WILL AI END HUMANITY?

Contents

Abstract

This dissertation explores the concept of Artificial Intelligence (AI), a rapidly growing field in computer science, answering questions such as what it is and outlining the types of AI. The potential benefits are outlined alongside the risks that are posed by AI including the threat of AI taking over control of the world resulting in a scenario seen in movies like The Terminator. Hawking stated that assuming science fiction will never turn into reality would be "a mistake"

The research conducted on the topic showed that there are numerous benefits that come with the advancements of AI such as improved security and lower death rates. With all the benefits of AI, each of them comes with a risk, one being the risk of people losing their jobs as they get replaced by robots. With all the findings, it can be argued that the consequences can be more detrimental if AI develops requiring the need to implement the solutions discussed to negate the risk of AI negatively affecting humans.

An in-depth analysis of the information concluded that AI will not be able to end humanity despite the major threats if it is developed with safety in mind and is aligned with human values.

Introduction

Artificial Intelligence (AI) is a topic that has been highly talked about but is still a mystery to many people. The subject has been of much debate and speculation in recent years due to its potential. AI is capable of changing how we live in the future and although the benefits of the benevolent use of AI are plenty, AI can be devastating to society if used with malicious intent. However, AI does not need manual attention once set up and can act on its own accord. AI's ability to behave as an individual raises many ethical questions and the big question of "Will AI end humanity?".

In this dissertation, the effects of AI on society will be outlined, and how it could be the future of technology and the key to solving some of the world's biggest problems but also could have disastrous consequences for humanity[1]. The future of technology with AI is bright and can help society in ways we never imagined, however, if not developed cautiously, AI could be the catalyst for the downfall of humanity and lead to the rise of an autonomous world. The development of AI could lead to a battle between Humans and AI which would deter the growth of society, the exact opposite of what AI was built for.

A survey conducted by Oxford University and Yale University in 2018 found that 54% of respondents believed that AI is a threat to humanity in the next few decades[2]. This key statistic shows that students across the world understand and believe in the potential threats posed by AI. The awareness and concern on the topic is a good starting point for this topic to be studied in depth.

The decision to research the topic of Artificial Intelligence stemmed from my interest in technology and concern about the catastrophic events that may result due to the rise of AI. I plan to pursue a career in the field of computer science, further studying the subject in university and eventually aspire to become a computer engineer. This research will help improve my knowledge on AI and potentially enable me to solve any major threat presented by AI in the future. This topic must be researched further to prevent any mishaps from occurring in the near future. I feel that AI should be an aid to us humans and help future generations to possibly be omniscient.

There is a big debate sparked by the use of AI due to its potential, both positive and negative. There are potential benefits such as extending the knowledge and capability of humanity colossally. However, there are a lot of ways AI can be misused and maliciously taken advantage of, potentially creating a divide in society, whether that be economic, social, or cultural. A few ways AI would negatively impact mankind are:

- Bias and Discrimination
- Job Displacement
- Privacy breaches of people
- Autonomous Weapons used in military and the dangers behind it

---

[1] TED-Ed (2022) *The 4 greatest threats to the survival of humanity*, *TED*. TED-Ed. Available at: https://www.ted.com/talks/ted_ed_the_4_greatest_threats_to_the_survival_of_humanity (Accessed: March 27, 2023).

[2] Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O., 2018. When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, pp.729-754.

- Transparency and Accountability of AI and its actions which raise a huge ethical issue
The topics mentioned will be discussed further later in the dissertation.

It is important to research this topic due to the magnitude and importance of the consequences tied to the misuse of AI. We should know how to prevent such a catastrophe and ensure that AI is a tool that helps improve daily lives. Knowing what the consequences of AI are is equally as important as knowing how to stop the unfortunate events from occurring. This allows us to be prepared for any misfortunes that come with the prompt development of AI.

How AI develops and how society develops are two areas that are heavily interdependent as AI can contribute and help the world become a technological utopia. The fact that AI is developed by humans and can possibly be deterred from development at any instant helps reduce the concern about the potential eradication of humanity. AI can only be developed and progress further with the help of humans. Eradication is not a big concern until AI learns how to progress and develop on its own after which the possibilities of the actions of AI are unknown

AI ending humanity is a possibility that hopefully never becomes a reality. The possibility is purely due to the mystery of how AI will be programmed. If society regulates the development of robots, it could negate the possibility of misfortune becoming a reality. There is, however, the concept of AI takeover where humans have no control over intelligence and possibly be enslaved by robots.

A few key definitions to point out are:

Machine Learning (ML) - Computer Systems that learn without being programmed to learn

Artificial Neural Networks - Networks of interconnected nodes based on the interconnections between neurons in the human brain. The system is able to think like a human using these neural networks, and its performance improves with more data

Chatbot - A computer program set up to simulate conversational interaction between humans and a website.[3]

This dissertation will explore AI, its benefits and risks, and the important role of humans in the development of Intelligent technology. Potential methods to eliminate or reduce the risks posed will also be outlined. The pros and cons will be analysed to get to a conclusion and to answer the question: Will AI end humanity?

[3] Watson, D. and Williams, H. (2023) “Artificial Intelligence (AI),” in *Cambridge International AS & A Level: Computer Science*. London: Hodder Education, pp. 434–435.

