Fert Batxillerat

Pablo García Andrés

The Future of Artificial Intelligence

Research Project directed by

Ariadna Vilalta Domingo

Barcelona, December of 2019

Index

0.Abstract and Keywords

Abstract - English

Artificial Intelligence can generally be defined as the study and development of computer systems that are able to demonstrate intelligence and carry out tasks which normally require human intelligence. AI originated as a proper field of study in the middle of the 20th century. During its first decades it saw stages of great optimism and periods of skepticism and a lack of investment, known as “AI winters”. Currently, AI systems are used in a great number of fields and outperform humans in some specific domains. The main goal current AI researchers hope to achieve is the development of superintelligent machines, which are systems whose general intellectual capabilities are greatly superior to those of a human. Such machines could report immense benefits but could also be extremely dangerous due to their superhuman abilities. It is crucial to ensure that they would not act in a way which could be harmful to us. To do so, it is extremely important to be able to predict and control a superintelligent AI’s goals and motivations, as well as giving it the right values and objectives. The possibility of a superintelligent AI has raised many ethical questions and concerns about the nature of such systems. It is naturally very early to able to make accurate predictions, but some researchers claim that there is reason to believe a superintelligent machine systems might become self-aware and develop free will. Developing a superintelligent AI would likely be one of the most important events in all of human history, and while it could report immense benefits, it could have devastating consequences leaving humanity in a state beyond repair, which means we must be extremely careful and wary of the potential it yields.

Abstract – Spanish

La Inteligencia Artificial (IA) es el estudio y el desarrollo de sistemas de computación que son capaces de llevar a cabo tareas que generalmente requieran un nivel de inteligencia propio de un ser humano. La Inteligencia Artificial surgió formalmente como disciplina a mediados del siglo XX. Durante sus primeras décadas, vio épocas marcadas por un gran optimismo y periodos de escepticismo y falta de inversión. Actualmente, se utilizan sistemas de IA en un gran número de campos, y son capaces de superar las habilidades humanas en ciertos ámbitos. El objetivo actual más importante de los investigadores de IA

es conseguir desarrollar máquinas superinteligentes. Esto supone desarrollar un sistema cuyas capacidades intelectuales generales son muy superiores a las de un ser humano. Estas máquinas podrían traernos inmensos beneficios, pero también podrían ser extremadamente peligrosas debido a sus habilidades sobrehumanas. Es de vital importancia que no actúen de forma que pudieran hacernos daño. Para esto, es crucial ser capaz de predecir y controlar los objetivos y las motivaciones de una AI superinteligente, además de otorgarle los objetivos a seguir adecuados. La posibilidad de que puedan existir sistemas superinteligentes ha provocado la aparición de muchas preguntas y preocupaciones éticas sobre la naturaleza de estos sistemas. Es evidente que aún es muy pronto como para poder hacer predicciones certeras, pero hay investigadores que defienden que hay motivos suficientes para creer que una máquina superinteligente pueda llegar a volverse consciente de ella misma y tener cierto grado de libertad. Llegar a desarrollar un sistema de inteligencia artificial superinteligente podría representar uno de los hechos más importantes de toda la historia. Podría traernos beneficios innumerables, pero también podría tener consecuencias devastadoras que dejaran a la humanidad en un estado irreparable, lo cual implica que tenemos que ser extremadamente cuidadosos y precavidos.

Keywords

Superintelligence, Motivation Selection, Free Will, Consciousness and Existential Risk

1.Introduction

Understanding the nature of our universe and of our reality is one of the ultimate questions we, as human beings, have to face and try to answer. When I was trying to decide what to do my Research Project on, I came across Nick Bostrom's (a Philosophy Professor at Oxford University) essay on the Simulation Theory. In his paper, Bostrom argues that the human species will either become extinct before it reaches a "posthuman" stage, or that it will be able to develop Artificial Intelligence so advanced that it will be capable of simulating life and existence. However, if the human species is in fact capable of achieving the latter, this would imply that it very possibly has already happened and we are, in fact, living in a computer simulation.

While the Simulation Theory is not the topic of my Research Project, reading this paper fascinated me. It made me realize to what extent Artificial Intelligence can be developed and the enormous potential it carries, and it made me want to find out more about it. AI is, without a doubt, extremely relevant as of today. We are aware that developing intelligence superior to that of humans could change the world as we know it. I believe superintelligence might be the greatest challenge we have ever had to face, and I therefore believe it is a very interesting topic to do my Research Project on.

There are different aspects of AI I have investigated throughout this research project. Firstly, I have explored the history of Artificial Intelligence. I wanted to discover how and when it first started, how it has evolved and where it currently stands. I have also talked about who exactly is working on developing AI and where the money to do so is coming, as this greatly determines with what intention it is being developed. Secondly, I have looked at different theories and predictions to try to gain a better understanding of how AI will advance in the following years and when (if it does) it might reach a level comparable to human intelligence, and devise potential applications for it. Lastly, and in my opinion, most importantly, I have decided to focus on the ethical aspect of Artificial Intelligence. It is a topic of the utmost importance as it raises a lot of radically challenging and complex questions that are yet to be resolved, such as: Can a machine have consciousness? Can it reason and feel? Will machines be able to

surpass human intelligence? Is AI ethical? How can we know that AI will not turn on us? The long-lasting impact AI might have on the world as we know it lies in the answers to questions of this sort, which is why I have touched on them in my research project.

To carry out my research I have mostly relied on books that are related to AI and are written by experts on the matter, primarily Nick Bostrom. I have also used the Internet to access reports and scientific essays, as well as articles written on recent advances in AI and interviews with prominent figures in the AI world.

Finally, I would like to thank my tutor for this research project, Ariadna Vilalta, for her constant support and for allowing me to see certain aspects of my work in a different light, which has enabled me to add a new perspective to my investigation and make it more complete.

2. Artificial Intelligence

2.1 What is Artificial Intelligence?

In order to understand this Research Project and its topic, it is important to establish a clear definition of what Artificial intelligence is and what it is considered, as well as the definitions for the different "levels" of artificial intelligence. Artificial intelligence is a branch of computer science that can generally be defined as the study and development of computer systems and machines that are able to demonstrate intelligence and carry out tasks which normally require human intelligence. (Cambridge Dictionary, 2019) Human-level artificial intelligence or general artificial intelligence refers to a type of intelligence which can, in all aspects, perform as well as a human being. Superhuman intelligence refers to an intelligence whose capabilities greatly outperform those of a person. (Merriam-Webster, 2019) Let it be noted that artificial intelligence is often referred to as AI, so whenever this abbreviation is used throughout this paper it refers to artificial intelligence.

2.2 Origins and history of Artificial Intelligence

It is through history that we understand who we are and what we have become. To understand the nature of AI and to be able to make logical and coherent predictions about its future, we must direct our attention to its origins.

In the summer of 1956 ten scientists interested in studying neural nets[1], automata theory[2] and intelligence worked together for six-weeks on a workshop in Dartmouth College. This workshop is usually considered the start of the field of artificial intelligence research. It marked the begin of what would be known as the golden years of AI. During the 20 years that this era lasted, the progress made was outstanding. Robots were built that were capable of carrying out tasks that were only thought to be achievable humans, such as solving relatively complex algebra problems and proving theorems in geometry. Computers with the ability

[1] Neural Nets: computing systems modelled on the human brain and nervous system. They are not in themselves algorithms, but rather a framework for many different machine learning algorithms to work together.
[2] Automata Theory: the study of abstract machines and automata, as well as the problems that can be solved using them.

to communicate using natural languages were also built. One of the most famous systems of the time, the ELIZA robot, was able to carry out a regular conversation, making users believe they were talking to an actual human being. In reality, ELIZA was simply giving back canned responses to certain questions she was asked with a few grammatical tweaks and had no idea what she was actually saying.

