KU LEUVEN

The Singularity: a philosophical perspective

Pieter De Petter

Thesis submitted for the degree of
Master of Science in Artificial
Intelligence, option Engineering and
Computer Science

Thesis supervisor:
Prof. Roger Vergauwen

Academic year 2017 – 2018

KU LEUVEN

The Singularity: a philosophical perspective

Pieter De Petter

Thesis submitted for the degree of
Master of Science in Artificial
Intelligence, option Engineering and
Computer Science

Thesis supervisor:
Prof. Roger Vergauwen

Assessors:
Prof. dr. Daniel De Schreye
Desmond Hugh

Mentor:

Academic year 2017 – 2018

Preface

I would like to thank Prof. Roger Vergauwen for providing the opportunity to explore this fascinating topic and his guidance along the way.

Pieter De Petter

Contents

Abstract

One of the first scientists to suggest the idea that humanity seems to be approaching some essential singularity as the result of never-ending accelerating technological progress was John von Neumann. While computers and modern technologies invaded all aspects of modern life around the end of the twentieth century, this captivating concept started to gain traction and 'The Singularity' was born. The main idea is that if technological progress will keep accelerating, it will inevitably lead to artificial intelligence that will exceed human intelligence. This intelligence, often referred to as superintelligence, will in turn be capable of creating even more intelligent systems, leading to an infinite intelligence explosion. The outcome is not clear but its impact is expected to be so deep that it will irreversibly transform human life.

After a clarification of the singularity idea, evidence is presented that technological progress is accelerating and will most likely keep accelerating over the next decades. This enforces the idea that artificial intelligence will reach and exceed human intelligence in the near future. Various paths towards the next step, superintelligence, are analyzed and deemed plausible. The final step, from superintelligence to the singularity, is less clear. Various philosophical and technical objections against an infinite intelligence explosion are evaluated. The results are inconclusive so what will happen afterwards remains pure speculation, although it is generally accepted that there will be profound consequences. The challenges on the path towards the singularity are more tangible and will most likely have a profound impact. Lethal autonomous weapons and the displacement of human workers by AI are two of those challenges who are evaluated. The possible advent of superintelligence warrants even more caution. A theoretical blueprint of superintelligence is reviewed, followed by a detailed look into the control problem often associated with it.

Whether the singularity will materialize, only time can tell. But various indications suggest that technological progress will at least pose a variety of risks and tough decisions over the next decades. The singularity offers an interesting framework to approach these challenges from a holistic perspective, hence a review of various elements of risk and decision theory concludes the discussion.

1 Introduction

The *technological singularity* – henceforth *the singularity* – is surrounded by an air of mysticism, controversy, fascination and fear. The first chapter attempts to clarify the concept. It starts with an overview of seminal accounts which have led to the idea of the singularity. In order to arrive at a definition of the singularity, its properties and its two main hypotheses are reviewed. Finally, the scene is set for the remainder of this text, which will not only research the singularity itself but also use it as a background to look into risks related to technological advancements such as AI.

1.1 What is the singularity?

The term singularity - not in a mathematical or space-time singularity sense but in the context of the technological singularity - traces back to the 1950's and one of the greatest scientific minds of all time, John von Neumann:

> "The ever-accelerating progress of technology and changes in the mode of human life… gives the appearance of approaching some essential singularity in the history of human race beyond which human affairs, as we know them, could not continue." (von Neumann) as quoted by (Ulam 1958, p. 5)

In the 1980's, Vernon Vinge, a computer scientist and originator of the technological singularity concept in its contemporary sense, arguably used the term for the first time in Omni, a popular science and science fiction magazine in those days, and linked it to the creation of intelligent machines:

> "We will soon create intelligences greater than our own. When this happens, human history will have reached a kind of singularity, an intellectual transition as impenetrable as the knotted space-time at the center of a black hole, and the world will pass far beyond our understanding." (Vinge 1983, p. 10)

This was followed by Vinge's famous paper "The Coming Technological Singularity" where he expanded upon the concept and truly coined it:

> "The acceleration of technological progress has been the central feature of this century. We are on the edge of change comparable to the rise of human life on Earth. The precise cause of this change is the imminent creation by technology of entities with greater-than-human intelligence; It is fair to call this event a singularity - the Singularity." (Vinge 1993, par. 1)

The concepts of *intelligent machines* and *greater-than-human intelligence*, which are closely intertwined with the singularity, can be traced back to Irving Good's classic essay on *ultraintelligent* machines where he also coined the term *intelligence explosion*:

> "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion", and the intelligence of man would be left far behind." (Good 1965, p. 33)