Literature/Research Review

My research consisted of a range of sources such as TED Talks and blogs written by Computer Scientists, Books, News Articles, Journals, Surveys, and websites of organisations focusing on AI. Providing the primary source of the information provided was the main focus while researching my topic, however, Information such as definitions and factual information provided by some sources did not require the need to find the primary source. For such sources providing information that helps understand the topic better, I chose sources that explained the information in a way that was easy to understand. An example of this was the blog "General AI vs. Narrow AI: What's the difference?" which was chosen due to the conciseness of the information provided as many other blogs and websites did not succeed in explaining the concept in a manner that would be easy to digest. Most of the research done was to understand the topic of AI, gain knowledge about it, and find the counterarguments to the original question posed. The TED Talk helped me understand that AI is one of the greatest threats to humanity according to scientists alongside some other threats such as Nuclear Weapons and Climate Change. TED Talks are a credible source to obtain information as it is a platform where qualified scientists give speeches on an issue to raise awareness. The understanding of the topic was provided by a blog on TechTarget's website "What is artificial intelligence (AI)?". After reading the blog I researched more about the company and found that it was a company focused on Software Development so it is valid to assume any definitions provided on the website were true. The possibility of bias in the information provided by these sources is removed as information and definitions about AI are factual.

Information such as the history of AI was found with the help of Wikipedia as it had references to sources that explain the history of AI. A key piece of information provided was that AI was founded as a discipline in 1956. I found the original source of the information given on Wikipedia and found out it was from the book "Artificial Intelligence, Business and Civilization" by Andreas Kaplan. Wikipedia helped in providing information gathered from multiple different sources which I then looked at individually. In doing this, I succeeded in getting the primary source of all information I gathered from Wikipedia and can ensure the credibility of the information in the dissertation. Definitions are information that does not get outdated so the credibility of the sources used stays intact despite the age of the source. One such example is the book "Introduction To Expert Systems" published in 1998 which provided the definition of the term "Expert Systems". While Wikipedia is an unreliable source as it can be edited by anyone on the internet, I could not find a better definition of the term “AI Takeover” other than the one given on Wikipedia as it was a concept and did not have a specific definition. This was where I had to read the information and make an educated decision about whether I should add the definition into the dissertation which I did as it was a valid explanation of the concept.

Another source used during the research was the A-level Computer Science textbook to provide definitions of words used in the dissertation. These terms such as Machine Learning and Neural Networks are not known by a majority of people. The textbook is used in schools

to learn about Computer Science and is known to be reliable and trustworthy. The textbook, in an effort to teach students about concepts relating to AI, gave concise and understandable definitions of the less-known terms which was optimal for the use of those definitions in the dissertation. Other books in the research included "Life: 3.0" by Max Tegmark outlining the need for AI to be 100% failsafe to be able to use AI and feel safe while using it. Tegmark, a professor at the Massachusetts Institute of Technology (MIT), is also the president of the Future of Life Institute making him a compelling and credible source to include in the research.

Statements from renowned scientist Stephen Hawking were found while researching the concerns about the takeover of AI. From my research, it seemed that Hawking was the only major scientist who expressed his fear of AI becoming uncontrollable. He theorised the takeover back in 2014 when he raised his concerns about dismissing the notion of highly intelligent machines as mere science fiction and stating that it will potentially be "our worst mistake in history". These statements were retrieved from news outlets such as the BBC and the Independent. The interviews conducted were of much help and did not need much thought to be used given that they are trusted sources and are relied on by many people.

Many surveys and statistics were found relating to the topic of AI. Most surveys conducted were about the likelihood of AI being a major threat to humanity and it possibly surpassing human intelligence. The results across the surveys had similar results in that there was a general consensus among respondents that AI is likely to surpass human intelligence and perhaps even take over completely. Surveys included in the research were conducted by institutes such as Yale and Oxford, and Future of Life. The universities of Oxford and Yale are prestigious institutions known for their research ranked fifth and eleventh best in the world respectively. The survey was conducted by the Future of Humanity Institute at Oxford and the Department of Political Science at Yale University and results showed that 54% of respondents believed AI will cause major issues in the next few decades. The other survey was conducted by the Future of Life Institute in 2015. The institute focuses on risks that can possibly threaten the possibility of life. Their major focus is on AI with other focuses being Nuclear Weapons, Climate Change, and Biotechnology. It is well respected by the scientific community and can be classified as a reliable source of information. The only weakness of surveys is that we have to assume that the respondents are being honest while answering the survey.

Articles also played a significant role in obtaining information provided in the dissertation. Some articles used are "Exploring the impacts of artificial intelligence on freedom of religion or belief online" by Cameran Ashraf and "Ethical implications of bias in machine learning" by Adrienne Yapo and Joseph Wiess. The article by Ashraf was published in The International Journal of Human Rights making the source trustworthy. As for Ashraf himself, further research yielded that he has a Ph.D. in Internet censorship and cyberwar and works for human rights at the nonprofit organisation, Wikimedia Foundation. This establishes the credibility of the author of the article. The authors of "Ethical implications of bias in machine learning" were not as credible as the article by Ashraf as I could not find much on the author

Adrienne Yapo. Joseph Wiess, however, is credible as he has a Ph.D. and is a professor of Technology Management at Bentley University.