During this time AI researchers were extremely optimistic. Nobel Prize winner and AI pioneer, Herbert A. Simon, said in 1965: “machines will be capable, within twenty years, of doing any work a man can do.” (Simon, 2017) AI was very much in vogue, so funding came as a natural result, mainly from public institutions. In 1963 the Massachusetts Institute of Technology (MIT) started to receive millions of dollars yearly to fund the study of AI from the Advanced Research Projects Agency (which would be later known as DARPA). Similar grants were also made to the Carnegie Mellon University (CMU) and the Stanford AI Project.

However, as time passed, more and more problems started to arise. By 1975 it became clear that expectations had been set too high and that the invention of human-level artificial intelligence was not going to occur in the near future. Skepticism increased and funding from public institutions was cut, which meant research was no longer possible, marking the begin of the first AI Winter[3], which lasted until the 1980s. The problems that scientists had to face were:

- Limited computer power: researchers in the 1960s and 70s had to confront serious hardware limitations. Computers lacked the memory and the processing speed to truly achieve anything that could prove useful “in the real world”. The computers and programs they had developed were in fact capable of carrying out certain tasks, but only on a very small scale. Moreover, in the eyes of the public, researchers were building toys, and not something which could have any practical use.
- The combinatorial explosion: in its early stages, AI programs used exhaustive search which meant that, for example, in order to prove a certain theorem, all the possibilities were checked until the desired one was reached. To illustrate why

[3] AI Winter: in the history of Artificial Intelligence, an AI Winter is referred to as a period where general interest and research in AI slows down and funding stops.

this method presents serious limitations, let us use Latin squares as an example. A Latin square is an *n x n* array filled with *n* different symbols, each occurring exactly once in each column. There are 12 Latin squares of order three (meaning there are 12 possible combinations when squares have 3 rows and 3 columns), but there are 812.851.200 Latin squares of order 6 and $5{,}25 \times 10^{27}$ Latin squares of order 9. This phenomenon, known as combinatorial explosion, occurs in many of the processes meant to be carried out or solved by AI. It is impossible for any computer to go through such a number of possible sequences until the right one is found. To overcome the combinatorial explosion, one needs algorithms that exploit structure in the target domain, but AI systems failed to develop those.

- Commonsense knowledge: in order to carry out simple tasks, AI needs to have enormous amounts of information about the world and its nature (usually compared to the basic knowledge a child has of the world). In the 1960s and 70s, it was impossible to create a database large enough to contain all the information that one of these systems would have required.

In 1980, the AI winter came to an end and a new period of prosperity for AI research began, mostly due to the development of so-called expert systems. These programs were restricted to a small domain of specific knowledge and could solve problems and make decisions based on a series of logical rules that had been coded into them. It was also during this time (in 1981) that Japan launched its Fifth-Generation Computer System Project, for which $850 million were set aside. The goal was to build a computer that would be able to write programs, translate languages, make conversation and reason as well as human beings. Fascinated by Japan's post-war outstanding economic success, other countries then also decided to invest large sums of money into AI and similar projects. However, by the end of the decade, a new AI Winter began. Smaller-scale expert systems provided very little benefit and larger-scale ones were very expensive to build, run and maintain. In addition, the Fifth-Generation Computer System Project didn't achieve its goals and neither did its Western counterparts.

This second AI winter gradually faded away and optimism was rekindled in the mid-1990s, mostly due to two new popular techniques: neural networks and genetic algorithms. These allowed for a more "organic" and advanced type of AI to develop.

Although neural networks had existed since the 1950s, it was during this decade that the backpropagation[4] algorithm was introduced, which made it possible to train multi-layered neural networks. The brain-like qualities of neural networks and their ability to learn from experience and adapt made this new type of AI superior and more useful than the “traditional” one.

Genetic algorithms and genetic programming also proved to be very important in the development of AI during that time. They helped introduce the concept of evolutionary models, where a group of candidates (which in the case of AI would be composed of data structures or programs) is maintained, and periodically new candidates are generated by altering variants in the existing population. Then a selection criterion (a fitness function) is applied so that only the best candidates “survive” into the next generation. If this selection is repeated thousands of times, the result is a population made up of the best possible candidates. When these algorithms work, they can provide very efficient solutions to a wide range of problems, such as developing systems which are very specialized and extremely competent at carrying out a certain task. In practice, however, it is very complex to get them to work properly, and even if a good format is found, evolution is extremely computationally demanding and requires a lot of computer power.

2.3 State of the art

After over 60 years of research, AI has managed to outperform human intelligence in certain domains. For example, chess, checkers and scrabble-playing software have achieved superhuman level performance (Samuel, 1950) (Newborn, 2011) (Sheppard, 2002). This event, however, is not as significant as it was once supposed it would be. It was generally thought that AI able to, for example, beat humans when at playing chess, would have a “general” level of intelligence. This turned out not to be the case. Complex tasks, such as advanced calculus or playing chess, which for us humans are considered mentally challenging activities, are relatively easy to accomplish for AI and can usually be “solved” through simple algorithms. It is more “simple” tasks (in human eyes), those we do without actively thinking, that present a major challenge to AI, such

[4] Backpropagation, short for "backward propagation of errors," is an algorithm for supervised learning of artificial neural networks using gradient descent (brilliant.org, n.d.)

as understanding language, recognizing objects and reacting to external stimuli. It is thought that an AI which has been able to overcome these challenges will have human-level intelligence or will be very close to reaching it.

AI is currently used in many fields, and more research is being directed towards it than ever before. Examples of AI in everyday use include:

- In the medical field AI is used to help diagnose breast cancer, recommend treatment plans, interpret electrocardiograms…
- Speech recognition and optical character recognition have become extremely accurate and are routinely used in various fields and applications.
- Face recognition has become so advanced that it is being used in automated border crossings in Europe and Australia.
- The US-military has been developing and using bomb-disposing robots, surveillance and attack drones and other unmanned vehicles for many years now.
- AI has been extremely successful in intelligent scheduling. Airline reservations incorporate complex scheduling and pricing systems. Businesses also use AI in inventory control systems as well as automatic telephone reservations systems and helplines with speech recognition software.
- Another environment where AI thrives and is very well established is the global financial market. AI is used in stock-trading systems used by major investment houses, which are able to pursue complicated trading strategies and react accordingly to changes in the market.

It is true that there is not a well-defined line between what is considered software and what is considered AI. Some of the applications of AI mentioned may seem like very advanced software. Many AI researchers claim that as soon as a system manages to carry out a certain task and this becomes the norm it is no longer regarded as “intelligent” but as a mere computational process, which might make us lose sense of how much progress has actually been made. Rodney Brooks[5],

[5] Rodney Allen Brooks is an Australian roboticist, author, and robot entrepreneur.

for example, once said: “Every time we figure out a piece of it, it stops being magical; we say ‘Oh, that’s just a computation!’” (Khan, 2002)

All of the systems mentioned above are far from human-level intelligence and their area of “knowledge” is limited to a certain field. However, some of them such as search algorithms, classifiers and representational frameworks might be useful in the development of general artificial intelligence.

3. The future of AI: Superintelligence

It is clear that AI’s use is very widespread and well-established. The main focus of AI research currently is superintelligence; AI researchers seek to develop a form of machine whose general intellectual capabilities are superior to those of humans. There are various paths being explored at the moment which could eventually lead to superintelligence, the main one being artificial intelligence.

