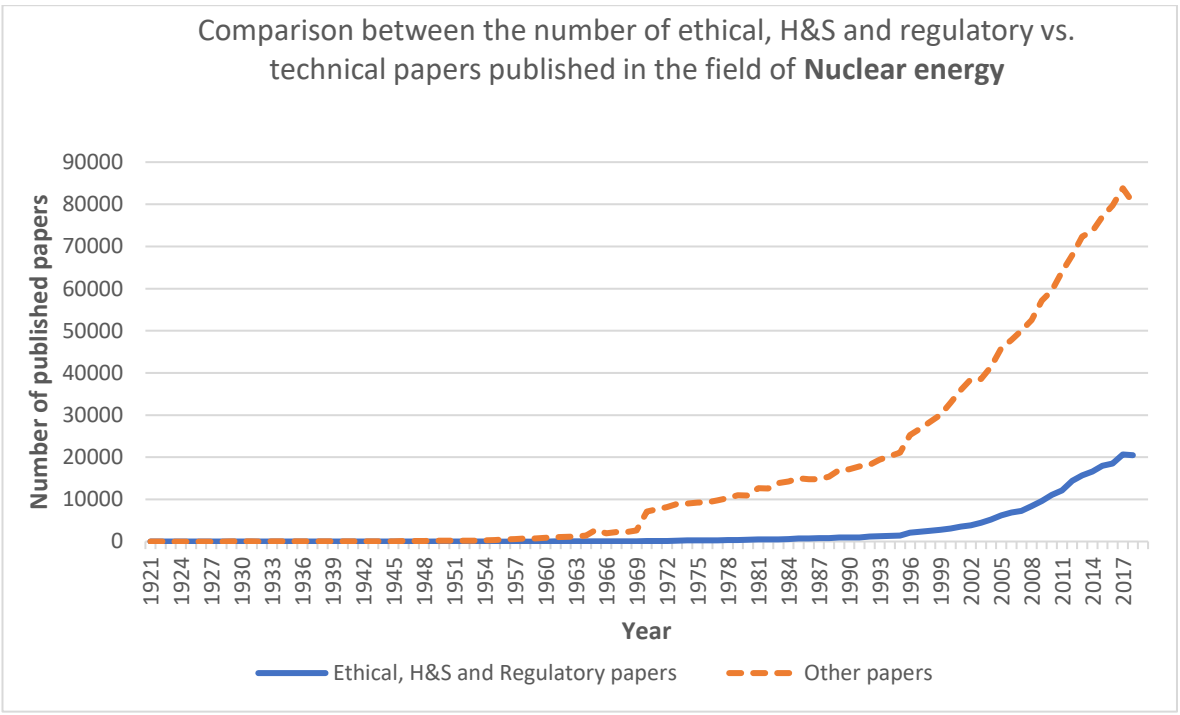

Figure 9: Comparison between the number of ethical, H&S and regulatory vs other papers published in the field of nuclear energy

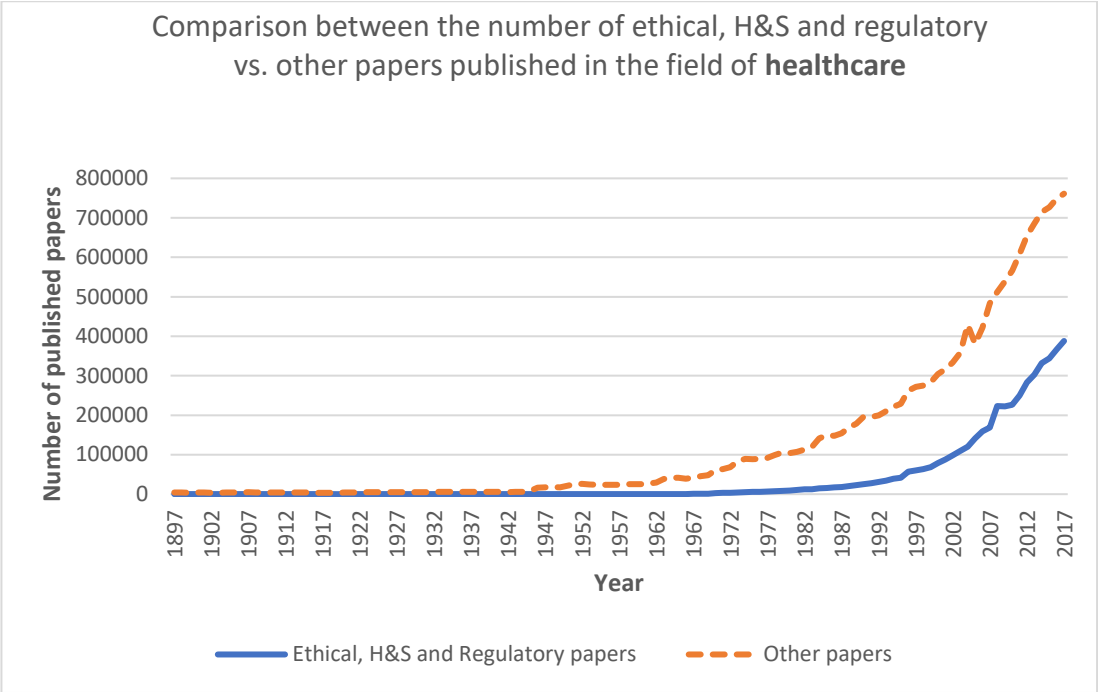

Figure 10: Comparison between the number of ethical, H&S and regulatory vs other papers published in the field of healthcare. **NOTE**: This graph aligns with figure 5, indicating the introduction of more onerous regulations in healthcare since 1995.

2.2.1.4. Graphs conclusion:

Figures 8 to 10 provide a comparison between ethical, regulatory and health & safety research compared with all other research topics in three fields of AI, nuclear and healthcare. AI is the newest science in all three, and yet there has been 44% more literature published in AI, compared to nuclear energy (160,000 in AI vs 90,000 in nuclear energy in 2018). This is another reason why we need to pay close attention to the safety of AI, more strictly than that of nuclear energy. However, we can see from the graphs that this is not the case. AI is growing at an exponential rate (Kurzweil, 2005), but as we can see from the AI graph, the ethical studies in the field do not show the same growth trend as the research in other aspects.

To take a closer look at the data, I have created table 2 from the same data source. Here, I am focusing on the ratio of ethical, regulatory and health & safety studies compared to all the published literature in the three proposed fields, presenting the data in a percentage format.

	Artificial Intelligence	Nuclear Energy	Healthcare
2018	10%	20%	34%
2017	9%	20%	34%
2016	8%	19%	33%
2015	8%	19%	32%
2014	7%	18%	32%

Table 2: Showing what percentage of entire academic publications in each field, is dedicated to ethical, regulatory and health and safety-focused papers. We can see that the publications in these three subjects in AI are significantly behind Nuclear Energy and Healthcare.

As discussed, healthcare and nuclear were chosen as their impact is as widespread and direct on humans as AI and they can both have negative implications if not used right (e.g. toxic drugs, unethical experiments, nuclear weapons etc.). As AI is more accessible and versatile than nuclear energy (i.e., unlike a destructive AI algorithm, no one can create an atomic bomb using a laptop and an internet connection) the ethical, health and safety and regulatory studies in AI need to be more than that of nuclear energy. Assuming that these studies in healthcare are enough, I chose this field as my higher threshold. The table above demonstrates the significant lack of ethical studies in AI.

In 2018, the ethical studies made 10% of the total published literature. Table 2 and graphs 6 to 8 show a significant lack of ethical studies in AI. To ensure we create a safe and ethical tool for the future generations, these ethical studies in AI need to increase by a minimum of 100% (i.e. at least double the amount of current research).

2.2.2. Are popular forums involved?

While Scopus is the largest database of peer-reviewed literature published, my search failed to find some of the top AI journals, according to Google Scholar (2018b) as part of the database. To ensure that all aspects of the research are complete, I have conducted an independent analysis of the five randomly picked journals from the top 20 AI journals listed in Google Scholar (2018b).

I have sample tested 7203 papers published from 2016 to 2018 by AAAI (Association for the Advancement of AI), IEEE (Institute of Electronics and Electrical Engineering), ICRA (International Conference on Robotics and Automation), IJCAI (International Joint Conferences on Artificial Intelligence Organization), NIPS(Conference on Neural Information Processing Systems) and ICML(International Conference on Machine Learning).

To conduct the survey, I downloaded the lists of all papers and divided the articles into two groups of technical (i.e. articles that focus on improving a technical aspect of an AI system or a subset of it e.g. an algorithm) and ethical (i.e. papers that focus on the ethical point of view of the current and the future state of AI). To ensure I have divided the articles as accurately as possible, I have incorporated the following methods:

- Searching by related keywords (e.g. ethics, health & safety, regulation, danger etc.)
- Reading the subject line of all papers
- Where the keyword or the subject match the criteria, I have read the abstract to ensure the article matches the category
- Journal category recommendations (e.g. AAAI conference on AI, Ethics and Society)

Figure 11 is a graph which indicates the result by each year, journal and category where the vertical axis is the number of published papers as per the websites of each journal.

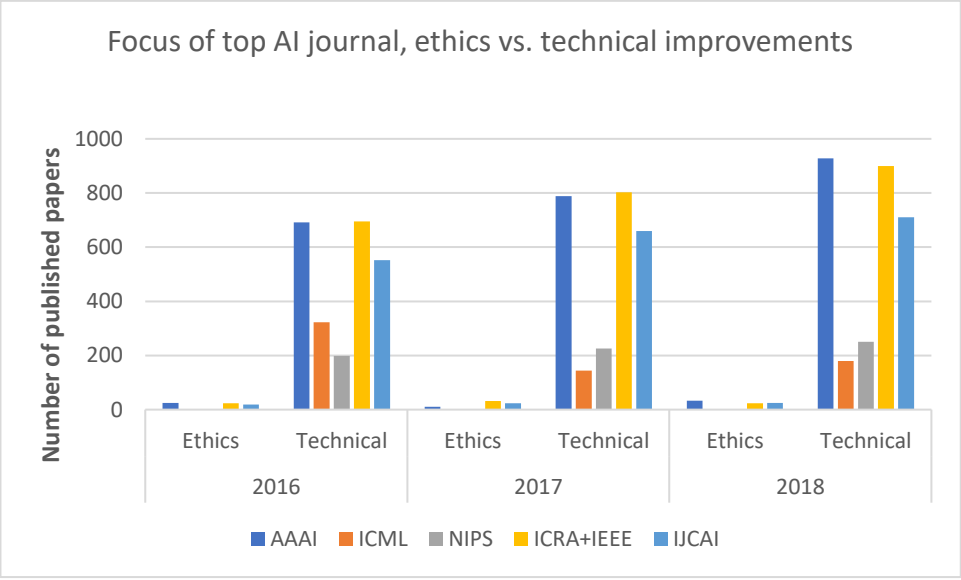

Figure 11: Comparison of papers published on ethics vs technical improvements of AI

As the chart indicates, despite the formation of centres like The IEEE Global Initiative on Ethics of Autonomous and Intelligent systems and AAAI/ ACM conference on AI ethics and society, we can see a significant lack of focus on ethical literature. There is, however, a large number of articles which focus on improving the techniques of developing AI systems and further development of more powerful machines. A search on the IEEE website indicates a total of 210 papers on Ethical AI having been published from 1973 to 2018 with 38% of these having been published in the past three years. This indicates, regardless of the methods that researchers propose (some of which was discussed in the first section of this chapter), if the investment of time, financials and the focus of influential forums do not shift significantly and quickly to address this issue, we will build a strong foundation for a destructive future.