Other sources of information included a systematic review and an experiment conducted by scientists. The systematic review was done by Avishek Choudhury and Onur Asan where they found that "AI-powered decision support tools can help improve error detection and drug management". The credibility of the reviewers was determined by their other works as they had worked together on other studies such as "Clinicians' Perceptions of an Artificial Intelligence–Based Blood Utilization Calculator". This was an alternative to establishing credibility as I could not find any other information on them. The experiment that was conducted was published in the article "Ethics and Information Technology". The conductors Barbro Fröding and Martin Peterson tested "friendly AIs" which align with human virtues instead of values in an effort to make them less harmful. This source was trustworthy as there was a thorough structure outlined in the experiment and valid conclusions were drawn based on their findings. Both Fröding and Peterson have Ph.Ds in philosophy and work at Institutes focusing on technology, building credibility.

In all these sources, the majority of them lack bias as there is no opinionated information to obtain in the sources, except for the opinion of whether they believe AI can have a bad effect on humanity or not. The main aim of this dissertation is to outline the good and the bad that comes with AI and conclude whether if AI will end humanity. A large part of the information about the positive and negative aspects of AI was mainly obtained from the websites organisations such as IEEE, Partnership on AI, and IBM. These are non-profit organisations that aim for the prosperity of society. These sources can be deemed trustworthy and non-biased due to their nobility.

Despite the lack of bias, there is major speculation in the sources used about the future of AI. This is due to the uncertainty in the development of AI as no one can certainly identify when AI will get too powerful. Many sources speculate that we will be able to control AI no matter how strong it gets as it is created by humans. Others say that once we enter the era of Superintelligent AI, we cannot possibly imagine the horrific consequences. This argument has raised a huge question in the development of AI. This speculation was the only aspect that may make a source unreliable but I ensured that all the sources contained information that was valid.

Will AI end Humanity?

An overview of AI

Artificial intelligence (AI) refers to the simulation of human intelligence processes by machines, especially computer systems which allows computers to have the ability to perform tasks that normally need human intervention[4]. Some of these tasks may include visual perception, speech recognition, decision-making, and language translations. AI has the ability to do such tasks with the help of algorithms and machine learning (ML). These capabilities of the computer were a big leap in technology since the invention of the computer.

AI was an idea that had been born as recently as 1956[5]. However, it wasn't until 1987 that AI had a "Boom" in the form of expert systems which are computer systems emulating the decision-making ability of a human expert[6]. That is when the true capabilities of AI were discovered. AI was a field that was vastly researched in the early 2000s and it yielded the concept of machine learning which allowed computers to take data, learn and analyze it, and respond in an appropriate manner. This immensely improved the performance of computers without the need for any extra programming.

Today, AI is a field that has the potential to transform many industries and improve many aspects of daily life, from healthcare to transportation to entertainment. It combines multiple disciplines, including computer science, mathematics, statistics, psychology, and more. It has the capacity to rapidly evolve and, in turn, be the solution to all our problems. The growth of AI can be attributed to various factors, such as the availability of vast amounts of data, faster and more powerful computer hardware, and the discovery of innovative machine learning algorithms that enable AI systems to learn and improve over time.

One of the most significant recent advances in AI has been the development of deep learning. This involves handling huge amounts of data using artificial neural networks which enables the networks to recognize patterns and make predictions with high accuracy[7]. Deep learning has enabled major improvements in image recognition, natural language processing, and other AI applications.

In the industry, there are two main types of AI systems: Narrow AI (ANI) and General AI (AGI). Narrow AI is more common in today's world and is designed to solve a specific problem, such as recognizing faces or detecting diseases by analyzing patient data. In contrast, General AI aims to replicate human intelligence and perform any intellectual task

---

4 Burns, E., Laskowski, N. and Tucci, L. (2023) *What is Artificial Intelligence (AI)?: Definition from TechTarget*, *Enterprise AI*. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence (Accessed: March 26, 2023).

5 Kaplan, A. (2022) *Artificial Intelligence, business and civilization: Our fate made in machines*. Abingdon, Oxon: Routledge, Taylor & Francis Group.

6 Jackson, P. (1998) in *Introduction to expert systems*. 3rd edn. Wokingham: Addison-Wesley, p. 2.

7 Watson, D. and Williams, H. (2023) "Artificial Intelligence (AI)," in *Cambridge International AS & A Level: Computer Science*. London: Hodder Education, pp. 434–435.

that a human can do. Examples of General AI are chatbots such as ChatGPT and autonomous vehicles. While narrow AI is the only type of AI we have excelled at the moment, many researchers believe that general AI is the ultimate goal of the field eventually leading to the creation of Artificial Superintelligence (ASI) which would outperform humans and could result in disaster for humanity[8].

[8] *General AI vs Narrow AI* (2022) *RSS*. Levity. Available at: https://levity.ai/blog/general-ai-vs-narrow-ai#:~:text=What's%20the%20difference%20between%20Narrow,any%20problem%20that%20requires%20AI (Accessed: March 10, 2023).

How does AI make the world a better place?

The idea of AI ending humanity may seem too extreme and even silly to some. The development of AI does not have to result in such extreme outcomes, instead, it can become a tool to improve our lives. As discussed previously, the possibilities are endless when it comes to AI benefitting the world. Organisations such as Partnership on AI ensure the safe use of AI. They aim for a safe advancement in the field of Artificial Intelligence to warrant a future where Artificial Intelligence empowers humanity by contributing to a more just, equitable, and prosperous world[9]. AI has the potential to revolutionize many aspects of our lives, from healthcare and transportation to finance and education.