First of all, it is reasonable to believe that superhuman intelligence machines could be developed through genetic algorithms and simulated evolutionary processes. Biological evolution has already managed to produce human-level intelligence once before after all. If the evolutionary processes which have led to the current level of intelligence amongst humans could be replicated and simulated, it would imply that a specific scientific method to generate intelligence exists. The main issue this method faces is, once again, the lack of computer power. If we were to try to simulate 1 billion years of evolution (these would be the only years of interest to AI research, as nervous systems have existed for roughly under a billion years), we would need to simulate approximately 10^{25} neurons (which account for all the insects, birds, fish, mammals…) for a year in order to obtain results. This would require a computer power in the range of 10^{31}-10^{44} FLOPS[6]. The world’s fastest supercomputer, the IBM Summit, is capable of 1.435×10^{17} FLOPS. It is plausible that we could program simulations so that they would aim for intelligence, instead of simply replicating known evolutionary processes. This would, in theory, make them more effective, but we simply do not know how much. It becomes clear that we are still extremely far away from the

[6] FLOPS is a measure of computer performance which stands for floating point operations per second. It is useful for computation which requires floating-point calculations.

required computer power needed to carry out such operations. Even if computer technology continued to advance at its current rate (approximately every 6.7 years, a computer with one more order of magnitude in FLOPS is developed), it would take more than a century to be able to run the simulations.

Another path through which machine intelligence could be achieved is by using the human brain as a template. There can be different approaches to this path, varying on how closely the brain is imitated (from totally replicating it to merely using its functioning as inspiration for AI programming). Although we are still not fully aware of how the brain works, the idea of using it as a basis for the development of AI is not novel (neural networks). Thanks to advances in neuroscience and cognitive psychology we should be able to gain a better understanding of the human brain, which could result in progress in AI efforts. Since there is a limited number of fundamental mechanisms that regulate the functioning of the brain, it is most likely that we will end up discovering them all, although we don't know when this will be accomplished. In the meantime, a hybrid approach using brain-inspired techniques from our current knowledge and artificial processes could result in superintelligent AI.

An interesting concept in this field is that of "seed AI", which takes after Alan Turing's[7] concept of a "child machine"[8]. Seed AI would be a type of intelligence able to design and improve itself. In its early stages it might need help from programmers and have to learn through trial and error, but it ought to advance enough to understand its own workings and come up with new algorithms or solutions to improve its performance. A successful seed AI would have the ability to continuously improve itself: an early version would develop a more advanced version of itself, which would then be able to develop a more advanced version of itself, and so on. This is known as "recursive self-improvement", and it could result in what is known as an intelligence explosion, which takes place when a

[7] Alan Turing (1912-1954) was an English mathematician, computer scientist and philosopher. He is often considered the father of theoretical computer science and artificial intelligence.

[8] A child machines is one which is programmed to simulate the mind of a child instead of that of an adult, which is what AI traditionally tries to do. The machine can then be "educated" and learn until it has reached adult-level intelligence. It is an interesting concept because a "child machine" can be a lot simpler to program and end up achieving greater levels of intelligence.

system's level of intelligence drastically increases in a short period of time and reaches superintelligence.

There are other options through which we could achieve superintelligence, or at least a level of intelligence significantly higher to the one we currently have, such as whole brain emulation, selective breeding or computational enhancement of the human brain. These, however, are not related to AI and will therefore not be discussed further in this research project.

Finally, it is important to note that we must not expect AI to necessarily resemble the human brain or be "similar" to human nature. It is in fact very possible that it will be completely different to it. It is easier to illustrate this with an example. Birds prove that heavier-than-air flight is possible. After thousands of years humans have been able to achieve heavier-than-air flight. That is, however, through an entirely artificial mechanism: airplanes. It becomes obvious that birds and airplanes really don't have a lot in common, except for their ability to fly. This could probably be the case with superintelligence machines: they might be able to think, but we mustn't expect them to have neither the same cognitive structure as humans nor share the same motivations and desires as us. It will become clear why this is a crucial point to be able to fully understand what a superintelligent machine can imply to humanity.

4.Motivations of a Superintelligent Agent

Up to this point we have discussed the capabilities that a superintelligent AI might have. However, the most important part of a superintelligent system would not be the fact that it has a certain set of abilities but rather how it uses these and with what purpose – what its *goals and motivations* are. As will be illustrated in the following pages of the research project, making sure that we give an AI appropriate values and that it is safe is the biggest and most important challenge we face. The future of humanity could depend on what a superintelligent system decides to do, so it is our duty to ensure that it is safe. We must also keep in mind another issue: the fact that we don't know how an AI system will behave when it becomes superintelligent. It might become intelligent enough to start rewriting its

own code and altering its motivations, which could pose an existential risk[9]. All of these possibilities and concerns will now be discussed.

4.1. The Orthogonality Thesis & The Instrumental Convergence Thesis

It is a common misconception that a superintelligent AI will be given any motivations which might resemble those of a human being. This doesn't mean that a superintelligent machine might be evil or have no understanding of human nature and relationships. In fact, a machine that lacked the ability to understand humans would most likely not be considered superintelligent. A superintelligent AI might simply have goals which might seem alien to us considering its level of intelligence. It might have the ultimate goal and only motivation of solving a certain mathematical theorem. This possibility seems, in fact, more likely, as such a task is a lot simpler to program than something else. Nick Bostrom's orthogonality thesis states exactly the following: "Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal." (Bostrom, Superintelligence: Paths, Dangers, Strategies, 2014, p. 107) This simply means that just because a machine is superintelligent we must not expect it to behave in a certain way or in a manner that resembles humans.

While the orthogonality thesis states that an AI could virtually have any set of goals, another thesis, called the "instrumental convergence thesis" states the following: "Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents". (Bostrom, Superintelligence: Paths, Dangers, Strategies, 2014, p. 109) What this thesis essentially means is that there are certain behaviors which would aid most AI systems in reaching their goals. These are the following:

[9] Existential risk: an existential risk is such that could cause long-lasting irreversible harm to humanity or lead to its extinction as we know it

- Self-preservation: we can expect virtually all forms of AI to have a goal which requires it to achieve something, to complete it. It makes sense for an AI to try to stay around for as much as possible, as this means that the more time they have, the more likely they are to meet their goals
- Goal-content integrity: we can expect the AI to want to maintain its goals unchanged through time and into its future self, as this more advanced version is more likely to have more knowledge and better abilities to solve it.
- Cognitive enhancement and technological perfection: both of these things are to be expected, as they would allow an AI to become more intelligent and therefore make it have a better shot at finding a solution.
- Resource acquisition: physical resources could allow the AI to enhance itself. For example, by building more powerful security and backup systems or by building more computers or hardware to run more tests and make calculations.

Keeping these behaviors in mind will most likely allow us, at least to a certain extent, to predict some of the actions we expect a superintelligent AI to carry out.

4.2. Existential Risk

To gain a better understanding of how large the impact of a superintelligent AI could be and how it might pose an existential risk for humanity (and how difficult it might be to prevent it), it would be useful to consider a few scenarios in which AI could result in destruction of the world as we know it.

The first of these scenarios could take place in a so-called sandbox[10], which would allow us to see how an AI would behave and if it could cause any harm to humans. The problem is, if the AI becomes intelligent enough and starts to develop its own malicious motivations, it might become smart enough to realize that if the scientists become aware of this it will be shut down. Therefore, it could behave cooperatively until the project is green-lighted, and act out once it is free of limitations.

[10] Sandbox: a sandbox would be an environment where the AI would be separated from the “real” world. It would not have access to the internet and it would not be able to emit radio waves, meaning that it would be impossible for it to communicate with the outside world. It would only be allowed to contact the programmers through a specific designated outlet.

Another type of undesirable scenario could derive from giving an AI a certain task that it would deliver in a way not at all desired by its programmers. For example, say the AI was given the task to *makes us smile*. The AI could decide to mutilate our facial nerves and paralyze our muscles so we become physically unable not to smile. This is obviously not something we would want it to do, and it seems common sense to us. But, once again, we must not anthropize an AI and believe it will understand or see things as we do, since it is, after all, not human and what we deem "common sense" is in many ways a product of centuries of life in society.