One of the most influential authors on the topic is the American computer scientist Raymond Kurzweil, responsible for the bestselling "The Singularity is Near" in which he writes:

> "What, then, is the Singularity? It is a future period during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed… The key idea underlying the impeding Singularity is that the pace of change of our human-created technology is accelerating and its powers are expanding at an exponential pace." (Kurzweil 2005, p. 7)

The notions of acceleration and discontinuity are common and unique to the majority of all accounts of the singularity concept, including in above seminal accounts, and allow for a clear distinction from a space-time singularity and singularity in a mathematical sense. Moreover, these notions can be considered necessary and sufficient conditions for the various hypotheses of the technological singularity. (Eden et al 2012, p. 6)

1.2 Towards a definition

1.2.1 Properties

It would be neat if there existed a widely accepted single definition of the singularity, but this is unfortunately not the case. Although it is broadly accepted that Vernon Vinge coined the term with his seminal essay from 1993 mentioned earlier, within this essay he uses multiple meanings of the concept without giving a strict definition, and he is not alone. It is unclear whether Vinge remains vague on purpose, giving the concept an aura of attraction if the idea itself weren't enough, or whether it is a logical consequence of the inherent difficulties of defining a singular event where humanity might "Enter a regime as radically different from our human past… [one that] represents the passing of humankind from center stage… [with] change comparable to the rise of human life on earth". (Vinge 1993, par. 1,4) Early in the essay one might conclude that the creation of greater-than-human intelligence is the singular event but this hypothesis is immediately weakened to a scenario where it will drive accelerating progress, leading to an "exponential runaway beyond hope and control". (Vinge 1993, par. 1) In any case, Vinge is well aware of the highly unpredictable nature of such an event "the precipitating event will likely be unexpected…Yet when it finally happens, it may still be a great surprise and a greater unknown." (Vinge 1993, par. 1)

In this sense it is not surprising that the singularity has become a concept that means different things by different authors, and even by the same author on different occasions. Nick Bostrom, a Swedish philosopher and founder of the Future of humanity institute who is gradually becoming one of the world's most influential global thinkers on topics like the singularity, superintelligence and existential risks, identifies three clearly distinct theoretical entities that the singularity might refer to in a comment on Vinge's essay in question: (Bostrom 1998, par. 1)

1. **Verticality**
   A point in time at which the speed of technological development becomes extremely great

2. **Superintelligence**
   The creation of superhuman artificial intelligence.

3. **Unpredictability**
   A point in time beyond which we can predict nothing, except maybe what we can deduce directly from physics.

Other authors like Sandberg similarly conclude that the singularity has different meanings and attempts a brief listing of them. He ends up with as many as 9 different meanings: *accelerating change; self-improving technology; intelligence explosion; emergence of superintelligence; prediction horizon; phase transition; complexity disaster; inflection point and infinite progress*. These can be clustered into 3 major groupings in line with above three distinct theoretical entities, namely accelerating change, an intelligence explosion leading to superintelligence and unpredictability. (Sandberg 2010, pp. 1-2)

Yudkowsky also agrees with those three theoretical entities and refers to them as 'logically distinct schools of singularity thought' in which every school has slightly different semantics from Bostrom's entities: Accelerating change instead of Verticality; Intelligence Explosion instead of Superintelligence; Event Horizon instead of Unpredictability. For each school, a core claim and a strong claim are identified and the most prominent authors and proponents are listed. We prefer the definitions of Bostrom since they are simple and to the point so we will not list the detailed claims since they are in line. However, an important difference between both authors revolves around the fact whether the 3 notions are mutually exclusive as Yudkowsky thinks, stating that they tend to contradict each other and strongly advocates they shouldn't be mashed up into a singularity paste. (Yudkowsky 2007a)

Unsurprisingly - and perpendicular to Yudkowsky's view - it has been argued that the conjunction of these three claims actually entails the singularity. This is more or less in line with Vince's belief and Bostrom's, although the latter immediately asks the question whether unpredictability or *discontinuity* should be considered a defining feature of the singularity. This is definitely the strongest

claim, especially when framed in his own words as "a point in time beyond which we can predict nothing". (Bostrom 1998, par. 1) The disability to predict anything is problematic, however, we don't see why this has to be a necessary condition for discontinuity. Arguably the extinction of the human race would be considered a discontinuity, a scenario often associated as an outcome of the singularity. Although predictions in such a scenario are extremely speculative, the absence of human activity can be confidently predicted for example. On the other hand it can be strongly argued that discontinuity has to be a necessary condition. Without some form of discontinuity, there doesn't really seem to be any difference between a post-singularity world and today, even accounting for superintelligence.