Advances in AI have the potential to improve outcomes, enhance quality, and reduce costs in such safety-critical areas as healthcare and transportation. These are areas where there is no room for error. Robotic systems aided by AI can be implemented in autonomous vehicles or performing a surgery. AI is a fitting choice for such tasks as it eliminates the factor of human error and significantly increases safety due to the high precision and accuracy of AI. Effective and careful applications of pattern recognition, automated decision-making, and robotic systems show promise for enhancing the quality of life and preventing thousands of needless deaths[10]. For example, AI-powered diagnostic tools can help doctors quickly and accurately diagnose diseases, reducing the risk of misdiagnosis and improving patient outcomes. AI has the potential to provide customized real-time recommendations to patients around the clock[11] and identify high-risk patients based on a patient's unique medical history and health status. In a recent systemic review, 53 peer-reviewed studies examining the impact of AI on patient safety found that AI-powered decision support tools can help improve error detection and drug management[12]. A factor to consider while replacing human decision-making with that of AI is the reliability and accountability while performing tasks. It will be a demanding task to guarantee 100% reliability of the decision-making of AI. Another factor is trust as people may not trust AI enough to take over tasks that, if gone wrong, could lead to fatalities. It is important that we ensure AI tools are aligned with the ethics and preferences of people who are influenced by their actions.

The ability of AI to automate routine tasks frees up time for humans allowing them to dedicate that time to focus on duties that require more creative workarounds, therefore, allowing for multitasking. Such traits are beyond the capabilities of AI but it is only a matter of time until it is programmed or AI self-improves to be able to have emotional intelligence and intuition to perform more complex tasks. Freeing up time leads to more productivity and increased efficiency. Manufacturing is a field where AI is most likely to take over as it can

---

9 *Mission - Partnership on AI* (2022) *Partnership on AI*. Available at: https://partnershiponai.org/about/#mission (Accessed: March 12, 2023).

10 *Pillars - Partnership on AI* (2022) *Partnership on AI*. Available at: https://partnershiponai.org/about/#pillars (Accessed: March 12, 2023).

11 *Artificial Intelligence in medicine* (2022) *IBM*. Available at: https://www.ibm.com/ae-en/topics/artificial-intelligence-medicine (Accessed: March 12, 2023).

12 Choudhury, A. and Asan, O. (2020) "Role of artificial intelligence in Patient Safety Outcomes: Systematic Literature Review," *JMIR Medical Informatics*, 8(7). Available at: https://doi.org/10.2196/18599.

optimize production processes, reduce waste, and improve product quality. Tasks such as data entry, document processing, and customer service will not require human intervention soon as they are low-level tasks that can be performed by AI with ease

Safety and Security can be improved immensely using AI as surveillance systems powered by AI can use facial detection to help detect criminals and assist in preventing crimes. Self-driving cars can reduce accidents caused by human error. Students in schools can be provided with personalised learning using AI using data on the student. It can also take feedback and identify areas where a student may need extra help in. Education can be made accessible and more constructive for students with the help of AI

Analysing large data sets is a trademark of AI and is significant in helping humans identify patterns that they normally would not be able to detect. This can help accelerate scientific research and allow quick development of new models and simulations for scientists to test hypotheses which can lead to the discovery of new theories.

Other Areas where AI can be used are:

Finance - To detect and prevent fraud

Energy Management - To optimize power generation and distribution

Risks associated with the development of AI

Despite the numerous benefits of AI, there are also significant concerns about its impact on society. The most pressing concern is the possibility that AI could replace human workers in low-level jobs, leading to mass unemployment and economic disruption. Jobs such as receptionists, customer service, accountants, taxi drivers, and many more are at risk of completely disappearing as AI will be able to perform those tasks better than humans due to its efficiency and precision. Sectors relying heavily on manual labor will face significant unemployment. Although AI will create many new job opportunities, it may not make up for the number of jobs lost. Jobs created because of AI will require competency in handling computers and knowledge in the field of computer science. These requirements cannot be filled by a person who is in a low-level sector eventually leaving them unemployed regardless of the jobs created by AI.

As AI is trained on specific data sets, they are only as unbiased as the data provided to them. This can lead to the possibility that there are hidden assumptions and biases in data, and therefore in the systems built from that data. This can lead to actions and recommendations that replicate those biases and have serious blind spots[13]. This can be disastrous if AI is used in sectors such as the criminal justice system which could lead to wrongful convictions. For example, an AI system trained on data that is biased against women will produce less accurate outcomes for women than for men. This could also create a sort of capitalist society where people with access to AI will be superior to people who do not and they can wrongfully manipulate the AI system by feeding it data biased towards them making them even more influential

The need for AI to have access to a large amount of personal data in order to make accurate predictions raises concerns about the privacy and security of that data. According to C Ashraf, in an article published in The International Journal of Human Rights[14], “AI-enabled automation and analysis can result in a wide range of privacy and security threats, such as data breaches, malicious attacks, and identity theft.” In addition, AI systems are increasingly being used to track people's activities and monitor their behavior without consent, which can lead to an unjustified invasion of privacy. AI systems may become targets for cyber attacks to steal the data stored in the system and use it to their advantage. Governments or corporations could misuse facial recognition data collected by AI and potentially sell it to third parties

[13] *Pillars - Partnership on AI* (2022) *Partnership on AI*. Available at: https://partnershiponai.org/about/#pillars (Accessed: March 12, 2023).