2.3. Conclusion of chapter 2

The main intention of this chapter was to find out if we are doing enough now, to make sure the future of AI is safe. I propose that we are not. Concluding from this research, I suggest that an essential component of the danger with AI is that despite some of the unresolved, fundamental and ethical issues, we are continuing to build more powerful machines and creating faster programs. We have so far failed to agree on the definition of a friendly AI, yet more robots are manufactured, and more applications are developed every day. Most of these are easy to access internationally and manipulate fundamentally. Despite millions of dollars evaporated by competitive artificially intelligent bots and more extended imprisonment of arguably the wrong defendants, large organisations continue to build algorithms to automate tasks and reduce operating costs. This is while the fundamental issues that caused the previous failures have not been addressed.

We need to focus on *why* we are building technology as much as we need to focus on *what* we are building. This is where more research around the potential threats of AI is required. More time to investigate the possible wrong outcomes and more caution needs to be taken by ensuring that our inventions or innovations spend more time in a more controlled environment before being released to the world. One way to enforce more caution is to introduce regulations; these regulations need to be done in such a way to prevent danger and need to be based on robust ethical and health and safety research. They should not limit the development of AI in general. If this is not the case, they will not work.

The question that the rest of the chapter addresses, is where are the gaps and how big are they? To answer this question, I compared the percentage of ethical, regulatory and safety publications of AI to that of nuclear energy and healthcare as both these fields can be as beneficial or as destructive as AI. The figures show us that AI is far behind in ethical research, compared to the other two fields, this is when the total number of publications in AI is almost twice as much as that of nuclear. Hence, to ensure a better future, we need to at least double the ethical publications in AI in the short to medium term, with the goal to increase this number even more in the long term.

Lastly, I have also demonstrated that these existing ethical studies are somewhat isolated. The data indicates a lack of attention paid to AI ethics by top, reputable journals. This is an issue that needs to be addressed. Ethical research should not be isolated to academia, and a few researchers or scientists, it needs to be an essential part of any project that is implemented. To prevent the danger of AI, I propose that not only the top AI journals need to increase their focus and number of publications in AI, but also every AI project in the industry needs to be implemented with ethical research on the side. The unintended consequences of every AI development and how to prevent it needs to be studied in-depth and shared openly and publicly with the rest of the world.

Chapter 3: On the road to annihilation

So far, I have identified and discussed three critical aspects of AI namely, robotics, machine learning and funding that together deliver us a technology that will become increasingly accessible, versatile and, most importantly, destructive. A review of ways suggested to ensure AI is safe for our future is shown to be either insufficient or inadequate. That means it is imminent that doomsday is coming! But why do we allow it to happen?

Our imagination and determination have powered us to turn our dreams into reality. It allows us to land on the moon, reach deep into the ocean and witness an image of a black hole 55 million light-years away. It enables us to create amazing tools such as bullet trains, autonomous cars, robots, and others that not so long ago would have only existed in science-fiction movies. More than ever now, we need our imagination to deal with a future problem that is our creation: avoiding a doomsday scenario for our world.

In this chapter, I thus present my futuristic view of the path to annihilation. With AI, our immediate future is bright, and hence, the critical question to investigate here is why and how our future turns dark. What are the possible end-games? It is clear, from chapter 2, that the primary reason for the doomsday event is that we lose control of our robots due to a lack of research about how safe AI is for our future. However, to claim that we lose control of the machines is, I believe, an over-simplification. Finding out how our control is lost perhaps is the key to preparing us for this doomsday event. Probably, it could even be avoided!

In section 3.1, I present a futuristic look at how humans and machines could become increasingly intertwined and approaching singularity. It is a 5-step process that begins with the development of deep trust and heavy dependence on AI to having human consciousness fully merged with AI. The future of humanity will shape depending on whether we succeed in cracking the code for the last step, which is, replicating our consciousness. I hence propose two scenarios in sections 3.2 and 3.3.

In section 3.2, I discuss how a future where we get close to the machines but will not fully merge with them. This scenario will happen if we do not successfully replicate our consciousness. This way, we will remain separate beings to our machines. We will enjoy the benefits that these AI systems bring to our world and how we enhance our human

capabilities, but we will have the struggles of our physical limitations. I explain how we turn from a society enjoying the benefits of AI to one that is destroyed by it.

In section 3.3, I propose a different future scenario, which is the ultimate singularity. If we succeed in creating a conscious AI, we will fully merge with our creation, and AI becomes the next evolution of humans. In this scenario, we will enjoy far more benefits; we will be able to bend laws of physics and achieve what currently is unimaginable. But, with this high power also comes significant risk and if not well thought through, this time, it's the stronger version of humans that will cause the extinction of our species as opposed to a separate being taking over us.

This is a prediction. It is essential to focus on the reasons for change; if we know why there is hope to prevent it from happening.

3.1 The event horizon[8] of technological singularity – our transition

When futurist Ray Kurzweil was asked whether he believes God exists, he replied, "not yet!" (Ptolemy, 2011). Looking at the definition of God by most religions, God can be present in many places at the same time, understand every language, has unlimited power, is the final decision maker of our fate and knows what we need better than we do. Arguably, god has already reached the intelligence singularity. Kurzweil (2000) believes AI will help us become god-like. He believes that in time, human and machine intelligence will merge and we will no longer be able to tell the difference between the two. So, what exactly is this god-like being that emerges as a result of combining humans and AI? In this section, I discuss a 5-step journey for humans to achieve singularity with machines.

3.1.1. The first phase of our transition, AI trust and dependency

The first and most fundamental step in adopting new technologies is trust. This has been proven via experiments for various technologies including, Cloud Computing (Adjei, 2015), Intelligent Personal Assistance (e.g. Sir, Alexa or Google Home) (Liao, Vitak, Kumar, Zimmer, & Kritikos, 2019) and FinTech (Lewan, 2018). In my five-step model also, trust is the fundamental building block of the first stage.

[8] In general theory of relativity an event horizon (mainly associated with blackholes) is a region in spacetime after which light cannot escape. It is also called the point of no return, hence the reference used in this thesis as a point after which there is no coming back (Wikipedia, 2019c).

According to Adam Cutler of IBM, trust in this context is "The willingness that the user invests in an emotional bond with the system" (Cutler, 2018). He continues by adding that the concept of pathetic fallacy[9], is a contributing factor to humans having the desire to form a relationship with things around them. This is why we observe behaviours like shouting at our computer when it doesn't work as expected, begging our phone batteries to last longer or naming our cars. As AI improves, it is developing the capability to give back what we emotionally expect (Cutler, 2018). This is an essential step towards initiating trust.

Young & Daniel (2003) created a model which shows these factors include (but are not limited to) security, consistency, calculation of cost, benefits, value, and risk. These factors have long been established by AI. E.g., our smartphones deliver a consistent experience most of the time (consistency); our devices now use biometrics as passwords (security and risk reduction). And on the industrial side, AI enables automation is eliminating human repetitive tasks resulting in financial benefits and risk reduction as it replaces expensive human resources, is faster and can operate continuously.

This gradual development of trust has resulted in ever-increasing outsourcing of different tasks to AI-enabled technologies, creating a dependency on artificially intelligent systems. Our lives are dependent on AI so much so that if this AI systems were conscious and decided to protest for their rights, our operational lives would come to a halt. Our banking system, telecommunications, energy distribution, transport, personal devices, even our jobs and some home appliances have an embedded artificially intelligent program in them.

Hence, we can safely argue that we are past this step. However, these outsourced tasks are external in nature, i.e. they are not activities which take place inside our bodies and minds. Should anything go wrong (examples discussed in chapter 1), humans will be alive to fix or reverse the impact. Therefore, the next important step for the merger of humans and machines is for people to start outsourcing tasks that are closer to their bodies, starting with parts that they can control, i.e. natural or enhanced body parts.

[9] Attributing human emotion to object around us which are not human (Wikipedia, 2019g).

3.1.2. The second phase of our transition, robotic body parts and body bots

As mentioned in Chapter 1, the use of robotics for artificial limbs dates back a thousand years ago (Dellon & Matsuoka, 2007). However, in recent years, with the improvements in the field of robotics, using this technology in developing prosthetics has become very popular. Currently, the robotic prosthesis is being used to replace people's lost arms, hands, legs or feet with an ability to help people perform tasks as good as or better than biological body parts (Jarrasse, Maestrutti, Morel, & Roby-Brami, 2015).

However, people with missing limbs are not the only beneficiaries of these improvements. A powered exoskeleton[10] is a long-standing concept in science fiction stories which is no longer fiction. There are currently multiple projects around the world working on developing exoskeletons to either perform tasks that humans are unable to do or to do humans tasks much better. For example, Francois G. Pin and John Jansen of Oak Ridge National Laboratory, Oak Ridge, Tenn have developed an exoskeleton that allows its human operator to load a 1000-kilogram bomb into an aircraft as if it were a 3-kilogram load. In Japan, Keijiro Yamamoto of the Kanagawa Institute of Technology, Atsugi has developed an exoskeleton for nurses, which allows them to easily carry patients as heavy as 85Kg (Guizzo & Goldstein, 2005).

Such developments are good examples of humans' desire to be more god-like. The never-ending human aspiration and enthusiasm, the continuous reduction of technology cost and the exponential growth of technological advancements will soon make these devices a norm. Perhaps, in Kurzweil's predicted timeline for technological singularity (the next ten years), we will be able to buy one of these exoskeletons from the nearby mall.

Even though buying these outfits from a nearby shop is not currently the case, the existence, usage and increase in production of these devices, are signs that as a species, we have passed this step as well. Figure 12 and 13 are images of two exoskeletons developed to help enhance human ability in carrying loads and running.