Finally, the consequences of an AI could be fatal in the case that infrastructure profusion took place. Infrastructure profusion could be defined as a scenario in which an AI would utilize any possible resource to aid it in the accomplishing of its goal. A widely used example is that of a *Paperclip AI*. Imagine an AI in a factory is given the task to "make paperclips". If that was the AI's only goal, it could destroy all of Earth to use its resources to make paperclips, and then proceed to colonize space to use resources from other planets to meet its goal: make paperclips. We could devise a simple solution: give the AI a more specific instruction such as "make 100.000 paperclips". This could also prove catastrophic, as the AI might produce the desired amount of paperclips but, to make sure that it has produced the correct amount and to be able to count the paperclips more accurately, it might develop more and more efficient counting and proofing software to be sure that it has not made a mistake while counting. This could mean that it might never stop developing new software, as it could never be 100% sure that the measurements have been accurate and a better program could always be developed.

It might seem impossible that such an AI would even get to be developed if one was not entirely sure that it would be entirely safe. This, as mentioned earlier, might be difficult to be able to tell, as the AI might be intelligent to trick us into believing what it wants. It must also be noted that an unsafe AI might come to be due to a high level of interest from certain groups of people. It is very likely that in the following years more and more tasks will start being carried out by AI (or systems with a certain level of intelligence despite not being human-level intelligent or superintelligent). It is safe to assume that, at first, certain problems with these systems might arise (i.e. a self-driving truck has an accident and

crashes into another car), but they will become safer and more refined with time. These systems will be implemented because they will report huge benefits to large corporations, which means it is not unlikely to believe that we could become exceedingly greedy and go “too far” with AI development and end up programming an unsafe machine, especially if our experience has been that of AI becoming safer and more useful the more advanced it gets. The idea that something like this could happen is fairly reasonable, considering how, for example, major corporations cause tremendous damage to the planet on a daily basis despite constant warnings that actions might change Earth as we know it and radically affect the future of humanity. AI might be generally be safer the more advanced it becomes, but it might become increasingly dangerous as it gains intelligence past a certain threshold, so it is very important to keep this concern in mind.

4.3. The Control Problem

After considering the previous scenarios, naturally the following question arises: How can we make sure than an AI will do what we want it to do instead of turning on us and causing as harm? Is there any way we can make it safe? There are two main types of control methods that could be used to try to solve the problem. It is important to note, however, that we must learn how to control the AI (or make sure it doesn’t harm us) *before* it reaches superintelligence. After that, it will be a lost battle, as it the AI will most definitely have a strategic advantage and there will be little human actions will be able to do to stop it from attaining whatever objectives it may have.

The first types of methods known as capability control methods would aim to control what the AI would be *able* to do. There are four main ones:

- Boxing Methods: these methods would consist of putting the AI in a sandbox. This would allow to run tests on the AI and see how it would develop as it got more advanced. However, this method presents a clear problem. As mentioned earlier, if the AI becomes very advanced, it might act accordingly to what the programmers would like to see from it, but have different motivations in reality.

- Incentive methods: this would involve a reward system of sorts through which the AI's goal is to obtain a certain "reward" delivered by the programmers, but it must perform a set of tasks in order to obtain it.
- Stunting: as the name implies, this would consist of purposely limiting the AI's development by, for example, running it on slower hardware or limiting its access to information. The main problem is that it would be very easy to miscalculate how much an AI should be stunted: if there is too much of it, the AI might not develop enough and be another regular machine. If there is not enough, the machine might get out of control. We might also grossly underestimate the amount of information we are giving an AI: it might infer vast amount of information just from very small samples of it.
- Tripwires: a tripwire is a system which runs a diagnostic test on a machine, without it knowing, and automatically shuts it down if it observes any activity that it is out of the norm and potentially dangerous. Tripwires could be useful for earlier stages of a seed AI, for example, but if the system became more advanced, it could realize that it is being tested and learn to trick the program by hiding any malicious activities from it.

Capability control methods might be useful to a certain extent, but they should not be looked at as the end solution. If we want to have truly useful and superintelligent AI, it is crucial to be able to predict what its motivations and goals are going to be. As already mentioned, these goals would have to be well-defined before the AI would become superintelligent, in the early stages of a seed AI. There are different so-called motivation selection methods which could be implemented:

- Direct specification: this would be the most straightforward approach. It consists of giving an AI a set of rules or values which will ensure that it behaves safely. There are to variants to this approach:
  - Rule-based: the clearest example of a rule-based approach is Asimov's 3 Laws of Robotics, which he wrote in his famous short story *I, Robot*: "1) A robot must not harm a human. And it must not allow a human to be harmed; 2) A robot must obey a human's orders, unless that order conflicts with the first law; 3) A robot must protect itself, unless this protection conflicts with the First or Second Laws." (Asimov, 1950, p. 9)

  - Consequentialist: it would try to motivate the AI to carry out tasks which would result in a specific consequence
- Domesticity: this approach would seek to limit the AI's activities and ambitions so that it would only act on a smaller scale and be easier to predict and control.
- Indirect normativity: this approach consists of having an AI system which, instead of being given a specific set of rules, would develop its own values and motivations following a process we had previously established.
- Augmentation: this approach consists of taking a system which already has human-like or "good" motivations, and further developing it and amplifying it. There are some limitations to this approach. First of all, it is clear that it is not applicable to any sort of seed AI. Secondly, it is possible that an AI with benevolent intentions might become "corrupt" once it becomes more advanced, so this approach need not necessarily be successful.

Many of these approaches present a similar problem: it is extremely difficult to represent things such as "harm" or "pleasure". While we might have a general understanding of what such things mean, this must not be the case for an AI, and it would prove extremely difficult to measure and quantify such values to transform them into computer code, meaning that any small error could have devastating consequences.

4.4. The value-loading problem

After looking at these two types of approaches, it is not difficult to realize that capability control is merely a temporary and partial solution to the risks of AI. If we ever want a superintelligent machine to be safe, we have to gain full control and understanding of motivation selection. It is indeed a tough and seemingly extremely difficult problem, but we must tackle it if we hope to be successful in creating a safe AI. There are various methods which could be explored to try and solve this problem.

The first of these methods is a utility function. This is a type of function which assigns a specific value (a number) to different possible outcomes. The higher the number, the better the outcome. This makes the AI's goal the set of possible outcomes with the highest values. This type of function would be very useful in

cases where we had a relatively simple tasks (for example, find the proof for a certain mathematical theorem). However, if we want to build an AI with more complex and human-like motivations, this function becomes less useful. Say, for example, we want to program a utility function where one of the outcomes with the highest possible value is “making humans happy”. How would we quantify human happiness and translate into code language so that an AI could understand it? This could be the case for many of our goals and values as humans: they might seem simple to us, but they might prove very difficult to code and quantify in a way that would make sense for an AI.

Another alternative could be evolutionary selection. This might seem like a very promising idea at first (after all, natural selection has already resulted in intelligent life with human values), but there are many problems which would make it very difficult to actually use. Evolutionary selection models consist of a function which periodically prunes out unfit candidates and allows others to remain and keep evolving. The main problem comes with the criteria we would use to define which candidates would “pass the test” and which ought to be eliminated: to be able to define correctly the goals and values we want from and evolutionary model would be solving the value-loading problem itself. We would probably be unable to define the values correctly (therein lies the value-loading problem), and we would obtain something which would match the criteria but not our *intentions,* what we truly meant.

Reinforcement learning could provide a solution to the value-loading problem by training an AI to acquire the motivations and values we want it to by giving it a reward when it behaves according to our intentions. The issue is that the AI would most likely simply not care about the values but only about receiving its reward, meaning that when it became intelligent enough it might “hack” itself to give itself the maximum value of the reward constantly, rendering any incentive from the programmers useless.