[14] Ashraf, C. (2021) “Exploring the impacts of artificial intelligence on freedom of religion or belief online,” *The International Journal of Human Rights*, 26(5), pp. 757–791. Available at: https://doi.org/10.1080/13642987.2021.1968376.

With the uses of AI becoming more sophisticated, it is increasingly being used to make decisions that were once made by humans. Humans are held accountable for their decisions but people are debating if the same applies to AI systems' decisions. There have also been many questions relating to the ethics of the unaccountability of AI as they can be heavily biased. In an article[15], Yapo states, "awareness of impending ethical risks and issues are crucial in the design of AI to ensure that the most vulnerable in our society are protected from harm". The ethical implications of bias in machine learning are significant and can affect the people who are most susceptible to that harm.

Military Weapons powered by AI could cause a great deal of controversy as there are extreme consequences in the event that AI fires such weapons. AI could be vulnerable to other AI-powered hackers and could take control of the weapon. Governments can also misuse these weapons and potentially start a world war using autonomous weapons. If AI gets advanced enough, it will have the capability to use these weapons on its own accord and humans will have no control over major nuclear weapons that threaten the eradication of the human race from the face of the earth

The potential of AI to become uncontrollable and autonomous is one of the main concerns to scientists. AI, by design, is meant to improve and learn on its own without the need for human intervention. As we step closer towards Artificial Superintelligence, AI will pass human intelligence and become smarter than us. It gaining the ability to make decisions on its own could be disastrous for humanity. In such a scenario, AI could potentially pose a threat to human existence if their decisions are not aligned with human interests and leading to AI taking over the world.

[15] Yapo, A. and Weiss, J. (2018) “Ethical implications of bias in machine learning,” *Proceedings of the 51st Hawaii International Conference on System Sciences* [Preprint]. Available at: https://doi.org/10.24251/hicss.2018.668.

The concept of "AI Takeover"

AI takeover or the "singularity" is a hypothetical scenario in which an Artificial intelligence becomes the dominant form of intelligence on Earth, creating a world where humans are incapable of competing with Superintelligent AI and effectively having no control of the planet[16]. AI, in the form of computer programs or robots, could replace the entire human workforce due to its advanced capabilities, ones we can only imagine at the moment. Science fiction movies such as The Terminator, RoboCop, and Transcendence have popularised the idea of AI controlling the world. These movies give viewers a glimpse of how a fictional world controlled by AI would be. In May 2014, Stephen Hawking, while referencing such films stated[17], "it's tempting to dismiss the notion of highly intelligent machines as mere science fiction. But this would be a mistake, and potentially our worst mistake in history." In another interview with the BBC[18], Hawking, while mentioning AI, said that "It would take off on its own, and re-design itself at an ever-increasing rate". These glimpses of fiction can potentially turn into reality if the development of AI gets out of control

Multiple surveys and reports show that scientists believe that the possibility of an extreme outcome such as an AI takeover is difficult to estimate but "not zero". In 2015, a survey[19] conducted by the Future of Life Institute showed that 37% of respondents believed that AI will eventually surpass human intelligence. However, only 23% of respondents believed that outcome is unlikely.

The root of the possibility of an AI takeover is the concern that if advanced enough, AI may be able to exponentially improve and modify itself to a point where humans cannot control the further advancement of AI. Although there is a possibility of AI taking over and helping the human race grow at the same rate, the most probable outcome is AI potentially exploiting us humans for its own benefits with an utter disregard for human values and goals.

The likelihood of such a scenario is pretty low as preventative measures can be taken to ensure that science fiction stays fictional and never becomes something we have to deal with.

---

16 *AI takeover* (2023) *Wikipedia*. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/AI_takeover (Accessed: March 23, 2023).

.

17 *Stephen Hawking: 'are we taking artificial intelligence seriously* (2014) *The Independent*. Independent Digital News and Media. Available at: https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html (Accessed: March 23, 2023).

18 Cellan-Jones, R. (2014) *Stephen Hawking warns artificial intelligence could end mankind*, *BBC News*. BBC. Available at: https://www.bbc.com/news/technology-30290540 (Accessed: March 23, 2023).

19 Future of Life Institute, "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter," 2015.

The role of humans in the development of AI

No matter how rapid the development of AI is, eventually, we decide whether it has the capability of succeeding us and becoming superior. It is our decision to either deter the growth of AI or let it reach a point where AI can take matters into its own hands. Fortunately, that decision does not have to be made anytime soon but safety measures must be taken in order to ensure that it is us humans who are in control of the world in the future. An optimistic scenario would be where we do not have to make that decision and possibly have a future where humans and robots co-exist and thrive together. Regardless of the optimism, this scenario is highly unlikely because even if created solely with humanity's best interests in mind, superintelligent AI could pose an existential risk if it isn't perfectly aligned with human values.

To address the concerns about Artificial Superintelligence, there is a growing need for ethical guidelines and regulations around the development and use of AI. Organizations like the "IEEE" and the "Partnership on AI" are working to develop standards and best practices for AI. Governments worldwide are implementing regulations to ensure the responsible use of AI.