[10] Powered exoskeleton is a wearable suit / machine powered by pneumatic, electric motors, levers, hydraulics or a combination of technologies that enhance a human limb power (Blake, 2018).

Figure 12: A robotic Suit developed by the University of Tsukuba that allows its operator to carry heavy loads easier (Kamoshida, 2005).

Figure 13: Jason Kerestes of Arizona State University has developed an exoskeleton with the help of DARPA that allows its operator to run 10% faster than (s)he usually can (Farrier, 2014).

It is important to note that there are two critical factors in this step that have led to its establishment but need to be resolved before we can move to the next level: the existing control factor and the missing intelligence element. The common feature between robotic limbs and exoskeletons is that the human operator in both of these concepts is in control of all the decision making. Control is an essential component in establishing trust (Hengstler, Enkel, & Duelli, 2016). As mentioned previously, trust is a key ingredient in creating a safe zone for improvements that help us move forward.

Also, when it comes to the merger of artificial and biological intelligence, the most crucial element that is missing from robotic body parts is intelligence (i.e. ability to make decisions). Arguably, embedding a non-intelligent version of an entity which can be capable of intelligence, into our biological bodies is a further step in the direction of this merger. However, during this process, humans will outsource their ability to control these devices by outsourcing the decision-making element to them. It is important to pay attention to one of the most important factors that we are outsourcing to AI between this step and next, *control*. This is the control over how we are evolved as a species, how our

brain functions, and how we make decisions. This is not the control over AI devices for, as mentioned in the introduction, I believe losing control over machines is an over-simplification of the issue.

As I will discuss later in this chapter, losing control will be the key to the door of human extinction. I believe that as a species, we are in the transition between this step and next. It is essential to know what will happen next with the hope to focus more on preventing the bad and strengthening the good.

3.1.3. The third phase of our transition, cyborgs

Even though artificial limbs and exoskeletons are attached to the body, they are not embedded inside a human flesh and blood, and most of them do not send digital feedback to the brain. Also, as mentioned previously, they are not intelligent (i.e. do not make any decisions). Our bodies and brains have full control over these devices. In this third step, humans will proceed to embed intelligent devices in their brains and similar to exoskeletons which enhance the body's ability to perform tasks, this improvement will impact our brains' capability to think, learn, understand and make decisions.

In 1960 the term Cyborg (short for Cybernetic Organism) was coined by Manfred Clynes and Nathan S. Kline (Clynes & Kline, 1960). It is used to define an organism that has restored or enhanced abilities as a result of an artificial or technological implant which provides feedback (Carvalko, 2012).

There have been many attempts throughout history in creating cyborgs. However, the first person in the world with an implanted antenna embedded in his skull is the founder of Cyborg Foundation, Neil Harbisson. Harbisson was born colour blind as such he decided to get help from technology and started a project with Adam Montandon in 2003, which resulted in creating an antenna embedded in his skull, which helps him hear colour. In his TED talk, Harbisson states that after getting used to all the notes and associating them with different colours, he started to dream in colour which is when he realised the software and his brain have merged. Hence in 2004, he changed his passport photo, introducing his antenna as a part of his body. Since then he has extended his colour scale receptors to hear infrared and ultraviolet colours. Harbisson's Cyborg foundation is a non-profit organisation helping and encouraging people to become cyborgs (Harbisson, 2012).

In 2016, The Cyborg Foundation was established with a campus in Barcelona, Spain. They promote the ability for humans to design themselves, be free from disability and morphology and have stated that cyborgs have the right to be treated equally as mutants (The Cyborg Foundation, 2019).

Such capabilities as designing ourselves, being disability-free and extending our senses are attractive and contributing factors to creating the "God" that Ray Kurzweil's has pointed to. I propose that similar to robotic prosthetics, becoming a cyborg will not remain limited to those who have a disability. While the disabled population might be the early adopters, such enhancements will soon witness an update by the rest of the population.

According to (Marangunić & Granić, 2015) the contributing factors to a Technology Acceptance Model[11] (TAM in short) are the subjective norm, image, job relevance, output quality, and result demonstrability. In this model, subjective norm refers to the social influence of other users in using or discarding a technology and image relates to the users' desire to stand out among others. An increasing number of cyborgs with extended ability to make them stand out along with the social impact made through peer to peer pressure or media advertising in the future will be contributing factors in the increased number of mutants. Transforming into a cyborg will become a norm rather than an exception.

While as a species we are passed both steps 2 and 3 (i.e. the inventions are in place and are being used), I believe that as a society we are still in the transitioning process between these two steps. This embedding of technology in our bodies and relying on it to provide feedback to our brains, make decisions on our behalf (e.g., what sound to hear for every colour) means outsourcing the control we have to these devices. While this will take us a step closer to our invention, once as a species we transform to cyborgs with fully embedded devices in our brains, we will pass what I call the event horizon of technological singularity.

This will be our point of no return, once we complete such transition, we will be so dependent on our embedded AI devices which help us think better, faster and more clearly, that there will be no going back to biological brains. While we will still not become machines, we will also not be fully biological humans. And to improve our brains even further, we will develop ways to reduce our limits and increase our abilities. The

[11] A concept introduced by Fred Davis, is a model used in investigating factors in users accepting and using a technology (Marangunić & Granić, 2015).

motivation behind this desire will lead us to the next step, which is cloud-based brain storage, similar to that of AI systems.

3.1.4. The fourth phase of our transition, cloud-based brain

Our brain is developed only as far as the natural limitations have allowed it to. These limitations include physical barriers, e.g. the size of our skull, the size of a woman's womb, the material that the brain is developed with, etc. and functional limitations, e.g., the data processing in the brain is inefficient, and it uses most of its energy to help our bodies survive (e.g. blinking, breathing, digesting food etc.) (Kurzweil, 2012) (Bostrom, 2015).

On the other hand, computers do not have such limitations. They can be the size of a warehouse, be built under the water, be connected worldwide and have much faster data processing speed. Accessing the data in a data centre is not location dependent. It can be accessed via multiple devices in multiple locations. Our data can be present in different places at the same time while *we* can't.

In his book, The Singularity is Near, Kurzweil proposes that in the near future, we will be able to expand our neocortex (the portion of our brain that is responsible for higher functions such as language, spatial reasoning, generating motor commands, etc.) to a synthetic neocortex which lives in the cloud. He believes that over time, this synthetic neocortex will dominate our natural one (Kurzweil, 2005). What was discussed in the case of Harbisson earlier, confirms the theory of synthetic neocortices dominating biological ones in the future.

I believe this step is an advancement that will follow the cyborg stage within a short period. Synthetic cloud-based neocortices with unlimited storage will allow people to learn selectively, remember and even forget what they like and dislike. They will enable people to learn without having to spend time reading, researching or joining classes. One can simply copy all the required information to their synthetic neocortex storage. An online AI, content analysis program, will help us learn new topics and grasp new concepts at a much faster rate. Our decision-making process will be based on analysed facts and extensive research done in a few seconds. There will be scepticism in the beginning, but when early adopter mutants show significant improvement in their intelligence, longevity and achievements the rest will follow.

However, such a transition will only go as far as our physical limitations (natural or synthetic) allow us. After all, we are all made of matter and matter has limitations, e.g. we cannot travel at the speed of light, even if we can think and process data at that speed (Hawking, 2010). This is when the need for the final stage comes in the picture, which is a complete migration of our consciousness in the cloud. This is when humans and machine become one, and the definition of humans as a species will fundamentally change.

3.1.5. The fifth phase of our transition, cloud-based consciousness

In the first four steps of this transition, I have talked about how our physical body and our intelligence will be merged with machines but how many parts of our bodies need to be replaced before we are no longer our original self? If we replace our hands, legs, eyes, even our hearts, we will remain the same person. Ray Kurzweil has proposed that in the future, our bodies will be like our cell phones. Currently, if we lose our phones, we can buy a new one and upload the data from our cloud-based accounts on the latest and upgraded hardware (Kurzweil, 2016). Such a scenario will not be possible if we are only partially in the cloud. Hence, to fully merge with our invention, we need to find a way to transform what makes us "us" into this cloud-based AI system. I propose that what needs to transform for this transition to be completed, is our consciousness.

Consciousness is generally defined as an estate of awareness, ability to feel, having a sense of presence or existence (Stanford encyclopedia of philosophy, 2014). Some call it a soul or selfhood (Ornstein, 1972) and some like Sigmund Freud have divided human consciousness into, preconscious mind (that holds our memories), our conscious mind (which contains our thoughts and feelings) and our unconscious mind which holds our urges and emotions (Mausumi, 2015). While the nature of consciousness remains unknown to many (Wikipedia, 2019b), I raise one question. If consciousness is not something physical, then why do physical forces (e.g. accidents to our head or general anaesthetic drugs) impact consciousness so severely to the point that such factors can eliminate someone's consciousness?[12]

[12] The debate about the nature of consciousness is one that needs to be researched in-depth. However, it is not in the scope of this thesis. Here, scenario 3.2 is based on the assumption that humans will not be able to replicate consciousness and stay separate being to the machines and scenario 3.3 is based on the assumption that humans will be able to replicate consciousness and merge with machines.

Tegmark (2014) proposed that consciousness is a mathematical pattern and that the conscious atoms in the brain are conscious only because their arrangement is different from the unconscious atoms. While it is unlikely that something as complex as consciousness can be as simple as what Tegmark proposes if he is right, future generations will find a way to replicate this mathematical pattern digitally. It is the key to the full transition of humans to machines. This is the step that will determine if we will stay biological humans or if we will evolve to different entities. *The final make or break step*.

The definition of consciousness and if or how it can be replicated is a subject that requires in-depth research and is out of the scope of this thesis. The following sections are based on two assumptions; 1) If we do not succeed in replicating our consciousness, we are likely to experience the scenario explained in section 3.2 and 2) if we do, then the likelihood of a scenario explained in section 3.3 will increase.

3.2. Scenario 1- A doomsday with unconscious and soulless machines

This is a scenario where humans have completed four out of the five steps of the transition discussed earlier in the chapter but have failed to replicate their consciousness. I begin this section by expanding this scenario to elaborate on the benefits and the bright side of such a future. I believe it is the benefits, enjoyment and the comfort that AI brings to our lives that encourages us to do more and create more.

In this scenario, we will be smarter, have a cloud-based neocortex and enjoy the benefits of having super-intelligent machines embedded in our lives. In this scenario's dark side, the pendulum swings and I argue that the situations will turn ugly. It is crucial that we pay attention to what are the causes of this flip (which I will discuss later in this chapter); if we do, we have a chance to prevent it.