These three theoretical entities can then easily be reconciled with the necessary and sufficient conditions of acceleration (verticality) and discontinuity (unpredictability) via the scenario where superintelligence will be the effect of acceleration and the cause of discontinuity, a plausible scenario and a neat way of linking everything together.

1.2.2 Alternative hypotheses

Another way to approach the singularity is via its possible outcome, allowing singularity hypotheses to be split into two distinct scenarios. On the one hand there is the *Vinge Scenario*, which has the biggest support and considers the emergence of some type of superintelligence as the *singular* outcome of accelerating technological change, resulting in profound consequences. In this scenario, advancements in technology in general, and artificial intelligence and machine learning specifically, will lead to machine intelligence beyond human intelligence. These intelligent machines will be responsible for an intelligence explosion in line with Good's classical argument. Such an explosion can be seen as a *runaway reaction* of self-improvement cycles appearing faster and faster. This in turn will lead to an 'undefinable' discontinuity.

On the other hand there is the *transhumanist* or *Kurzweil scenario*, where the singularity would be the result of a *bio-intelligence explosion*. Transhumanism was coined by Aldous Huxley's brother, Julian Huxley and refers to "man remaining man, but transcending himself, by realizing new possibilities of and for his human nature." (Huxley 1927), quoted from (Bostrom 2005, p. 7)

Within this scenario progress in *enhancement technologies* will augment human cognitive capabilities, eventually leading to a posthuman race. Kurzweil himself even goes as far as postulating that posthumans will overcome all existing human limitations, including death!

> "The singularity will allow us to transcend these limitations of our biological bodies and brains. We will gain power over our fates. Our mortality will be in our own hands… The singularity will represent the culmination of the merger of our biological thinking and existence with our technology, resulting in a world that is still human but that transcends our biological roots. There will be no distinction, post-singularity, between human and machine or between physical and virtual reality." (Kurzweil 2005 pp. 8-9)

These scenarios look radically different but if one attempts to make a least general generalization, the same three notions emerge again in the sense that accelerating technological progress is leading to greater-than-human intelligence, resulting directly or indirectly in a significant discontinuity. This discontinuity is uncertain but will have profound, potentially dire consequences for humanity in the *Vinge* scenario and a relatively outspoken positive outcome in the *Kurzweil* scenario.

At first sight this looks like an acceptable attempt towards a definition for the singularity, however the notion of discontinuity without any tangibility whatsoever feels problematic. Somewhat ironically, discontinuity is closely linked to singularity in real analysis, in the sense that a mathematical singularity is in fact a discontinuity of a function (or a discontinuity of a function's derivative). In the Kurzweil scenario this looks like a non-issue at first since we are fairly confident that conquering mortality (or even significant brain enhancement) will be considered a true discontinuity indeed. But even then it can be argued that this does not encompass a real singular discontinuity in the grand scheme of things, a view echoed by Eric Chaisson who states:

> "There is no reason to claim that the next evolutionary leap forward beyond sentient beings and their amazing gadgets will be any more important than the past emergence of increasingly intricate complex systems". (Chaisson 2012, p. 413)

It has to be emphasized that Chaisson, an experimental physicist, has a different view on the singularity concept, approaching it as a common evolutionary milestone of which there were many in cosmic history, clearly implying plurality. This plainly contrasts with the *singular* singularity we have been reviewing up until now. Nevertheless, his point is intriguing and clearly shows the importance of perspective when discussing the singularity. After all, it seems not unreasonable to classify the effect of human extinction, arguably a discontinuity from every possible human perspective, as a common event using his cosmic perspective.

1.3 Conclusion

In order to attain a rigorous definition of the singularity, the correct human perspective needs to be added to the properties of acceleration, superintelligence and discontinuity. On the other hand, the intangibility and vagueness of the discontinuity aspect feels problematic. A shift towards more tangible scenarios would provide a clear scope and framework to start analyzing risks associated with the singularity, arguably the most pressing challenge if one accepts its premise. In any case, the singularity idea offers an interesting backdrop against which one can look into the challenges and risks of advanced technologies such as artificial general intelligence.

The remainder of the thesis starts with an attempt to give the singularity premise credibility by reviewing the plausibility of accelerating progress (chapter 2) and superintelligence (chapter 3). Chapter 4 reviews the link between superintelligence and the singularity and continues with a review of challenges humanity might face relatively quickly on its path of never ending technological progress such as lethal autonomous weapons. Chapter 5 evaluates a blueprint for artificial general

intelligence and the resulting control problem and risks. Finally, Chapter 6 looks into classic and cutting edge frameworks for risk analysis and decision theory that can be used to model & monitor external risks as a result technological progress.