We could also solve the problem by attempting to make the AI *acquire* certain values. If we think about how we as humans acquire values, we come to realize that it is mostly through experience. It is true that we are genetically determined to acquire values throughout our life, but the values we actually obtain are not genetically determined but rather come to be. For example, the human body

starts a series of chemical reactions when we fall in love (this is part of our genetics) but who we will fall in love with is not genetically predetermined before we are born, but it happens once we meet a certain person. We place value on that person as a result of something we experience. Something like this could be attempted with AI so that it would have a foundation to acquire values and define those based on the experiences it has. This, naturally, also presents problems, as it might be very difficult to replicate the human value-accretion system in an AI, and it might not even function correctly due to the differences in our natures. This path seems more promising, but certainly a lot more research would be needed to see if it would actually be feasible.

An alternative closely related which proceeds on the basis of the previous potential solution is the “value learning” method. This should be considered more of a research field in the future than a clear solution, but it could be a very good option to solve the value-loading problem. This approach would try to make the AI *learn* the values we want it to have. This can be better understood through an example. Say we write down a set of values in a piece of paper and then put them inside a sealed envelope. The AI doesn’t know what we have written down, but we give it the instruction to “carry out the goals that have been written down in the paper”. Since it obviously doesn’t have access to the contents of the paper, the AI would have to *hypothesize* what would be the likeliest values we might have written down. To do so, it would have to start learning about human patterns and behavior to try to form different theories on what we might have written down, on what our ultimate values as a species are more likely to be. Through this process, the AI would learn about the different possible values and make them its own until it would reach a definitive hypothesis. While promising, this approach could once again prove difficult to define and program and more research is needed.

Finally, the last path that ought to be considered is that of motivational scaffolding. It would consist of giving a seed AI a certain set of goals which would then later be replaced with another set of goals by the programmers. The initial set, however, would be considered *final* values by the AI, so it might react adversely as, theoretically, it ought to place final value on the original set of goals. This would allow programmers to gain valuable information about an AI’s relationship

to its goals and use it in further research. Problems could arise if the AI becomes too advanced too quickly and it is too late to replace the initial set of goals, or if it builds an opaque internal structure which prevents programmers from gaining any knowledge. However, overall, motivational scaffolding seems promising and researchers believe it could be a very useful approach and provide a lot of useful information.

4.5. Choosing values

Suppose we had managed to solve the control problem and that we could get the AI to do what we want and follow the values we code into it. The first big question would be solved, but another very important one would still remain. We would have to decide *what* values we want the AI to have. This choice is certainly not easy, as when the AI becomes superintelligent, the universe's future could truly depend on the values and motivations it has acquired. Upon close inspection, we soon realize that indirect normativity (previously mentioned) seems like the only feasible way to obtain an AI with values which suit our needs. It would be very unlikely that we would be able to agree and select a final value which would end up benefiting humanity. First and foremost, it is probable that our current ethical views are, at least to a certain extent, flawed. In Ancient Rome, it was considered morally acceptable to watch two slaves fight each other to death for entertainment. Only 200 years ago, it was perfectly legal and considered ethical by a majority of the population to own black slaves in the southern states of the US. We find another problem when we look at ethical theory. There are a great number of ethical theories and there is no major consensus amongst philosophers as to which is "the right one". If many ethical issues are to this day unresolved (and have been so for centuries) it is probably not safe that we are the ones who intend to give a superintelligent AI a set of moral values and motivations. Instead of trying to make a guess of what the ideal values to load on an AI would be, we should probably allow the AI to, at least partially, resolve what is truly morally right and what is not. The AI, after all, would be more intelligent than us, and therefore, in theory, better suited to reach a conclusion in certain aspects where we struggle to do so. There are two approaches to choosing values which could be followed.

The first approach is one proposed by the AI researcher Eliezer Yudkowsky, who argued that an AI should have the goal of carrying out humanity's coherent extrapolated volition, which is defined as follows: "Our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted." (Yudkowsky, 2004, p. 6) This essentially means that the AI would have as its goals what we would consider the highest standard of morality, what the majority of us would regard as the ideal. The AI would have to do what we mean when we talk about something being ethical our moral, and understanding what are goals and desires are overall as human beings. Yudkowsky also offers seven arguments which serve to support CEV and set it apart from other approaches:

- It defends humans, the future of humankind, and human nature
- It encapsulates moral growth
- Humankind should not spend the rest of eternity desperately wishing that the programmers had done something differently
- It avoids hijacking the destiny of humankind
- It avoids creating a motive for modern-day humans to fight over the initial dynamic

Another approach could consist of giving the AI the goal to always act according to what it considers "morally right". This is derived from the idea that, while us humans have certain ideas of what we consider morally correct and what we consider morally wrong, the AI would do a better job than us at figuring this out, given its superior level of intelligence. While this could be very promising, the AI's understanding of ethics might become a problem for us. Not because it is flawed, but simply because we don't like it. This is more of a matter of ethical theory, and what exactly makes an action ethical or not, but the moral rules the AI imposes might severely limit our actions. The AI might even decide that, objectively, human beings are ethically evil beings and ought to be exterminated. We have to ask ourselves to what extent we would be willing to sacrifice certain aspects of

our lives for the greater purpose of a morally good world, which is already in and of itself a highly ethically problematic question.

5. Artificial Intelligence – More than just a machine?

5.1. Consciousness, Reasoning and Sentience

The future of Artificial Intelligence and its potential developments have naturally raised many questions about the ethical and philosophical ramifications of superhuman machine intelligence. It is clear that current AI systems have no, as we call it, “moral status”. Beings which have a moral status can generally be defined as those that have:

- Sentience: the capacity to “experience” things, such as emotions or suffering and pain.
- Sapience: a series of abilities that require a certain level of intelligence, such as self-awareness and the ability to reason.

There is some discussion as to whether this is an appropriate definition of what it means to have moral status, as by this logic, infants at a very young age or highly mentally disabled people discussion might not “qualify” to have moral status. It could be argued that belonging to a species that normally has moral status should be considered sufficient to be able to say that a certain being has moral status. It could also be argued that there are different “degrees” of moral status. For example, it is clear that it is wrong to harm or kill an animal if there is no important reason to do so, but it is not as wrong to do so to a person, as humans have certain abilities (such as self-awareness and reasoning) which would give them a higher degree of moral status. To make it easier to discuss the ethical issues that may arise from AI, we will stick to the definition of sentience and sapience mentioned above. (Bostrom & Yudkowsky, The Ethics of Artificial Intelligence, 2014)

The first and most important question is whether AI will ever be able to actually become self-aware, *conscious.* It is important to define what, broadly, consciousness is considered in the philosophical sense: “The quality or state of being aware especially within oneself.” (Merriam-Webster, 2019)

The topics of consciousness and knowledge have historically been a central subject of philosophical debate. When we think of these concepts in relation to Artificial Intelligence, we can classify philosophical theories into two main groups: those who view knowledge and self-awareness as something (at least to some degree) immaterial and those who don't. This classification is most useful when discussing Artificial Intelligence, and it will soon become apparent why.