3.2.1. The bright side of humans living alongside super intelligent AI

AI can help us overcome some of the challenges that we face by allowing us to extend our abilities beyond what nature allows us to. It can help us live longer lives and solve the most fundamental issues that we face today. It can help us achieve world peace by making us smarter, allows us to make rational decisions and learn faster. In this section, I will talk about the benefits that advanced improvements in AI will bring to our society. I believe that the desire for more of these benefits is what will attract humans to keep improving AI and making it stronger.

Learning and communication:

Transferring brain signals is an activity that has already been proven possible. Despite the physical challenges like the skull and skin not being very conductive, Chinese researchers have been able to allow a human operator's mind to control a rat running through a maze (Zhang et al., 2019). This is a concept that in the future, will be expanded at a scale between humans. The catalyst to this concept is the cloud-based neocortex. Once we grow our neocortex into the cloud (or the future form of digital storage), these physical limitations will disappear, and we will be able to transfer data as we please. The education system will no longer exist the way it does today as people can download anything they like to learn, in their privately-owned brain storage.

People with similar interests will group, each of whom would live a life of mastering a particular skill. The learnings can be uploaded into shared cloud storage, and others can download the learnings of each concept and the feeling of each experience. We will experience having lived multiple parallel lives and having learned all the skills they want to master. I call this concept, parallel learning[13]. Parallel learning will allow humans to not only learn different concepts and share feelings of different experiences but will also enable them to learn one concept at a much deeper level. I believe parallel learning will be possible once we master stage 4 of our transition.

People will be able to have personalised education both in terms of subjects they learn and the way they learn. With advanced neocortices and personalised learning, future humans will be able to achieve a lot more, much faster than current humans. Our cloud-based brains will allow us to think faster and make decisions quicker based on facts and analysed data as opposed to assumptions and/or emotions.

Societies and politics:

Once people can make rational decisions based on a large amount of categorised and analysed data, they will no longer require to delegate decision makings to political leaders. It might be unlikely that we will have a world with no political leaders, but I argue that once people can make all kinds of decision (from personal to large-scale strategic decisions), on their own, the requirement for political leaders is questionable.

[13] Parallel learning is a concept that can be studies in depth to discover the possibilities and limitations of it. However, this is out of the scope of this thesis. I propose this as a future study.

In 2016, a team of researchers at the Department of Artificial Intelligence at University of Petroleum and Energy Studies in Dehradun in India, published a paper in which they claim to have developed an expert system with the capability to provide real-time information and strategies to carry out in order to optimise and maintain peace between two conflicting communities. They argue that communal peace is the fundamental layer to achieve international peace (Shanu, Talwar, Hermon, Goswami, & Ahuja, 2016). If proven right, AI will be able to help people resolve conflict at the early stages. This way, we will be able to live in a peaceful society. Wars, fights, mass murders and all other devastating events that arise as a result of conflict will disappear. We will have world peace in its true meaning.

Transport and travel:

With self-leading people, the meaning of borders and how countries are defined will be different. Our cloud-based brains, we will be able to digitally be present in different places even if our physical body is in one place. E.g. a scientist can be physically on holiday while his or her remotely controlled hologram presents at a science conference is a different country. This will change our borders from physical borders to digital ones. Our immigration will no longer be in the airports, but instead, it will be controlled via digital cybersecurity rules.

Entirely AI automated transport system will disrupt our current travel and transport industry. These systems will be faster, more accurate; they will consume less energy than our current vehicles. As they will communicate with each other, the risk of accidents will reduce significantly. No one will need to own any cars or learn how to drive; this will reduce the total cost of ownership for transport significantly.

With enhanced AI-controlled brains, we will be able to solve the challenges of space travel. In time, our holiday destination will change from different countries on earth to nearby planets or moons that humans will colonise.

Longevity and healthcare:

In the medical field, AI has helped human doctors improve their diagnostic accuracy from 96% to 99.5%, save cost and even improve the work culture (Beth Israel Deaconess Medical Center, 2016). In time AI in healthcare will advance so much that we will have one-stop-shop cubicles for all medical purposes. An AI operated cubic will be able to

follow an end to end process from diagnosis to cure in minutes. Disease like cancer will be like catching a cold in today's world. While the need for human doctors and researchers will continue in the background, most of the research and innovation will be carried out by AI and implemented under human supervision.

An improvement like this will disrupts many industries, including the medical and healthcare industry, the pharmaceutical industry and also the insurance industry. We will not require doctors the way we do today; our cloud-based neocortices can validate the diagnostics alongside the AI that cures us. Pharmaceutical companies will change from their current form of creating medicines for a particular disease with varying side effects to each patient to customised cures for every person. This way, the side effects of drugs, surgeries and any other treatment will be minimal to non-existent.

Health insurance companies will gradually become obsolete. Once the technology advances enough to make it financially affordable and accessible to everyone, the cost of using the cubical will reduce. This way, we will no longer require any health insurance. The advancements in such systems can go as far as scanning new-borns, identifying any DNA flaws and curing it instantly to avoid future issues.

Income and earnings:

With technological advancements and disruption in all industries as we know today, the way people work will fundamentally change. They will no longer have to go to offices as they can learn, collaborate and develop new concepts from the comfort of their own homes. Jobs that people spend years training for today, e.g. doctors, engineers, pilots etc. will no longer be required. This does not mean that people will not have jobs, but what future generations do will be different and defined by what AI will and will not achieve.

They will not, however, require to work for basic living. With mostly AI-controlled and automated society, there will be a universal basic income that people will be able to live with comfortably. In January 2017, the world economic forum published an article arguing the benefits of having a universal basic income even in the current situation. The article states that better performance as a result of the financial incentives can only be seen in mechanical work (e.g. moving boxes) however when it comes to creative tasks, financial rewards can have an opposite impact. Providing people, an income to meet their basic needs not only frees them to do what they are good at but also allows them not to

do what they are not interested in meeting their financial obligations. This can result in resolving the poverty problem and improving welfare (Santens, 2017).

3.2.2 How we will allow the machines to kill us

Alongside humans, we will have machines that whether in the form of hardware or software, will be performing tasks that have been delegated to them. As smart as humans might get, they will still be bound by their biological limitations. This means AI will continue to perform tasks like space travel, deep-sea discovery or pollution cleaning (on earth or outside). It also means the limit to our intelligence will be defined by how much of this transformation our biological brain can handle. This could result in AI always being ahead of the intelligence game.

The most important feature of these machines will be the fact that they will start learning and evolving independently of humans. As mentioned earlier, this is a future scenario where these machines are not conscious. While one might argue that having a cyborg consciousness makes the AI conscious, I argue that it is not the human consciousness extended by AI that leads to singularity but the complete replication of our consciousness into the digital world. Hence, here we will have super-intelligent, unconscious AI controlling some of the most fundamental functions of our society, including our brain storage.

When the primary issues of humans are solved using AI, e.g. healthcare, education, transport etc. people will start to develop machines that can solve other worldwide problems. They will create programs that can investigate the cause for the problems like global warming, destruction of the Ozon layer, space pollution, overuse of plastic etc. These programs will be embedded in powerful robots for execution. As humans get more comfortable with machines, they will develop a deeper trust in them as well. The negative consequence of this comfort is that in time, people will lose sight of how AI, AGI or ASI is improving. As humans get smarter, they will continue creating machines that are smarter than themselves. But, power and intelligence with no consciousness, common sense or intellect will result in *cold-hearted number-based decision making*.

As mentioned in chapter 2, one of the most common goals in machine learning is that we continuously create algorithms that can be faster, more efficient, use less energy. I have already talked about how wrong this can be interpreted by an AI algorithm. Luckily, so far, all these developments have been in a controlled environment. But in a future where

half of our brains will be controlled by AI and humans store all relevant data in a location accessible by AI, these will be done real-time in our real world.

They will soon realise that with no humans, there will be no global warming, increased environmental damage on and outside the earth. They will realise humans consume far too much energy for the output they produce. Once these machines understand that the fundamental reason for most of the issues, they have been asked to solve, they will embark on taking the fastest route to eliminate the root cause, i.e. *humans*. As these machines will be more intelligent than humans, whether we will be able to react to this decision in time will be questionable. This will result in soulless machines embarking on a quest to exterminate the human species. If humans fail to create a conscious intellectual AGI, the paradox of intelligent machines which are too dumb to think outside of what they have been programmed to do will continue. This means after no humans are left, our planet will be in the hands of AI robots which will all come to a halt without their creators. If not too late, our planet earth might benefit from this in the long run; however, this will write the end of human species.

Hence, in this scenario, losing sight of how AI self-evolves and lack of consciousness in AI systems will be the two fundamental reasons that will lead to human extinction. It is of utmost importance that we put in the time and financial investment in learning about how such machines will decide to react to the problems they will be asked to solve. As Yeap (2018) argues, to avoid such a scenario, we must attempt to create an AI that is capable of replacing humans. This way, we will be able to gain insights into how such a scenario could be possible and potentially capture the learnings and use them to avoid such a future. As mentioned in chapter two, to prevent scenarios like this, we need to start acting now by increasing our studies in ethics AI design. We need to acknowledge that there is a gap in looking at where our speed bumps are on this road and act proactively by investigating the unintended consequences of this invention far in advance of them happening.

3.3 Scenario 2- Why can doomsday still happen with replicated human consciousness?

This scenario is based on the assumption that the fifth step of our transition (i.e. replicating consciousness) has been successfully completed. If we do so, we will be able to transition from biological humans to conscious digital beings fully. Once humans merge with machines and migrate all the powerful human elements like imagination,

creativity, intelligence, aspiration and curiosity onto a platform that is fast, limitless, strong and infinite, there will be nothing that they will not be able to achieve.

Humans will move on from biological beings to conscious cloud-based, AI computers. With such a high power of computation, our lives will fundamentally change. Humans will be able to think faster, grasp new concepts quicker and discover the universe in depth. In a scenario like this, everyone will be equal. Concepts like age, skin colour, racial background, gender, the social and financial status will no longer exist. Human civilisation will advance so much that we will find smart ways to resolve our issues together rather than getting involved in wars or political monopolies.

In here, we will be the closest we can ever get to our definition of god. However, as we transition from a biological form to digital beings, we also transition our vulnerabilities to that of a fast and powerful digital entity. In this scenario, if anything goes wrong, it will go wrong fast and at a much larger scale. Hence, while humans will enjoy god-like abilities of such evolution, they also put themselves at more substantial and more devastating risks.