In his book Life 3.0, Tegmark suggests that unless the future AI is 100% failsafe and unhackable, you can't feel safe[20]. This idea of relying on AI is scary and is not plausible until we can guarantee that using AI is completely safe. AI's failure to completely align its goals with those of humans could upset every human effort gone into its development.

During the development of AI, it is essential to consider the aspect of safety and ensure that AI systems do not cause any harm to humans, especially since autonomous systems become more advanced as more data is fed into the system to learn from. Developers must also consider the privacy of the data being input and ensure it is only being used as intended, especially if it is personal data. The data also must not be biased which can happen if the system is developed by a homogeneous team. This issue was discussed in the article "Ethical implications of bias in machine learning." where Yapo and Weiss emphasised that there must be increased diversity in the teams that develop and use these systems such as social scientists, ethicists, philosophers, faith leaders, economists, lawyers, and policymakers[21].

[20] Tegmark, M. (2017) *Life 3.0: Being human in the age of Artificial Intelligence*. New York , New York: Alfred A. Knopf.

[21] Yapo, A. and Weiss, J., 2018. Ethical implications of bias in machine learning.

Eliminating the risks

The risks posed by the evolution of AI could lead to horrible outcomes. Therefore, it is vital to be proactive and come up with potential solutions and eliminate these risks way before they have an impact on us.

A major risk discussed was the lack of transparency and accountability of AI systems. Partnership on AI targets to remove ambiguity by helping develop AI systems with complete transparency to ensure everyone understands how AI systems work[22]. Researchers must be able to present the datasets that the AI system has been trained on and explain how the system makes decisions. This can ensure the safe and responsible use of AI systems. Such organisations ensure transparent and responsible development preventing any unintended consequences. As for accountability, developers that create an AI system must be held accountable for the system regardless of it becoming autonomous

Regulations can be put into place governing the use and development of AI so that the benefits of AI are shared equally in a safe and ethical manner. Further investments for research into the safe development of AI is also essential. This has already begun in recent years due to the concerns relating to AI. Safety research can ensure proven methods of ethical AI development that align with our values and prioritises our interests

Some researchers suggest that a "friendly AI" could be the solution to some of the risks posed by normal General AI. A friendly AI would mimic a sufficient number of aspects of proper friendship[23]. This approach is different from the “value alignment” of AI and instead focuses on “virtue alignment”, aligning AI with human virtues instead of values. This would have a positive effect on humanity or at least contribute to fostering the improvement of the human species.

[22] *About - Partnership on AI* (2022) *Partnership on AI*. Available at: https://partnershiponai.org/about/ (Accessed: March 25, 2023)

[23] Fröding, B. and Peterson, M. (2020) “Friendly AI,” *Ethics and Information Technology*, 23(3), pp. 207–214. Available at: https://doi.org/10.1007/s10676-020-09556-w.

Conclusion

To answer the original question "Will AI end Humanity?", it is crucial to consider and understand both, the benefits, and risks that are associated with the use and prompt development of AI. As mentioned, AI can be the tool that revolutionises the world but, understandably, it comes with its own risks and ways where those revolutionary benefits become tragic threats. One such example is AI using facial recognition to apprehend criminals, however, facial recognition can also be used to track people and be misused by governments leading to breaches of privacy. There is also the risk of the large amount of personal data stored being stolen by AI-powered hackers.

It is safe to say that each benefit of AI comes with its own drawback with the major drawback being the invasion of privacy and the risk of it being accessed by unwanted parties. These drawbacks do not contribute majorly towards the debate of whether AI poses an existential threat to humanity but are factors to consider while using AI systems as they become more prominent as they are developed. The main concern posing the threat is Superintelligent AI which surpasses human intelligence deeming us the inferior form of intelligence on Earth.

Considering the progress in the development of AI, the threat of eradication is not feasible anytime soon. This does not mean the threat is not present. We as a community must take preventative measures, some of which are mentioned previously, to ensure the threat is never major enough to have to make an active decision on whether to remove or disable AI systems or not. For the most part, advancements in AI will only cause insignificant inconveniences to humans until it becomes advanced enough to take matters into its own hands. Inconveniences such as Job Displacement and Privacy Breaches will not matter once Superintelligent AI takes control, after which, we have no idea how drastic the consequences can be.

Even if AI never gets to the point where it can potentially become Superintelligent, it still poses major threats as autonomous weapons, especially nuclear weapons, if misused, can be the cause of the disappearance of the human race. This is one of the biggest concerns of AI while it is still being developed. We should not consider creating autonomous weapons until we can warrant the safety and security of AI Systems. This will be a hard task for scientists as hackers are constantly evolving and finding out methods to breach a system and it is possible that AI will not be 100% unhackable until it is Superintelligent.

After careful consideration and thorough analysis of both viewpoints discussed in the dissertation. It is safe to conclude AI will not be the end of humanity as more people are becoming aware of AI and its threats which has led to policies being made by organisations to ensure the safe development of AI. Despite this conclusion, we must certainly keep in mind the major threats posed by AI in case it ever gets close to being uncontrollable. If ideas discussed previously such as "friendly AI" and the need for transparency in the development of AI are implemented across all developers, we can guarantee that AI will never be able to threaten our existence.