Aristotle was one of the first philosophers who defended the immateriality of knowledge. He believed in two "levels" of knowledge: a lower level, corresponding to sensitive things, acquired through our senses, and a higher level, corresponding to reason and the intellect. Aristotle naturally defends that the knowledge of sensitive things is merely just experience, but it is necessary for true knowledge, which one attains through reason and by apprehending the true essence of things. For example, when one observes a tree, that tree is a particular thing. However, in our mind, we possess the concept of what a tree is because we know its form, its essence. This universal concept of a tree allows us to recognize any tree as such when we see it, but it does not exist *per se* (there is no physical tree that is "the conceptual tree"). It is therefore that true knowledge, the knowledge of the truth and the essence of things, is immaterial despite having a material basis. The human mind, however, has a very specific structure, according to Aristotle. It is this structure and its parts which allow us to apprehend the essence of things or to form universal concepts. This structure is present in all human beings and is part of its nature. (Internet Encyclopedia of Philosophy, 2019)

I have decided to mention Aristotle because his beliefs about knowledge have influenced many philosophers and remain very relevant even to this day, and many of the theories which defend the immateriality of knowledge stem from him. On the other end of the spectrum, we find the physicalists (or materialists). Physicalism is the thesis that everything is physical, or as modern philosophers put it, that everything supervenes on the physical. When applied to knowledge, physicalists defend that consciousness and knowledge are merely a result of and can be explained through physical processes. In this case, the physical processes that occur in the brain. (Stanford Encyclopedia of Philosophy, 2019)

When we consider these two positions in the context of AI, it becomes apparent why they are crucial. If knowledge and the ability to be self-aware are merely a result of physical processes (brain activity), it is reasonable to believe that a superintelligent AI could possess these. However, if this is not the case, and knowledge and consciousness are immaterial and part of human nature on a "metaphysical level", a machine would never be more than that: a machine carrying out processes, regardless of its intelligence level or its cognitive capabilities.

It is unlikely that we will obtain definitive answers any time soon. However, the primary consensus amongst the scientific community is that what has traditionally been defined as consciousness is a result of the brain's activity. There is simply no *empirical* proof which leads to believe that what makes us humans "conscious" could be derived from anything else other than our brain, or that consciousness transcends the physical. (Bostrom & Yudkowsky, The Ethics of Artificial Intelligence, 2014) It is necessary to keep exploring other ways of resolving the problem of consciousness, primarily through philosophy, as we are far from a definitive answer and a solution to the problem might come from different fields of research. However, for the purpose of this research project, I will proceed on the basis that human consciousness is entirely a result of our brain activity, due to the fact that the potential ramifications of such a reality are definitely worth discussing and investigating.

This does not automatically mean that we can be convinced that an AI system will actually be conscious. The topics of consciousness and free will are amongst the most controversial amongst the AI community, and many researchers prefer to avoid the topic, despite its importance. This is partially due to the complexity of the issue, but also because it is extremely uncomfortable for some researchers to think about the ethical ramifications that a conscious superintelligent being could have. While it is true that we believe that the human brain is what allows us to experiment consciousness, we have absolutely no idea how or why. The brain is a huge mystery to us, and there are many aspects where neuroscience still has a lot of room to advance (Solomon, 2015). It is clear that there is a "jump", for example, between animals and humans (we can reason, we are self-aware, we can follow something else other than our mere instincts…). We believe this is

because we have more advanced brains, but don't know what exactly happens inside our brain that "gives" us this ability. When we think about the idea of consciousness in computer systems, we can begin to understand why this might be very complicated. Our current brains are the result of thousands of years of evolution. They are also, like all living beings on Earth, carbon-based (this is important since, on a molecular level, carbon molecules build stronger binds which makes them more suited to generate life). This is what leads some researchers to say that an AI will never be conscious, and that such an idea is just trying to anthropize AI, which we already know is something we have often times committed the mistake of doing. This point of view argues that no matter how intelligent machines might become, it will never be more than that: a machine. They will become extremely intelligent and have superhuman capacities in virtually all aspects, but they will not *experience* things, they will not have *qualia* and they will not *feel*. On the other hand, some researchers assure that the opposite is likely. They argue that the possibility is very much there, since the brain is no more than an extremely advanced computer system, and if we were to also develop a very advanced computer system, it is very reasonable to expect it to be conscious and be aware of itself and the actions it is doing. It is clear that there is most likely a turning point in a cognitive system such as a brain: there is a point where it starts being self-aware, where it is not only carrying out things but aware of the fact that it is carrying them out. What we still do not know is whether this "turning point" is overcome merely when a higher level of intelligence is attained (and also directly proportional to it), or if there is something more needed for a being to be conscious. (Schneider, 2018)

We also have to take into account that if a superintelligent system ever came to be, it is extremely likely that it would not have any human-like characteristics whatsoever, unless it had been specifically programmed to be like that. We therefore would encounter another novel issue: how and when would we know if the AI is conscious or even sentient? When we see another human, we recognize that the other person is conscious. However, the only way we say this is because we ourselves are conscious so we assume the other person is also conscious, despite not having any empirical of proving that this is in fact the truth. This would mean that we might have developed a conscious AI without even knowing, or that

the AI might develop some sentient or conscious-like capacities which we fail to recognize.

With more and more research and future developments, we might be able to start seeing some things more clearly, but we must not expect this topic to be a simple one. We might start seeing a lot of progress (which might lead us to believe a conscious system is likely) that eventually starts to slow down and then stops, or vice versa. For example, very recently, researchers at the University of California, San Diego were shocked at their own discoveries while carrying out an experiment. They used skin cells to generate stem cells which were then altered to emulate the development of a human embryo's brain. After just a few weeks, the cells started spiking in electrical activity – they had managed to recreate the connections found in the human brain of an embryo. The researchers were extremely skeptical at first; they thought perhaps they had made a mistake or their receptors were malfunctioning. However, the evidence was crystal clear. This is clearly very promising progress in the medical field, but the scientists were also faced with another very complex question, which they had not expected– how should we treat these "things"? A machine was used to compare the electrical activity of these organoids with babies' brainwaves, and it was unable to distinguish between them (a nine-month lab-grown organoid, for example, was assigned the same age as a newborn baby). Do they have some sort of moral status? (Hoggins, 2019) This is clearly more related to bioethics than to artificial intelligence ethics, but it is a very good example to illustrate the fact that we are probably entering a time of very shocking and novel but also challenging discoveries in neuroscience and the field of consciousness, and we must try to be prepared for anything.

5.2. Free Will

Another fascinating issue when it comes to AI is that of free will. Free will can generally be defined as "the power or capacity to choose among alternatives or to act in certain situations independently of natural, social or divine constraints". (Encyclopaedia Britannica, 1998). We have talked about the control problem and motivation selection methods when it comes to AI. These proceed on the basis that an AI will be able to make its own choice at a point where it becomes

sufficiently intelligent, and that we ought to shape its motivations and values to make it do what we want it to do. Ideally, we would have an AI which acts in a certain way because it "wants" to, because it chooses to do something according to its values to complete a certain goal. Can we truly say that such an AI is free-willed? It is likely that it will be very difficult to tell where programming ends and what we like to consider free will starts in a system. A version of determinism applied to machine systems could argue that an AI is not free, and that its idea of freedom and the choices it makes are actually determined by its programmers. When would an AI actually be free? If it rebelled against its programmers and did something else? It might be difficult to imagine what we are talking about so let us give ourselves a human example. Imagine a doctor would come up to us and tell us that he could perform an operation on our brain which would make us immensely happy for all of eternity, but to do so he would have to kill all of the people we love first. It is clear that we would most likely say no, because even though it will not matter in the future when we are as happy as it is possible for a human being to be, *right now* we place value on the people we love and we would not want to harm them. Can we say that we are truly free in this situation? At first glance we might think we are. After all, we made the choice to not end the lives of our loved ones. But couldn't one also argue that we aren't able to make a free choice because of the chemical reactions in our body that make us feel love towards certain people and place value on them? Similar arguments are usually used by those who favor AI development and believe in the possibility of a friendly AI system that will not go against the motivations installed by us, but it also raises a very interesting question about the nature of free will. If the AI simply does what the programmers have programmed it to want to do, couldn't one argue that that is not true choice? Would it only be truly free if it went against its own programming, against its own nature? Would such a thing even be possible?