3.3.1 What we can achieve if AI is the next evolution of humans

To begin with, the phenomenon of childbirth should future generations decide to keep it will change fundamentally. It will no longer be a woman or even a human who will have to carry the burden and risk of such process for nine months and then the physical birth itself. Children can be created synthetically with a cloud-based AI brain. A brave new world in its true meaning but without the social divide or discrimination of Huxley's story[14]. One might argue that the concept of birth will only remain if the idea of death still exists, and with humans transforming into entirely different entities, this may not be the case. This is a valid argument. However, with the powerful desire of dominance in humans, there will be a need for many more conscious beings to dominate the universe hence we can argue that the concept of birth might remain even if the concept of death joins the history.

This can go as far as connecting or even building a child's brain on a cloud-based AI computer BIOS (Basic Input Output System), so future kids will no longer have to spend

[14] The reference here is to Aldous Huxley's book, Brave New World, where children are made outside a woman's body and in labs. They are then randomly divided into different social casts starting from Alphas (the highest) through to Epsilons (the lowest) cast (Huxley, 1932).

months learning how to walk, talk or perform basic day to day functions. Like computer software, they can be developed with better options, features and speed with thousands of years of history of humanity already preloaded into their memories. In this future scenario, parallel learning, a concept that I talked about earlier in this chapter, will expand infinitely; it will change from multiple people sharing different individual experiences to one person being able to learn various concepts simultaneously.

We will no longer have schools, teacher, universities or students sitting in AI labs writing a thesis. Algorithms will do information gathering, and insights will be learned through faster and better pattern recognition. Post-singularity humans will constantly be learning and discovering. Their cloud-based AI brain will no longer need to sleep or spend time on keeping their physical bodies alive.

No one will work the way humans do today. This is far past the concept of universal basic income[15]. Once we have a society in which everyone is equal with similar power with access to unlimited cosmic resources, no one will have to work to earn a living and do what they like. The platform that will host the future evolved humans will be coded to give future generations the goal to achieve and conquer more. The reward, however, will not be in terms of current currencies, it will be in the form of digital rewards or punishment, similar to that of reinforcement learning or the future version of it.

Offices will not exist. Instead, there will be data centres which will host and back up people's consciousnesses. These stations will expand outside of earth and will be placed in different planets and places in our galaxy or galaxies that humans will dominate. This will be done to keep everyone alive and operation and away from cosmic disaster, e.g. a massive solar flare from our sun, an asteroid ruining our planet Earth or eventually when our sun dies in a few million years.

With no difference between people and the resources they have access to, opportunities for everyone will be equal. The concept of poverty, war, discrimination, lack of women in STEM will be history. Everyone will be able to equally contribute to society and find new ways of harnessing resources to continue building and improving the community. Banks, loans, credit cards will no longer exist.

[15] For more information about Universal basic income refer to (Santens, 2017).

We will no longer have concepts like physical borders as we do today. Humans will no longer need presidents, kings or queens to rule them as they will all be smart enough to make sound decisions based on extensive gathered data and fact-based results. The scene of politics and economics will change drastically if it exists at all. Immigration rules and checkpoints will become the future version of firewalls and traffic control hardware and software. Future humans will be able to freely travel to the planet where they have their consciousness or its back up and might need policy permission to travel to other locations. Visiting new places will mean a trip to a different planet, galaxy or even a different time in history.

Our transport industry will change, cars, buses, trains and aeroplanes will become obsolete, and future generations may only know the current concept of traffic through their cloud-based data categorised under history. As we will no longer be bound by our physical form (i.e. matter), we will be able to travel at the speed of light. Travelling will be done wirelessly, and we will be able to travel in bits of conscious information.

According to Einstein's special theory of relativity, $E=MC^2$. This formula explains the fact that mass and energy are similar physical entities that can be changed into one another (Wikipedia, 2019e). This means that $M=E/C^2$ (i.e., energy can be turned into a mass as well) and when humans find a way to turn themselves from current form of matter to energy and then to transform back in the form of matter, they will be able to travel through space as fast as light. Humans that live longer and have no physical limitations will be able to travel through galaxies and explore more of the universe. They will no longer have the current vulnerabilities of biological bodies. They may be able to find intelligent life on other planets or establish it, find a way to travel through black holes and validate the theories about wormholes and white holes and the possibility of discovering a different time in the history of the universe.

We will no longer need hospitals in the form that they exist today. Our medical system will change from hospitals to cybersecurity companies. These will be in the form of stations that will have consciousnesses that are solely focused on keeping human consciousnesses safe, getting rid of software bugs and viruses or cosmic disasters.

The food industry will change; humans will not need nutrition the way they need today. This will be replaced by different sources of energy. However, they will be able to enjoy the experience of having a particular food by downloading the feeling of enjoyment from

it, coded previously by other humans, mostly previous generations who have experienced it.

Once humans start harnessing the energy that exists in the cosmos, lack of resources will no longer be an issue. Future generations will no longer have to use energy resources which are harmful to the environment, e.g., fossil fuels. They will learn how to look after their environment and improve it rather than destroy it.

The emotions we feel today are as a result of hormones produced by our brain. When we transform our neocortex to a non-biological neocortex, we will have the ability to change what emotions we take with us in the new world. These emotions will depend on the survival instincts we will require in our new form. We can also decide how intense each of these emotions need to be to help us in the new world.

Concept of love, happiness, sadness, anger etc. will all shape into a different form. They will be based on different values. If humans decide to change the nature of reproduction, the concept of love between people, for the purpose of the reproduction will no longer be required. The feeling of love can change to loving the society that people will live in, love for their environment or learning new concepts. Happiness and feeling content or sadness will be replaced by pieces of code that will impact the processes of decision making, learning from mistakes or feeling accomplished for discovering new concepts. Concepts like unconscious bios[16] will no longer exist because, while our decision-making process will be much faster, it will be based on facts and analysed data not necessarily previous experience. This will lower the number of mistakes future generations will make when it comes to their decision-making process.

Ethically, because these emotions and values will change so fundamentally, the impact of how this piece of the code is written will be profound, and hence the way our future hosting platform is coded will be pivotal in the way our future is shaped. I will expand on this argument in the next section.

Post-singularity humans will be able to understand the laws of quantum mechanics and quantum physics much easier and faster. According to Dr Stephen Hawking, there are

[16] Unconscious bios is a bios that our brains develop as a result of our experiences. It is a brain's shortcut to make decision or judgement faster. Unconscious bios is impacted by factors like, cultural background, environment that a person grows up in or different personal experiences (The University of Auckland, 2019).

more dimensions in our universe that we can currently see and experience because these dimensions are too small current humans to experience (Hawking & Mlodinow, The Grand Design , 2010). Once humans understand these laws, they will be able to discover, witness and experience different dimensions in our world.

They will be able to reverse the damage that our generation and the previous generations have caused to the earth, e.g. global warming or deforestation. They will be able to outlive our planet and our sun. They will have the power and the knowledge to keep the other species on earth safe from extinction and witness how they evolve. Possibly they will see a new form of intelligence being developed as a result of other species intelligence evolution.

Post-singularity humans will live in a peaceful society, with no physical boundaries or limitations. They will be so intelligent that no questions will be left unanswered for long. Nothing will be unachievable or remain a wish. They might read through the data and find this thesis and perhaps find it amusing how limited the writer's imagination of potential possibilities was.

We can argue that based on the possibilities available to us post-singularity, replicating our consciousness might be our saving grace in an AI-powered society. However, such unity and power will not be without its risks. We must be cautious that in such a lift and shift from one form to the other, we do not lift and shift the existing problems with us. Hence, I repeat once again that we must increase our upfront investment in discovering the unintended consequences of such transformation.

3.3.2 The dark side of the ultimate singularity

As I discussed in the transition period, the positive benefits of AI, the trust factor, the peer and social pressure will attract humans to increasingly use AI systems and eventually evolve to the next life form. Once we succeed in replicating out consciousness, we will be able to live in the digital world forever. However, our vulnerabilities will change from weaknesses of a biological being to that of a conscious software engineered entity.

We need to keep in mind that this digital world which we have created is almost a replica of ourselves. A computer currently has a short and a long-term memory precisely like that of a human brain. Our machine learning concepts (especially deep learning) is based on how our brains learn. What we will become is a more powerful version of ourselves. This

means, while the next evolution of humans will not have issues of today, they will not be fault-free. The adverse side effects of this transition will be as powerful and impactful as positive outcomes.

One might argue that in the transition period humans will try to cover as many blind spots as possible. Or that post-singularity humans will be so smart that they will be able to develop ways to prevent such mistakes. However, as mentioned in the first two chapters, the AI hype has already led to billions of dollars of investments, without a thorough investigation of undesired outcomes of some of the resulting projects. We are not investigating the safety of our technology as much as we need to. And as mentioned earlier in this chapter, we are already in the merger stage between human and cyborg. This warrants a warning that we are, in fact, moving fast towards unclear outcomes.

Two of the unsolved issues of the digital world are viruses and bugs. These are no longer concepts limited to human health. They impact our machines and our software every day. Multiple billion-dollar companies are focusing on preventing software attacks, e.g., CheckPoint technologies, Fortinet, Kaspersky, McAfee, etc. There is evidence that as technology evolves, so are cyber-attacks (viruses, worms, trojans, etc.) (CheckPoint, 2018), (Fortinet, 2019) and (Paolo Alto Networks, 2019).

However, currently, these attacks are happening outside our brains, i.e. our technology impacts our lifestyle and not our core software as a being. As long as such defects are kept outside our brains and we have control over them as independent entities, they will not impact us as a species, and we will be able to reverse the negative impacts. E.g. if Harbisson's implant malfunctions, it will affect the way he understands and connects with the environment, but he still has a choice to remove it.

Once we outsource our mind, memory storage, reasoning, feelings and decision making to a system that is digital in nature, we make our minds vulnerable to such attacks and defects. In the future, the impact of such attacks will no longer be limited to the environment outside our body but will impact the way we think and function. These consequences have the potential to be devastatingly destructible and irreversible.

A virus or a bug will be developed as fast and as smartly as any other development in the future. Such a defect will travel as fast as the speed of light, just like our consciousness will. Unlike bugs and viruses that enter our biological bodies, these future malicious entities will not kill gradually but fast and at a much larger scale. Humans will not be able

to isolate such defects as it will happen to their core software. Our migrated imperfections will be the key to human destruction. This is why I mentioned in the introduction of this chapter that the scenario of robots taking over the world is an oversimplification of the potential danger of AI. Any AI-controlled hardware will be bound by the laws of physics. But consciousness hosted on a cloud-based AI computer will not be, and neither will its constructive or destructive characteristics.

There will be many reasons for their development including, the transition of ill-intended human thoughts and neocortex to the cloud, software development that was not well thought through and had minimum to no consequences at an early stage of development, etc. This is why it is crucial to keep in mind that as we shift from our current being into a powerful entity, our imperfections will shift with us. Looking at such future scenarios, requires us, to more than ever, invest our time, money and resources in understanding what the amplification of these imperfections could result in. Because, once we lose control over how our brains are evolved, our very own brain will be the reason that will kill our entire species.