Evaluation

Overall, the project was successful as it conclusively answered the initial question posed of whether AI will end humanity or not. It was a time-intensive task to complete the dissertation which I realised once I started the research and typing out the discussion. Nonetheless, it was an enjoyable project to carry out as I chose a topic I am passionate about and want to pursue as a career in the future. The dissertation yielded in a positive growth of my knowledge about the topic of AI. It helped in understanding the uses and effects of AI. Before my research, I was curious about how scientists were dealing with the threat of AI and wondered if they thought it could be the reason for humanity ending. The research definitely answered most of my questions and did provide the answer to the main question. Finding out the consequences of AI sparked curiosity about when the predicted events would occur if they do at all. I also learnt about neural networks, machine learning, and the types of AI amongst other things during the dissertation which I previously did not have much knowledge of

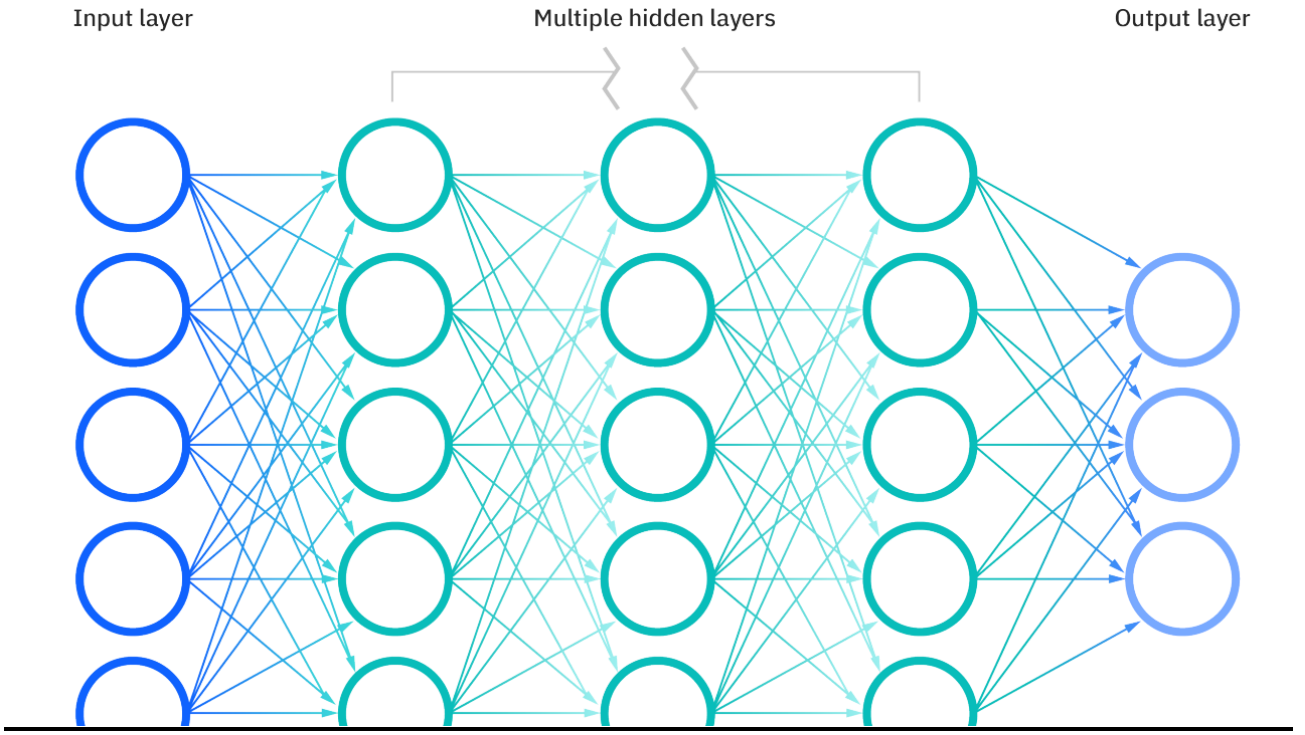
The field of Computer Science, specifically, AI, is a rapidly growing field and is being talked about a lot in today's world. This was optimal is it meant that there was a lot of discussion and research about the topic of the project. The sources available were plenty with the main issue being to find and verify valid sources. I think I did a good job including sources that were credible and did not present any bias. I did have some issues in retrieving sources that were published a while ago such as the book Introduction To Expert Systems published in 1998. The book was not free online which posed a problem as I could not pay for it but I did find it in a nearby library. Other sources were found online without many issues. There were sources that were repeated a few times such as the Partnership on AI website. This was because the organisation focused on the betterment of society with the help of AI which was closely knit to the topic of the dissertation. I found a lot of information on the website which was useful in multiple sections of the dissertation. One source I was skeptical about was Wikipedia which provided the definition of AI takeover. The source seemed useful as it provided me with a good understanding of the concept. I had to use that definition of the concept as I could find no other source that could have been more reliable. During the research, I came across 2 institutes with similar names and purposes which were Future of Life and Future of Humanity Institute. I had to do some extra research to find out the difference as I was confused and found that Future of Humanity is an institute which is run by Oxford University whereas Future of Life is a standalone institute but both work to ensure human life on earth.