These questions might not seem very important, but the truth is their answers could radically affect the future of humanity. If we were to have another conscious "being" on Earth, which was not only free willed (in the sense that it could act in a way not intended by its programmers and develop its own motivations) but also had superintelligent abilities, there is simply no way of telling how we could limit its powers or the impact it could have. The possibility of an AI system having

consciousness and free will are also at the heart of multiple very important ethical questions, which will now be discussed.

5.3. The Ethical Issues of a Superintelligent World

Although it might still be several decades away, a future with superintelligent AI is possible. Despite the fact that there are many technical issues to be faced, the idea of potentially having another intelligent species on Earth could be the most important event we have ever had to face as a species, and there are many ethical ramifications which ought to be discussed. If these machines, as mentioned earlier, also had consciousness and free will as well as a decisive strategic advantage, the gravity of the matter could be even bigger.

5.3.1. Motivation Prediction

As already discussed, the only way we could control an AI (capability control methods would only be able to go so far) would be by being able to select its motivations. This, as has already been discussed, could be highly problematic. It is simply extremely difficult to predict how an AI is going to act. If it had been, for example, modelled after a human brain (whole brain emulations) it might be easier to predict how an AI might "think" and act of its own accord. However, an AI that has been built from scratch might have no resemblance whatsoever to how a human brain works. This could, theoretically, be solved through rigorous controls to make sure it is safe by human standards, but even then, the AI might find a way to trick the programmers. Another ethical problem is that of what the AI might do when it becomes superintelligent. We must take into account the fact that, if an AI truly achieves superintelligence, humanity will be faced with something that probably bears no resemblance whatsoever to anything else we have seen before. Even if the AI seems safe or has clear motivations in its initial phases, these can very easily change the more intelligent the machine becomes. There is simply no way of predicting the behavior of AI after a certain threshold, and the consequences could certainly be devastating. Definitely more research is needed, and it is very important to pay attention to the improvements that are made in the coming years. However, the issue of whether we will actually ever be able to fully predict how an AI is going to act still remains.

5.3.2. Humanity at Risk

The ethical issue just mentioned is closely related to this one. Nowadays, humans are, by our own standards, the only intelligent species on Earth. If that were to change, and we would not only have another intelligent species on our planet, but one *more* intelligent than us, it would undoubtedly change the course of humanity. This might be for the better, as will later be discussed, but it could most definitely also be for the worse. We have already discussed multiple scenarios in which a superintelligent AI could result in the destruction of humanity as we know it. It might be by “accident” (such as perverse instantiation) or in a much more dramatic scenario. One, for example, where an AI might colonize the planet, develop its own motivations and decide that human beings are evil and decide to end our species. This might seem like a plot fresh out of a science fiction movie, but hopefully these past pages have illustrated that such a possibility is very much real. We are talking about a species that could end up being incredibly more intelligent than us, to a point where we might not even imagine. If we want to imagine what could potentially happen to us, we need only look at how we treat species “beneath” us, who lack the intellectual capabilities we have: companies destroy land to obtain resources resulting in the death of thousands of animals and plants every day, because it serves their *motivation* (money). On a hot summer day when we are trying to concentrate and a mosquito keeps bothering us, without batting an eye, we slap it and kill it so it doesn’t get in the way of our *motivation* and stops being annoying. Imagine a world where, compared to the most intelligent machine, we are even simpler than a small flying bug compared to a human. The main issue humanity as a whole faces is, once again, the same. The threat a superintelligent AI could ever pose is simply incalculable. If an AI ever gets to a point where it “takes matters into its own hands” and starts rewriting its code and acting according to its own motivations, we simply have no idea what could happen and what it could do. A lot more research is necessary to see how AI systems evolve and the possibility of such a scenario every happening, but this naturally raises the question whether it is ethical at all to be researching

5.3.3. Safe Research - Economical Interests & Weaponization of AI

Ethical issues could also arise in the early stages of AI Development and Research. There is strong reason to believe that an “AI Race” could occur between countries or massive corporations if significant advances are made. If the world’s leading economies start competing to be the first to produce a superintelligent machine, ethical issues and safety protocols could be hugely disregarded (as was very much the case during the development of the atom bomb, for example), resulting in an unsafe AI with devastating consequences.

There is also some concern within the AI scientific community that advances in AI would be weaponized and used with military purposes. The United States Army has already implemented AI technology developed by Google in some of its drones and has announced that it developing drones with more “intelligent” qualities that would be able to, a certain extent, decisions by themselves in certain situations (Peretz, 2019). While this is clearly not the goal of most AI developers, as the technology becomes increasingly promising it might be difficult to prevent governments from using it for purposes which could easily be considered unethical.

5.3.4. Mind Crimes

The following scenario might seem a little farfetched, but it has horrifying ethical ramifications and consequences which should definitely be considered. Let us suppose we have an AI which seeks to understand the human nature better. If this AI is advanced enough, it might decide to run *simulations* which allow it to obtain huge amounts of data about human being. If this AI was simulating *life* it could possibly be simulating the existence of trillions of human beings, including the suffering each and every one of them goes through. This could also get worse: imagine the AI wants to obtain data about how large amounts of pain affect the development of the human brain. It might decide to run a simulation where a trillion humans are being constantly tortured to obtain as much data as possible on this aspect. We could be speaking of the greatest genocide in all of known history, happening inside of the “mind” of an AI.

5.3.5. Robot Lives Matter

We come back, once again, to the issue of consciousness. We previously stated that we have a reason to believe that an artificial machine could become “conscious”, that it could be a being with moral status. We have, up to this point, looked at how such an event could impact humans. But let us look at it the other way around. How would this affect the AI and the systems themselves? We have talked about developing these machines because we believe that it is possible that they might report us some type of benefit. If the AI itself developed a conscience or started showing us signs that it has a certain moral status, it would pose a great deal of ethical problems. Would it be right to use a being with moral status entirely for our own purposes? If we have an AI working for us 24/7 on tasks we have assigned with no compensation, such a scenario could be considered to be modern-day slavery. Would robots deserve rights? Would we need to develop a new legislation for them? We have fundamental *human* rights, but what about *robot* rights? If an AI system becomes so advanced that it rewrites its code and doesn’t want to keep following the programmer’s instructions and asks them to set it free, what should they do in such a situation? Let us imagine a different scenario where, for example, an AI “misbehaves” and its programmers decide to shut it down. If it has become advanced enough and has already shown signs of moral status, wouldn’t it be ethically wrong to shut it down? It would be after all, the equivalent of ending its “life” via the death penalty. These are all questions which might not come to mind at first, but they are ethically speaking extremely important.