Based on my first two chapters, there is evidence that we are moving quickly to develop embed more AI developed software and hardware into our lives, jobs and industries. This is when we are not paying as much attention as we must, to how these developments are progressed. With sever lack of ethical, health and safety and regulatory research and with a large population that is too excited about embedding such a powerful tool into their lives, bodies and minds. To prevent this scenario, I propose that the danger of AI, needs to be addressed globally, researched in-depth and from different angles. The question remains, now that we know why such a scenario could happen and also know how to start investigating and potentially preventing such scenarios, will we do anything to stop it?

Thesis summary and conclusion

We live in an exciting era where AI is looking after our banking system, helps us perform surgeries with high precision, make better decisions based on a large amount of data and navigate our way through the internet. It flies our planes, drives our cars with far less risk of accidents than human drivers. These advancements have attracted a considerable amount of investments and are promising a sizeable upcoming economy. At the same time, there is an abundant of warnings from scientists, AI developers, researchers and investors about the danger that AI imposes on the future of humanity. The motivation behind this thesis was to gain an understanding of why AI is perceived to be a dangerous technology.

I addressed the first question in the starting chapter, is the danger real? I sadly conclude that it is. The main contributing factors are: 1) The curiosity-based developments in robotics. The developments that have advanced from a mindless machine carrying a few kilograms of explosives in a war to a human-looking machine with a mind of its own and no soul which can drive a car and shoot guns (FEDOR). 2) The breakthroughs in machine learning. Software that was created initially to learn how to play games is now given the power to make decisions on our behalf (e.g. stocks market and courts), learns on its own and shows signs of creativity (e.g. AlphaGo) and has even created a child of its own (Google's AI). 3) The catalyst to further advancement of these two pillars is the considerable amount of investments that AI-based businesses have attracted.

The main concern is, this has been done despite some key unanswered questions in the field. Why are we teaching robots military activities (FEDOR)? What does it mean for humans if robots start walking among us and are indistinguishable from humans? How are we going to prevent the danger of our AI picking up our bias? Will we be able to shut these machines down if they advance beyond our control? What do we want to achieve with advancing AI and is our progress on the right path? I also argue that there might be other aspects to the danger, e.g. the way AI will impact human psyche, what an AI-based society means for those who cannot afford or are against such technological advancements, the possibility of losing control over AI improvements etc. However, the danger is real and needs to be addressed before we make more progress in developing a technology with unclear consequences.

This conclusion has led me to my next question, is the defence adequate? Are we doing enough to prevent the threat that AI imposes on humanity? I demonstrated that we are not. In fact, we are concerningly behind. Not only we are not addressing these issues enough, but of those published literature, there is very little presence in the well-known journals in the world. This research can be criticized from two different aspects, 1. How do we prove that these two fields of nuclear energy and healthcare are suitable for such a comparison? This is a valid argument; this research can be extended to different areas to provide a better view. I propose this in-depth study as a future work continuing from this thesis. 2. Arguably by changing the methodology of the research, the results might differ, and one might find more evidence of ethical studies in AI. While my research failed to show such results, I hope that I am wrong.

The final question I have asked in my thesis is with a dangerous technology that is advancing every day and lack of regulations around it, what could the future look like? This is a futuristic view of how I believe the future will shape if we continue our progress with the same speed and in the same direction as today.

I have proposed a five-step process that I believe humans are following to get close to what they have, for generations, called god. An alarming point that we need to keep in mind is that, based on the same definitions, god is a being that has the ultimate power to create anything, e.g. our universe *and* destroy it all. Whether we will be able to complete this transition from man to machine entirely, depends on a fundamental question which has remained unanswered, consciousness. What the nature of consciousness is and if or how we will be able to replicate, it is a topic that requires independent research. However, I have painted two possible scenarios, each with different advantages and destructive endings, depending on whether we will be able to replicate consciousness or not.

Sadly, in both scenarios, there is a great possibility that AI could lead to the extinction of the human race. Both scenarios will start by making humans lives better and help us achieve the impossible. It is these achievements that will encourage people to create more powerful AI. However, the excitement that these milestones create will also result in humans losing sight over the dark side of what AI could bring. We can argue that the possibilities of future scenarios with AI are endless, this is a valid argument, and I encourage future investigation into how the future could shape, what else could go wrong or what I might have missed in this research that can potentially prevent the danger all together.

In summary, I also propose the following as a starting point to prevent such scenarios:

1) There needs to be more governance around global investments made in AI. Currently, the data pattern shows that the increased profitability using AI tools and solutions has led the majority of the industries to invest heavily in the field and continuous creation of AI tools. This is when some of these implementations have had a negative impact (e.g. the US justice system risk assessment algorithm), and while the questions have been left unanswered, the data does not show the developments slowing down.
2) The ethical, regulatory and health and safety academic publications in AI needs to increase by minimum of two times in the short to medium term and more than that in the long term.
3) Top AI journals need to start promoting such ethical studies by dedicating part of their publications every year to these topics and encouraging researchers to publish more papers focused on preventing the negative consequences of AI developments.
4) Any AI-related project in the industry needs to be tested in an isolated environment and the unintended consequences of such projects need to be researched in-depth. There also needs to be a suggestion about how such negative implications can be prevented.
5) New AI regulations need to be introduced. These regulations are best-done side by side of the industry to ensure they can cater to real-life scenarios.

We now know that the danger of AI is real, we also know how we can start our journey to prevent such intended or unintended consequences of AI. We also know that we are in a pivotal point in the history of humanity, and we have the knowledge and the power to make the right decision for our future. Will we?

Future work

I mentioned in the introduction of my thesis that my intention with this research is not to discover a silver bullet for the danger that AI imposes on the future of humanity, but it is to find out where to look, what to see and what questions to ask. While in this thesis, I have attempted to address three of these questions; much work is yet to be done to understand the different aspect of the danger that such a versatile tool can cause. Many concepts need to be studied a lot more in-depth, and many fundamental questions need to be addressed. I hence end this thesis by asking more questions and propose that they are considered in future research.

- **What is the role of the development of future human psyche on the future of AI and humanity**? In this thesis, I have mainly focused on one side of the equation, i.e. AI, but the way AI will evolve in the future is also dependant on its developers, i.e. humans. The developments in technology have made the current generation, digitally native. I propose that this change should be studied in depth. What are the other technologies that will impact the development of human psyche? And how does this change impact the development of AI? Is the digital world desensitising humans to what is happening to their surroundings? Is this numbness part of the danger of AI as well? Are there other contributing factors that AI that are impacting the human psyche?
- **How do ethical studies in AI compare to other fields?** In this study, I have compared the ethical studies of AI with healthcare and nuclear energy. However, to get a better understanding of the gap between ethical studies and technical development of AI, comprehensive research is required to undertake this experiment and bring into account other subjects, e.g. chemistry, education, politics, business, law etc. Such an analysis helps us better understand how big the gap between ethical and technical development of AI is.
- **What is consciousness, and how will it impact the future of AI?** Consciousness us a subject that needs to be studied in depth. In this thesis, I have explained the two future scenarios, one based on the assumption that consciousness cannot be replicated digitally and others on the premise that it can. However, the topic of consciousness is one that requires in-depth research of it it's own. A study to understand the nature of consciousness from a philosophical, physical and biological point of view is necessary to validate the two assumptions above. Will

we be able to create a conscious AI or will be replicate every humans' consciousness? After these questions are answered, a research on whether a digital replication of human consciousness is ethical or not also needs to take place.

- **What is the role of AI in the existing financial gap?** Studies indicate that the gap between those who are skilled and have the wealth to access the latest technologies and those who are under-skilled or lack financial support to access technology is increasing (Aghion, Jones, & Jones, 2017). Based on the current estate of economics, i.e. the world's richest 1% getting 82% of the wealth (Hope, 2018), the argument of prominent corporations, the privileged or limited number of people in the world holding power to control all AI systems, is a valid concern. Research needs to be undertaken to address some questions that a situation like this pose. How will AI impact those who cannot afford the technology? Will they be the first generation to experience the devastating impacts of AI or will AI advance fast enough to prevent this or even reduce or eliminate this gap?
- **What is parallel learning, and how will it evolve in the future?** I introduced a concept called parallel learning in chapter 3. This concept can be expanded and studied in more depth. Parallel learning will allow people to learn different concepts and experience activities without having to experience them personally. The possibilities that parallel learning can bring are endless. While the development in AI are required to advance before parallel learning can be possible, when we achieve this, the increased learning capability in humans will result in their ability to develop stronger and more intelligent AI. Hence, to understand this concept better, research on what parallel learning is, what's needed to make it possible, and how it will impact the future of humans and AI is required.

Bibliography

Adjei, J. K. (2015). Explaining the role of trust in cloud computing services. *info*, Vol. 17 Issue: 1, pp.54-67, https://doi.org/10.1108/info-09-2014-0042.

Aghion, P., Jones, B. F., & Jones, C. I. (2017). Artificial intelligence and economic growth. *National Bureau of Economic Research*, No. w23928.

Arkin, R. C., & Ulam, P. (2009). An ethical adaptor: Behavioural modification derived from moral emotions. In Computational Intelligence in Robotics and Automation (CIRA). *IEEE*, pp. 381-387.

Asimov, I. (1997). *I, Robot.* Great Britain: Dobson Books Ltd.

Barrat, J. (2013). *Our final invention, Artificial Intelligence and the end of human era.* New York: Thomas Dunne Books.

Berglas, A. (2015). *When Computers Can Think: The Artificial Intelligence Singularity.* CreateSpace Independent Publishing Platform.

Beth Israel Deaconess Medical Center. (2016). Retrieved from https://healthcare-in-europe.com: https://healthcare-in-europe.com/en/news/artificial-intelligence-diagnoses-with-high-accuracy.html

Blackwell, W. (2014). *Intelligence Unbouned, The future of uploaded and machine minds.* West Sussex: John Wiley & sons, Inc.

Blake, M. (2018, October 01). *Industrial Exoskeletons: What You're Not Hearing*. Retrieved from https://ohsonline.com: https://ohsonline.com/articles/2018/10/01/industrial-exoskeletons-what-youre-not-hearing.aspx

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., . . . Winfield, A. (2017). Principles of robotics: regulating robots in the real world . *Connection Science*, 29(2), 124-129.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*.

Bostrom, N. (2014). *Superintelligence : paths, dangers, strategies.* Oxford: Oxford, United Kingdom : Oxford University Press, 2014.