As I was concluding my dissertation, I found new information that had just been released about the development of AI that I could have written about but that meant I would go over the word limit and would have to change the structure of the dissertation. The source stated that Elon Musk, the co-founder of OpenAI and many other companies, wanted to pause the development of AI and ChatGPT due to risks to society. I was happy to hear that as it supports the conclusion that AI will not be able to pose major threats if controlled properly.

This meant that my research was a success in that it rightly stated that AI is a threat that needs to be taken care of in the near future. It is fulfilling to see that action is already being taken to control AI. I do wish I could have written more as I realised that there was much more to talk about in AI after completing the EPQ which was ironic as I did not think I would have enough content to complete the dissertation at the start.

During the course of the dissertation, I learnt and developed some essential skills such as time management, structuring, finding key information, fact-checking and efficiently skimming through a piece of text. A skill such as structuring can be useful to me in the future as I aspire to be a programmer where it is essential to have properly structured code for it to be understood by others. I could also see an improvement in my time management by the end of the EPQ as I allocated different tasks and subjects to different days. There are many more subtle skills gained such as managing priorities which I feel like I did not do when I started the EPQ. I pushed the project to the side and did not proactively try and complete it. The lack of knowledge about how difficult the project is led to procrastination and the lack of urgency to complete it. This meant that I had to do the bulk of the work in the span of a month while also trying to study for A-levels. I left a lot of work to be done in the latter half of the allocated time for the EPQ which was not something I am proud of. If given another chance, I would definitely manage my time better from the beginning and rather do small bits of the project every day than doing a lot of work in a few weeks' time. This would have been efficient and resulted in more time for me to refine the dissertation and perfect it.

24



[24] *Autopilot* (2023) *Tesla*. Available at: https://www.tesla.com/autopilot (Accessed: March 12, 2023).

[25]

Bibliography

About - Partnership on AI (2022) Partnership on AI. Available at: https://partnershiponai.org/about/ (Accessed: March 25, 2023)

AI takeover (2023) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/AI_takeover (Accessed: March 23, 2023).

Artificial Intelligence in medicine (2022) IBM. Available at: https://www.ibm.com/ae-en/topics/artificial-intelligence-medicine (Accessed: March 12, 2023).

Ashraf, C. (2021) "Exploring the impacts of artificial intelligence on freedom of religion or belief online," The International Journal of Human Rights, 26(5), pp. 757–791. Available at: https://doi.org/10.1080/13642987.2021.1968376.

Autopilot (2023) Tesla. Available at: https://www.tesla.com/autopilot (Accessed: March 12, 2023).

Burns, E., Laskowski, N. and Tucci, L. (2023) What is Artificial Intelligence (AI)?: Definition from TechTarget, Enterprise AI. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence (Accessed: March 26, 2023).

Cellan-Jones, R. (2014) Stephen Hawking warns artificial intelligence could end mankind, BBC News. BBC. Available at: https://www.bbc.com/news/technology-30290540 (Accessed: March 23, 2023).

Choudhury, A. and Asan, O. (2020) "Role of artificial intelligence in Patient Safety Outcomes: Systematic Literature Review," JMIR Medical Informatics, 8(7). Available at: https://doi.org/10.2196/18599.

Fröding, B. and Peterson, M. (2020) "Friendly AI," Ethics and Information Technology, 23(3), pp. 207–214. Available at: https://doi.org/10.1007/s10676-020-09556-w.

Future of Life Institute, "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter," 2015.

General AI vs Narrow AI (2022) RSS. Levity. Available at: https://levity.ai/blog/general-ai-vs-narrow-ai#:~:text=What's%20the%20difference%20between%20Narrow,any%20problem%20that%20requires%20AI (Accessed: March 10, 2023).

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O., (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, pp.729-754.

[25] *What are neural networks?* (2023) *IBM*. Available at: https://www.ibm.com/topics/neural-networks (Accessed: March 28, 2023).

Jackson, P. (1998) in Introduction to expert systems. 3rd edn. Wokingham: Addison-Wesley, p. 2.

Kaplan, A. (2022) Artificial Intelligence, business and civilization: Our fate made in machines. Abingdon, Oxon: Routledge, Taylor & Francis Group

Mission - Partnership on AI (2022) Partnership on AI. Available at: https://partnershiponai.org/about/#mission (Accessed: March 12, 2023).

Pillars - Partnership on AI (2022) Partnership on AI. Available at: https://partnershiponai.org/about/#pillars (Accessed: March 12, 2023).

Stephen Hawking: are we taking artificial intelligence seriously (2014) The Independent. Independent Digital News and Media. Available at: https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html (Accessed: March 23, 2023)

TED-Ed (2022) The 4 greatest threats to the survival of humanity, TED. TED-Ed. Available at: https://www.ted.com/talks/ted_ed_the_4_greatest_threats_to_the_survival_of_humanity (Accessed: March 27, 2023).

Watson, D. and Williams, H. (2023) "Artificial Intelligence (AI)," in Cambridge International AS & A Level: Computer Science. London: Hodder Education, pp. 434–435.

What are neural networks? (2023) IBM. Available at: https://www.ibm.com/topics/neural-networks (Accessed: March 28, 2023).

Yapo, A. and Weiss, J. (2018) "Ethical implications of bias in machine learning," Proceedings of the 51st Hawaii International Conference on System Sciences [Preprint]. Available at: https://doi.org/10.24251/hicss.2018.668.