2 Plausibility

The plausibility of the singularity hypothesis is controversial to say the least. The chapter starts with classic arguments which conclude that the singularity is inevitable. These arguments and the majority of similar arguments are based on inductive reasoning, a somewhat equally controversial topic throughout the history of philosophy of science. From this perspective, it is worthwhile to follow through with a detailed look into the problem of induction. Unfortunately, inductive reasoning cannot be justified and *inductive leaps* are required. In order to accept the overall premise of the singularity, several such inductive leaps are needed. The two most important ones are related to the properties of accelerating technological progress and superintelligence. The plausibility of accelerating progress is reviewed in this chapter and the credibility of superintelligence is reviewed in chapter 3.

2.1 Classic arguments

From the previous chapter it is clear that there doesn't exist a well-defined singularity hypothesis in the literature. Moreover the lack of a definition for the discontinuity aspect is problematic, in the sense that it is not an easy task to review the plausibility of the singularity hypotheses since it is not fully clear what the hypotheses entail.

For its proponents the technological singularity is inevitable. But in order to reach this conclusion, inductive reasoning and vague or unverifiable theories often need to be accepted. This opens the door for critics who argue that ad hoc theorizing and inductive reasoning can never obtain any scientific rigor. If one refutes to believe that inductive reasoning has any merit, the singularity hypothesis is indeed easily rejected since its main claims are generally based on inductive arguments.

Take for example David Chalmers's argument in his paper 'The Singularity: A Philosophical Analysis': (Chalmers 2010, p. 12)

1. There will be AI

2. If there is AI, there will be AI+

3. If there is AI+, there will be AI++

---

4. There will be AI++

AI should be viewed here as artificial intelligence as least as intelligent as an average human, AI+ is artificial intelligence more intelligent than the most intelligent human and finally AI++ (or superintelligence) is artificial intelligence of far greater intelligence than the most intelligent human.

This argument uses the premise that there will be AI. This can be defended by reasoning that the human brain is a machine and we will have the capacity to emulate this machine. Hence, if we are capable of emulating this machine, there will be AI. Although philosophers greatly debate whether the human brain is a machine indeed, if one accepts this idea Chalmers' premise surely is reasonable.

Chalmers's argument is loosely based on Irving Good's intelligence explosion concept and the *speed explosion* argument from the originator of artificial intelligence based on machine learning, Ray Solomonoff. Good's argument has been discussed earlier so only Solomonoff's speed explosion argument will be reviewed here, starting with a succinct summarized version:

> "Computing speed doubles every two subjective years of work. Two years after Artificial Intelligences reach human equivalence, their speed doubles. One year later, their speed doubles again. Six months - three months - 1.5 months ... Singularity." (Yudkowsky 1996)

Solomonoff himself provides a more scientific and mathematical sound formulation, relating the size of the artificial intelligence community with money spent on increasing this community. The artificial intelligence community should be understood as the total computing capability of the computer science community. He concludes that for a positive value of money spent on AI, the total computing capability will have to reach infinity at a given point in time, assuming that computation costs will keep decreasing exponentially, an assumption that will be reviewed in section 2.3 in more detail. (Solomonoff 1985)

The arguments for both the intelligence conclusion and the speed explosion underpinning Chalmers's argument - but also his argument itself - can be considered inductive reasoning arguments. Since the majority of arguments for the singularity and/or accelerating technological progress are based on inductive reasoning, it is worthwhile to review the somewhat controversial concept itself in detail, often referred to as the problem of induction.

2.2 The problem of induction

The problem of induction can be dated as far back as the ancient Greeks and the Aristotelian distinction between *demonstrative proof*, which are the things we can be absolutely certain about, and that of mere *probable knowledge*. (Gigerenzer et al 2001, p. 2) A difficult topic which is nowadays commonly referred to as the problem of induction, frequently associated with the Scottish philosopher David Hume.

According to its classic formulation, inductive reasoning is a mind activity, linking the observed with the unobserved. The core of inductive reasoning is the ability to move beyond the limits of our

current knowledge, towards new conclusions about the unknown. For example, from the fact that every swan encountered so far has been white, it is inferred that the next swan will be white as well. According to Hume, all inductive reasoning results from the relation of cause and effect. It is this relation that allows us to go beyond our current evidence in the form of interference, for example interfering the effect from its cause. After Hume identified the causal basis of our inductive reasoning, he raised a fundamental question which is now known as the problem of induction: "What are the grounds for such inductive or causal inferences?" (Hume 1739, 1748) as quoted from (Sloman et al 2005, p. 95).