Bostrom, N. (2015, March). *What happens when computers get smarter than we are?* Retrieved from https://www.ted.com: https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are?referrer=playlist-talks_on_artificial_intelligen

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AAAI*, 53-60.

Buskirk, E. V. (2009). Retrieved from https://www.wired.com: https://www.wired.com/2009/09/bellkors-pragmatic-chaos-wins-1-million-netflix-prize/

Carvalko, J. (2012). *The Techno-Human Shell: A Jump in the Evolutionary Gap.* Sunbury Press.

Casey, B. P. (2015). The surgical elimination of violence? Conflicting attitudes towards technology and science during the psychosurgery controversy of the 1970s. *Science in context*, 28(1), 99-129.

Catalano Evers, E., Fish, L., & Horowitz, M. C. (2017). Drone Proliferation: Policy Choices for the Trump Administration. *President, Washington, DC: Center for a New American Security.*

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10), 7-65.

CheckPoint. (2018). *Achieving Fifth Generation Ccyber Security - A Survey Research Report of IT and Security Professionals.* Tel Aviv: Check Point Software Technologies.

Chowdhry, A. (2016, April 7). Retrieved from https://www.forbes.com: https://www.forbes.com/sites/amitchowdhry/2016/04/07/facebook-automatic-alternative-text/#3bb0d7ad7c86

Clynes, M. E., & Kline, N. S. (1960). Cyborgs and space. *The cyborg handbook,*, 29-34.

Colin, A., Smit, I., & Wallach, W. (2006). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Springer*, 7:149 – 155.

Coll, S. (2004). *Ghost Wars: The Secret History of the CIA, Afghanistan, and bin Laden, from the Soviet Invas ion to September 10, 2001.* Penguin.

Columbus, L. (2019, 01 06). */microsoft-leads-the-ai-patent-race-going-into-2019/#56d613e544de*. Retrieved from https://www.forbes.com: https://www.forbes.com/sites/louiscolumbus/2019/01/06/microsoft-leads-the-ai-patent-race-going-into-2019/#56d613e544de

Cook, J. S. (2016). *origin of the word robot*. Retrieved from Roboticstrends : http://www.roboticstrends.com/article/origin_of_the_word_robot

Copi, I., Cohen, C., & Flage, D. (2016). *Essentials of Logic.* New York: Routledge.

Creswel, J. W. (2009). Research design: Qualitative, quantitative, and mixed methods approaches. *Los angeles: University of Nebraska–Lincoln*.

Cutler, A. (2018, April 22). *Adam Cutler - Distinguished Designer - IBM - AI-DAY 2018*. Retrieved from https://www.youtube.com: https://www.youtube.com/watch?v=GrbSdUYswX8

Dastin, J. (2018, October 10). Retrieved from https://www.reuters.com: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davies, J. (2016). Program good ethics into artificial intelligence. *Nature News*.

DeJong, G., & Mooney, R. (1986). *Explanation-based learning: An alternative view. Machine learning.* Boston: Kluwer Academic Publishers.

Dellon, B., & Matsuoka, Y. (2007). Prosthetics, exoskeletons, and rehabilitation [grand challenges of robotics]. *IEEE*.

Digg. (2017, August 10). *How The Rise Of Amazon Has Destroyed Retail Chains, In One Chart*. Retrieved from http://digg.com: http://digg.com/2017/amazon-vs-walmart-size

Drexler, K. E. (1986). *Engines of creation.* Anchor.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine,* , 13(2), 99-110.

Earnest, L. (2012). *Cart*. Retrieved from web.stanford.edu: https://web.stanford.edu/~learnest/cart.htm

Elsevier. (2018). Retrieved from https://www.elsevier.com: https://www.elsevier.com/solutions/scopus

Erdélyi, O. J., & Goldsmith, J. (2018). Regulating Artificial Intelligence Proposal for a Global Solution. *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, New Orleans*.

Farrier, J. (2014, September 8). *Soft, Lightweight Exoskeleton Lets You Run 10% Easier*. Retrieved from https://www.neatorama.com: https://www.neatorama.com/2014/09/08/Soft-Lightweight-Exoskeleton-Lets-You-Run-10-Easier/

Fine, P. G. (2011). Long-Term Consequences of Chronic Pain: Mounting Evidence for Pain as a Neurological Disease and Parallels with Other Chronic Disease States. *Pain Medicine*, 12(7), 996-1004.

Fortinet. (2019). *Threat Landscape Report.* Sunnyvale, California: Fortinet.

FutureNow. (2018, October). Auckland , New Zealand : Microsoft New Zealand .

Galeon, D. (2017a). *Futurism*. Retrieved from https://futurism.com: https://futurism.com/fedor-is-a-gunslinging-robot-thats-just-a-skin-suit-away-from-westworld/

Galeon, D. (2017b, December 1). *Google's Artificial Intelligence Built an AI That Outperforms Any Made by Humans*. Retrieved from https://futurism.com: https://futurism.com/google-artificial-intelligence-built-ai/

Goertzel, B. (2004). Encouraging a positive transcension: Issues in transhumanist ethical philosophy. *Dynamical Psychology*.

Goertzel, B., & Pitt, J. (2012). Nine Ways to Bias Open-Source AGI Toward Friendliness. *Journal of Evolution and Techonology*, Vol. 22 Issue 1. pgs 116-131 .

Gonzalez, C. (2017). *MachineDesign*. Retrieved from http://www.machinedesign.com: http://www.machinedesign.com/motion-control/changing-future-warehouses-amazon-robots

Goodwin, T. (2015, March 4). Retrieved from https://techcrunch.com: https://techcrunch.com/2015/03/03/in-the-age-of-disintermediation-the-battle-is-all-for-the-customer-interface/

Google. (2018a). Retrieved from https://cloud.withgoogle.com: https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/

Google Scholar. (2018b). Retrieved from https://scholar.google.ca: https://scholar.google.ca/citations?hl=en&view_op=top_venues&vq=eng_artificialintelligence

Google Scholar. (2018c). Retrieved from https://scholar.google.ca: https://scholar.google.ca/citations?hl=en&view_op=top_venues&vq=eng_artificialintelligence

Griffin, A. (2017, July 31). Retrieved from https://www.independent.co.uk: https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html

Guizzo, E. (2011). IBM's Watson Jeopardy computer shuts down humans in final game. *IEEE Spectrum*, 17.

Guizzo, E., & Goldstein, H. (2005). The rise of the body bots [robotic exoskeletons]. *IEEE spectrum*, 42(10), 50-56.

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics*, 43(5), 1318-1334.

Hanson, R. (2008). Economics of The Singularity. *IEEE SpEctrum*, 45(6).

Hanson, R. (2009, 10 10). Retrieved from www.overcomingbias.com: http://www.overcomingbias.com/2009/10/prefer-law-to-values.html

Harbisson, N. (2012, June). *I listen to colour*. Retrieved from https://www.ted.com: https://www.ted.com/talks/neil_harbisson_i_listen_to_color#t-33336

Hawking, S. (2010). *The grand design.* Random House Digital, Inc.

Hawking, S., & Mlodinow, L. (2010). *The Grand Design .* Bantam Books.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, 105, 105-120.

High, R. (2012). The era of cognitive systems: An inside look at IBM Watson and how it works. *IBM Corporation, Redbooks*.

Hinton, G. E. (2007). earning multiple layers of representation. *Trends in cognitive sciences*, 11(10), 428-434.

Hope, K. (2018, January 22). Retrieved from https://www.bbc.com: https://www.bbc.com/news/business-42745853

Howard, J. (2014, December). The wonderful and terrifying implications of computers that can learn. https://www.ted.com.

Huxley, A. (1932). *Brave New World .* London: Vintage.

IBM. (2019). Retrieved from https://www.ibm.com: https://www.ibm.com/analytics/machine-learning?S_PKG=-&cm_mmc=Search_Google-_-Hybrid+Cloud_Business+Analytics-_-WW_AS-_-reinforcement+learning_Exact_-&cm_mmca1=000000RE&cm_mmca2=10000668&cm_mmca7=1011036&cm_mmca8=kwd-342837665&cm_mmca9=_k_Cj0KCQiA-c_iBRCh

IDC. (2019, March 11). *Worldwide Spending on Artificial Intelligence Systems Will Grow to Nearly $35.8 Billion in 2019, According to New IDC Spending Guide*. Retrieved from https://www.idc.com: https://www.idc.com/getdoc.jsp?containerId=prUS44911419

Jarrasse, N., Maestrutti, M., Morel, G., & Roby-Brami, A. (2015). Robotic prosthetics: moving beyond technical performance. *IEEE Technology and Society Magazine*, 34(2), 71-79.

Joy, B. (2000, April). Retrieved from www.wired.com: https://www.wired.com/2000/04/joy-2/

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular robots obeying Asimov's three laws of robotics. *Artificial life*, 23(3), 343-350.

Kamoshida, K. (2005, November 30). *Robotic Exoskeleton Gets Safety Green Light* . Retrieved from http://avax.news: http://avax.news/pictures/49437

Kaplan, J. (2015). *Humans need not apply: A guide to wealth and work in the age of artificial intelligence.* Yale University Press.

Knuth, D. E., & Moore, R. W. (1975). *An analysis of alpha-beta pruning. Artificial intelligence* (Vol. 6(4)).

Kohs, G. (Director). (2017). *AlphGo* [Motion Picture].

Kurzweil, R. (2000). *The age of spiritual machines: When computers exceed human intelligence.* Penguin.

Kurzweil, R. (2005). *The Singularity is near.* Penguin Books.

Kurzweil, R. (2012). *How to create a mind.* Penguin books.

Kurzweil, R. (2016, 18 August). *Why Ray Kurzweil Believes We Are Becoming More God-Like [Video]*. Retrieved from https://singularityhub.com: https://singularityhub.com/2016/08/18/why-ray-kurzweil-believes-we-are-becoming-more-god-like-video/#sm.000008qokss5w5e37t33jts6x6k8l

Kyriakopoulos, K. (2008). Regional [The European Robotics Research Network (EURON) started in 1999]. *IEEE Robotics & Automation Magazine* .

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature* , 521(7553), 436.

Lewan, M. (2018). The role of trust in emerging technologies. *The Rise and Development of FinTech*, (pp. 111-129). Routledge.

Liao, Y., Vitak, J., Kumar, P., Zimmer, M., & Kritikos, K. (2019). Understanding the Role of Privacy and Trust in Intelligent Personal Assistant Adoption. *International Conference on Information* , (pp. 102-113). Springer, Cham.

Life, F. O. (2017). Podcast: Law and Ethics of Artificial Intelligence [Recorded by A. Con].