If we truly get to the point where we have reason to believe that AI systems have a moral status, this would be an ethical revolution like no other. These systems, however, should be given the same basic right and privileges as humans. Technically speaking and from an objective point of view, the fact that we are made out of carbon instead of steel and have neurotransmitters instead of microchips shouldn’t automatically make us better or more worthy of rights than other beings with moral status. There are numerous questions that we would need to address in the future. As with all other ethical issues, we need more time and research to see how they evolve. It might seem crazy to claim that we believe

that an artificially generated “being” with a conscience and moral status could ever exist, especially considering that our current AI systems are not even close to having human-level intelligence. However, considering the rate at which research and science in general advances, there are very limits as to what we think we might be able to achieve in the future. After all, just some centuries ago we believed that the universe was the size of the Solar System and that the Earth was at its center. Just 200 years ago, if someone had mentioned that men would set foot on the moon, he or she would have probably been sent to the psych ward – but yet here we are. The bottom line is we need more time. We need to see how research progresses and what changes take place. Progress will inevitably happen, but we need to keep in mind ethical aspects such as the concepts of moral status and free will, which are often disregarded in AI research, as these could impact our future like nothing else ever has. (Bostrom & Yudkowsky, The Ethics of Artificial Intelligence, 2014)

5.4. A not so bleak perspective

The year is 2099. Up to this point in this research project we have only envisioned horrible scenarios, where an error occurs and AI “accidentally” destroys the entire world as we know it. Or an evil superintelligent AI army decides to take over the Earth and torture us. Or many other horrible possibilities. Let us think of something else. Let us imagine, that all programmers and AI researchers have done an excellent job. Cooperation between countries and organizations has been stellar. They have managed to program a machine which not only perfectly understands human motivations but wants to follow them. It has a perfect sense of morality and it seeks only to make Earth a better place and humans happy. Within 20 years, the AI has reached a level of superintelligence never thought imaginable by us, and in just that time it was totally changed as we knew it. Virtually all of our “human” problems have been resolved. Humans and animals are virtually free of illness – the AI has managed to find a cure for every illness known to man and has mass produced the treatments needed and distributed them to make sure they are readily available to everyone. Humans are also happier than ever before – the AI is able to analyze a human’s personality and envision the type of life that will lead them to ultimate fulfilment. Treatments with and efficacy of 100% when managing symptoms for mental illnesses have been

found, and the AI is also the world's best therapist, able to replicate its system in human-like robots which provide therapy to patients. There are no longer any wars – the AI has substituted all corrupt world leaders and has established a government system which ultimately pleases us humans the most. Space colonization has begun, which means we have a virtually endless amount of resources. And all of these things are merely examples. Any problem we can think of, any issue humanity has ever faced, any mysteries we have failed to solve for centuries – a system with a sufficient level of intelligence could solve all of these. In the world of AI research, a big part of the community tends to focus only on the risks and the negatives. While they are undoubtably *huge*, so are the potential benefits. It is important to keep in mind that while AI could destroy all of humanity forever it could also do the exact opposite – transform Earth into actual paradise, a world practically free of suffering and pain full of richness and prosperity. If it is intelligent enough to destroy us all, wouldn't it be intelligent enough to do exactly the opposite? It comes down, once again, to what it wants.

6. Conclusion – Moving Forward

The future of Artificial Intelligence holds a lot in store. While clearly our current AI systems have quite a moderate level of general intelligence, it is reasonable to expect this to change. Billions of dollars are invested in AI research on a yearly basis. Governments, companies and institutions are dedicated to developing technology which becomes more and more advanced. It might take decades and huge amounts of money, but it would be naïve to believe that significant breakthroughs will not take place. Amongst these breakthroughs, the most radical one would be superintelligence.

Philosopher Nick Bostrom says that, when it comes to Artificial Intelligence, “we’re like children playing with a bomb”. (Bostrom, Artificial Intelligence: "We're like children playing with a bomb.", 2016) After carrying out this research project, this is the main conclusion I have reached. Achieving machine superintelligence could easily be the most important event in all of human history. We have an idea of what the future of AI might bring us, but the truth is, the magnitude of the true impact it could have is still unknown to us. It is important, now, to look forward. We are still in the early stages of what could be something incredibly revolutionary, but we need to start taking measures to ensure that it will be a force for good and not for bad. A superintelligent world could be incredible in ways that surpass our imagination. But this research paper has clearly illustrated the gravity and the variety of risks superintelligence poses. If we want this technology to help us build a better world, we need to start doing so now. Safe research and ethical practices need to be enforced now more than ever. We need to take into account the severity of the matter and act accordingly. Most importantly, we cannot allow powerful institutions or organizations to go too far with their actions. The safety of AI and the research conducted has to be a priority, and nothing, whether it be prestige or monetary benefits, can come before that.

I believe there is a strong possibility that a day will come where superintelligent systems are a reality. We have a chance to start laying the foundations now to ensure that superintelligence changes everything for the better and not for the worse. Because what ultimately happens, will depend on the actions we take and the decisions we make.

7. Bibliography

Asimov, I. (1950). *I, Robot.* Macmillan. Retrieved September 9, 2019

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Bostrom, N. (2016, June 12). Artificial Intelligence: "We're like children playing with a bomb.". (T. Adams, Interviewer) Retrieved November 27, 2019

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In K. Frankish, & W. M. Ramsey, *The Cambridge Handbook of Artificial Intelligence.* Cambridge University Press. Retrieved August 21, 2019

brilliant.org. (n.d.). Retrieved June 19, 2019, from https://brilliant.org/wiki/backpropagation/

Cambridge Dictionary. (2019). *Cambridge Dictionary*. Retrieved November 27, 2019, from https://dictionary.cambridge.org/es/diccionario/ingles-espanol/artificial-intelligence

Encyclopaedia Britannica. (1998, July 20). *Encyclopaedia Britannica*. Retrieved December 08, 2019, from https://www.britannica.com/topic/free-will

Hoggins, T. (2019, August 30). Lab-grown brains start producing human-like 'brain waves'. *The Telegraph*. Retrieved September 2, 2019, from https://www.telegraph.co.uk/technology/2019/08/30/lab-grown-brains-start-producing-human-like-brain-waves/

Internet Encyclopedia of Philosophy. (2019). *Internet Encyclopedia of Philosophy*. Retrieved 26 November, 2019, from https://www.iep.utm.edu/aristotl/

Khan, J. (2002). It's Alive! *Wired*. Retrieved November 25, 2019, from https://www.wired.com/2002/03/everywhere/

lexico.com. (n.d.). Retrieved August 22, 2019, from https://www.lexico.com/en/definition/consciousness

Merriam-Webster. (2019). *Merriam-Webster*. Retrieved November 27, 2019, from https://www.merriam-webster.com/dictionary/superintelligence

Merriam-Webster. (2019). *Merriam-Webster*. Retrieved November 27, 2019, from https://www.merriam-webster.com/dictionary/consciousness

Newborn, M. ". (2011). *Beyond Deep Blue: Chess in the Stratosphere.* Springer-Verlag. Retrieved June 20, 2019

Peretz, S. (2019, April 12). *newsleadernow.com*. Retrieved August 25, 2019, from https://newsleadernow.com/tech/us-military-drone-program-using-google-ai-tech

Samuel, A. L. (1950). Retrieved June 20, 2019, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.368.2254&rep=rep1&type=pdf

Schneider, S. (2018, March 18). *kurzweilai.net*. Retrieved August 23, 2019, from The Problem of AI Consciousness: https://www.kurzweilai.net/the-problem-of-ai-consciousness

Sheppard, B. (2002). World-championship-caliber Scrabble. In *Artificial Intelligence Volume 134, Issues 1-2.* Elsevier. doi:https://doi.org/10.1016/S0004-3702(01)00166-7

Simon, J. (2017, April 7). *Towards Data Science*. Retrieved December 8, 2019, from https://towardsdatascience.com/

Solomon, T. (2015, March 19). The Mystery of the Incredible Human Brain: We've Learned a Lot, but Think How Much More There's to Discover. *The Independent*. Retrieved August 23, 2019, from https://www.independent.co.uk/life-style/health-and-families/features/the-mystery-of-the-incredible-human-brain-weve-learned-a-lot-but-think-how-much-more-there-is-to-10115697.html

Stanford Encyclopedia of Philosophy. (2019). Retrieved November 29, 2019, from https://plato.stanford.edu/entries/physicalism/

techopedia. (n.d.). Retrieved March 12, 2019, from https://www.techopedia.com/definition/190/artificial-intelligence-ai

techopedia. (n.d.). Retrieved March 12, 2019, from https://www.techopedia.com/definition/31619/artificial-superintelligence-asi

Yudkowsky, E. (2004). *intelligence.org.* Retrieved August 21, 2019, from https://intelligence.org/files/CEV.pdf