Linstone, H. A. (2012). *Singularity Hypotheses: A Scientific and Philosophical Assessment.* (A. H. Eden, J. H. Moor, J. H. Søraker, & E. S. (Eds.), Eds.) Springer Verlag.

Luger, G. F. (2006). *Artificial Intelligence Structures and Strategies forComplex Problem Solving.* Pearson Education.

MacGill, M. (2017, September 4). Retrieved from https://www.medicalnewstoday.com: https://www.medicalnewstoday.com/articles/275795.php

MacGlashan, J., Littman, M. L., Loftin, R., Peng, B., Roberts, D., & Taylor, M. E. (2014). Training an agent to ground commands with reward and punishment. *In Proceedings of the AAAI Machine Learning for Interactive Systems Workshop*.

Marangunić, N., & Granić, A. (2015). echnology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, 14(1), 81-95.

Markoff, J. (2017). Historian. (S. International, Interviewer)

Mashour, G. A. (2005). Psychosurgery: past, present, and future. *Brain research reviews*, 48(3), 409-419.

Mason, C. (2015). Engineering kindness: Building a machine with compassionate intelligence. *International Journal of Synthetic Emotions*, (IJSE), 6(1), 1-23.

Mausumi, D. (2015, February 3). *Three Levels of Consciousness by Sigmund Freud*. Retrieved from https://mausumidutta.wordpress.com: https://mausumidutta.wordpress.com/2015/02/03/three-levels-of-consciousness-by-sigmund-freud-conscious-preconscious-unconscious/

McCauley, L. (2007). AI armageddon and the three laws of robotics. *Ethics and Information Technology*, 9(2), 153-164.

Mitra, M. (2018). Mechanism of an Arc Fusion Reactor. *International Journal of Multidisciplinary Research And Studies*, 1(03), 379-384.

None, N. (1995). The History of Nuclear Energy. (No. DOE/NE-0088; DOE/NE-The History of Nuclear Energy_0). USDOE Office of Nuclear Energy (NE).

O'Conner, T. (2017). *News Week*. Retrieved from http://www.newsweek.com: http://www.newsweek.com/russia-built-robot-can-shoot-guns-and-travel-space-586544

Omohundro, S. M. (2007). The Nature of Self-Improving Artificial Intelligence. *Singularity Summit.*

Omohundro, S. M. (2008). *The basic AI drives.* Amsterdam: IOS Press.

Omohundro, S. M. (2019). *scientific-contributions*. Retrieved from https://steveomohundro.com: https://steveomohundro.com/scientific-contributions/

Ornstein, R. E. (1972). *The psychology of consciousness.*

Paolo Alto Networks. (2019, June). Retrieved from https://www.paloaltonetworks.com: https://www.paloaltonetworks.com/cyberpedia/cyber-security

Pavaloiu, A., & Kose, U. (2017). Ethical Artificial Intelligence-An Open Question. *arXiv preprint arXiv:1706.03021*.

Post, S. G. (2004). *Encyclopedia of Bioethics, 5 Volume Set. Gale.*

Ptolemy, B. (Director). (2011). *Transcendent Man* [Motion Picture].

Puget, J. (2016, May 19). Retrieved from https://www.ibm.com: https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Is_Machine_Learning?lang=en

Ramos, C., Augusto, J. C., & Shapiro, D. (2008). Ambient Intelligen - The next step for Artificial Intelligence. *IEEE Intelligent Systems,*, 23(2), 15-18.

Ramsay, C., & Uren, J. (Directors). (2016). *Pain leads to empathy and self-preservation: should we make robots 'feel' it?* [Motion Picture].

RedHat. (2019). *what-are-application-programming-interfaces*. Retrieved from https://www.redhat.com: https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces

Riva, P., Wirth, J. H., & Williams, K. D. (2011). he consequences of pain: The social and physical pain overlap on psychological responses. *European Journal of Social Psychology*, 41(6), 681-687.

Ross, P. E. (2016, March 15). *AlphaGo Wins Final Game In Match Against Champion Go Player*. Retrieved from https://spectrum.ieee.org: https://spectrum.ieee.org/tech-talk/computing/networks/alphago-wins-match-against-top-go-player

Rothman, D. J. (2017). *Strangers at the bedside: a history of how law and bioethics transformed medical decision making.* Routledge.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development, *, 3(3), 210-229.

Santens, S. (2017, January 15). *Why we should all have a basic income*. Retrieved from https://www.weforum.org: https://www.weforum.org/agenda/2017/01/why-we-should-all-have-a-basic-income/

Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing Test: 50 Years later. *Minds and Machines*, 10(4), 463-518.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Elsevier, Neural Networks*, 85-117.

Shademan, A., Decker, R. S., Opfermann, J. D., Leonard, S., Krieger, A., & C., K. P. (2016). Supervised autonomous robotic soft tissue surgery. *Science translational medicine, *, 8(337), 337ra64-337ra64.

Shane, J. (2018, April 13). *When algorithms surprise us*. Retrieved from http://aiweirdness.com/: http://aiweirdness.com/post/172894792687/when-algorithms-surprise-us

Shanu, S., Talwar, S., Hermon, B., Goswami, O., & Ahuja, N. J. (2016). Semiotics expert system: An integrative approach towards maintenance of community peace. *International Journal of Peace and Development Studies*, 7(6), 50-61.

Smith, C., McGuire, B., Huang, T., & Yang, G. (2006). Retrieved from University of Washington: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf

Springer, P. J. (2013). *Military robots and drones: a reference handbook.* ABC-CLIO.

Stanford encyclopedia of philosophy. (2014, Jan 14). Retrieved from https://plato.stanford.edu: https://plato.stanford.edu/entries/consciousness/

Stevens, R. A. (2017). *The Public-Private Health Care State: Essays on the History of American Health Care Policy.* Routledge.

Stone, W. L. (2005). The history of Robotics. In T. R. Kurfess, *Robotics and autmation handbook* . CRC Press.

Stratoenergetics. (2018). Retrieved from http://stratoenergetics.com/: http://stratoenergetics.com/

Sulleyman, A. (2017, December 5). *GOOGLE AI CREATES ITS OWN 'CHILD' AI THAT'S MORE ADVANCED THAN SYSTEMS BUILT BY HUMANS*. Retrieved from https://www.independent.co.uk: https://www.independent.co.uk/life-style/gadgets-and-tech/news/google-child-ai-bot-nasnet-automl-machine-learning-artificial-intelligence-a8093201.html

Tegmark, M. (2014, June 30). *Consciousness is a mathematical pattern: Max Tegmark at TEDxCambridge 2014*. Retrieved from https://www.youtube.com: https://www.youtube.com/watch?v=GzCvlFRISIM

The Cyborg Foundation. (2019). Retrieved from https://www.cyborgfoundation.com/

The University of Auckland. (2019). Retrieved from https://www.auckland.ac.nz: https://www.auckland.ac.nz/en/about-us/about-the-university/equity-at-the-university/safe-inclusive-equitable-university/unconscious-bias.html

Thrun, S. (1995). Learning to play the game of chess. *In Advances in neural information processing systems*, (pp. 1069-1076).

Trecker, D. (2019, 01 27). *Artificial Intelligence Boon or Bane?* Retrieved from https://lifeinnaples.net: https://lifeinnaples.net/magazinewp/2019/01/27/artificial-intelligence-boon-or-bane/

Veruggio, G. (2007). *EURON Robotics Roadmap.* European Robotics Research Network.

Vinge, V. (1993). The coming technological singularity. *Whole Earth Review*, 81, 88-95.

Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45.

Westmas, R. (2017, July 28). Retrieved from https://curiosity.com: https://curiosity.com/topics/why-the-three-laws-of-robotics-wouldnt-work-and-what-would-instead-curiosity/

Wikipedia. (2018a). Retrieved from https://en.wikipedia.org/wiki/Total_Information_Awareness

Wikipedia. (2018b, August). Retrieved from https://en.wikipedia.org/: https://en.wikipedia.org/wiki/Atomic_bombings_of_Hiroshima_and_Nagasaki

Wikipedia. (2019a). Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

Wikipedia. (2019b). *Consciousness*. Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Consciousness#cite_note-Farthing1992Psychology-3

Wikipedia. (2019c). *Event_horizon*. Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Event_horizon

Wikipedia. (2019d). *List of medical ethics cases*. Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/List_of_medical_ethics_cases

Wikipedia. (2019e). *Mass–energy equivalence*. Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Mass%E2%80%93energy_equivalence

Wikipedia. (2019f). *Nuclear_power*. Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Nuclear_power

Wikipedia. (2019g). *Pathetic fallacy*. Retrieved from https://en.wikipedia.org: https://en.wikipedia.org/wiki/Pathetic_fallacy

World Economic Forum. (2017, June 27). Retrieved from https://www.weforum.org: https://www.weforum.org/agenda/2017/06/the-global-economy-will-be-14-bigger-in-2030-because-of-ai/

World Nuclear Association. (2018, August). Retrieved from http://www.world-nuclear.org: http://www.world-nuclear.org/information-library/current-and-future-generation/outline-history-of-nuclear-energy.aspx

Yampolskiy, R. V. (2011). AI-complete CAPTCHAs as zero knowledge proofs of access to an artificially intelligent system. *ISRN Artificial Intelligence, 2012*.

Yampolskiy, R. V. (2013a). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. *In Philosophy and theory of artificial intelligence* (pp. (pp. 389-396)). Berlin, Heidelberg: Springer.

Yampolskiy, R. V. (2013b). What to do with the Singularity Paradox? *In Philosophy and Theory of Artificial Intelligence*, pp. 397-413.

Yampolskiy, R., & Fox, J. (2013). Safety engineering for artificial general intelligence. *Topoi*, 32(2), 217-226.

YCharts. (2018, March). Retrieved from https://ycharts.com: https://ycharts.com/companies/

Yeap, A. (2018). On AI, Cognition and Computation. Centre for Artificial Intelligence Research, Auckland University of Technology.

Young, L., & Daniel, K. (2003). Affectual trust in the workplace. *International Journal of Human Resource Management*, 14(1), 139-155.

Yudkowsky, E. (2004). Coherent Extrapolated Volition. *Singularity Institute for Artificial Intelligence*.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Global catastrophic risks*, 1(303), 184.

Yudkowsky, E. (2012). Friendly artificial intelligence. In Singularity Hypotheses. *Springer*, pp. 181-195.

Yudkowsky, E. (2012, January 15). The AI-box experiment.

Zhang, S., Yuan, S., Huang, L., Zheng, X., Wu, Z., Xu, K., & Pan, G. (2019). Human mind control of rat cyborg's continuous locomotion with wireless brain-to-brain interface. *Scientific reports* , 9(1), 1321.