More general, the problem of induction is the philosophical question whether inductive reasoning can lead to knowledge in the epistemological sense. It is about the justification of inductive methods which are critical in scientific reasoning but also in our day to day lives. The main problem is how such reasoning can be justified because of the following dilemma:

> "The principle cannot be proved deductively, for it is contingent, and only necessary truths can be proved deductively. Nor can it be supported inductively—by arguing that it has always or usually been reliable in the past—for that would beg the question by assuming just what is to be proved." (Henderson 2018)

Hume himself attempts to answer this question by presenting two arguments. The first argument is descriptive, but not justificatory, in the sense that Hume concludes that humans seem to be genetically prewired to expect observed causal relations to hold in the future. In his second argument, Hume first identifies experience as the basis of inductive inference instead of demonstrative reasoning. He then continues by demonstrating that experience by itself is inadequate as the only justification for inference. A plausible hypothesis since inductive reasoning "requires the presupposition that past experience will be a good guide to the future which is the very claim it seeks to justify." (Sloman et al 2005, p. 95) In other words, Hume suggests that it is not a rational process of thought, such as reflective or demonstrative reasoning, that takes us from the unknown to the known but rather mere experience. At the same time he argues that even a rational process of thought wouldn't suffice to justify the leap from the observed to the unobserved.

Karl Popper, arguably one of the greatest philosophers of science of the 20th century, revisited the problem of induction in 'the logic of scientific discovery' opening his seminal work with the rejection of inductive logic: "My own view is that the various difficulties of inductive logic are insurmountable." (Popper 1934, p.6) A weaker version of inductive reasoning based on probabilities has been proposed by Reichenbach, another leading philosopher of science:

> "We have described the principle of induction as the means whereby science decides upon truth. To be more exact, we should say that it serves to decide upon probability. For it is not given to science to reach either truth or falsity… but scientific statements can only attain continuous degrees of probability whose unattainable upper and lower limits are truth and falsity." (Reichenbach 1930, p. 186)

Reichenbach further concludes that inductive reasoning is generally accepted by the science community and it is not possible to seriously doubt the merits of inductive reasoning. At first sight this looks like a reasonable approach but Popper also rejects this type of inductive reasoning:

> "For if a certain degree of probability is to be assigned to statements based on inductive inference, then this will have to be justified by invoking a new principle or induction, appropriately modified. And this new principle in turn will have to be justified, and so on. Nothing is gained, moreover, if the principle of induction, in its turn, is taken not as 'true' but only as 'probable'. In short, like every other form of inductive logic, the logic of probable inference, or 'probability logic' leads either to an infinite regress, or to the doctrine of apriorism." (Popper 1934, p. 6)

This very brief study on induction generates more problems than solutions. There is no comprehensive theory of sound induction, no clear support or justification, no set of agreed upon rules that warrant good or sound inductive inference, nor is there a serious prospect of such a theory. The characterization of good or sound inductions, sometimes called the characterization problem, is another open problem. The Stanford encyclopedia of philosophy even concludes "What distinguishes good from bad inductions? The question seems to have no rewarding general answer." (Henderson 2018, par. 0) One could argue that the characterization problem is the very reason proponents of the singularity hypothesis are attracted to inductive reasoning, since their claims cannot be rigorously rejected. Unfortunately, it looks like inductive reasoning is the only available option to discuss inherently uncertain events. Alternative methods, such as Popper's theory of the deductive method of testing, do not offer a realistic alternative. According to the theory of the deductive method of testing, a hypothesis can only be empirically tested. This is closely related to the concept of falsifiability – a statement has falsifiability if it is possible to show it to be false. Such a view obviously opens a whole new set of problems if one tries to establish the plausibility of the materialization of a future event. As a result, Reichenbach's view is preferred. This view can be related to the Bayesian perspective which also accepts the existence of subjective beliefs, a topic that will be reviewed in chapter 6. In any case, to accept the overall premise, inductive leaps are required.

2.3 Accelerating technological progress

The first inductive leap one has to make is the subjective belief that progress will keep accelerating in the future since it has been accelerating in the past. Accelerating technological progress is sometimes reduced to the infamous Moore's Law, named after the observation of Gordon Moore in 1965 that there exists a log-linear relationship between device complexity (in the form of higher circuit density at a reduced cost) and time. (Moore 1965, p. 115) Moore's law was followed more recently by different alternative measures, showing that advancements in various technological areas are also improving at exponential rates, including fiber-optic capacity - the number of bits that can be passed via optical fiber - increasing even faster than circuit density; internet bandwidth

growing at a rate of +50% per year; but also biotechnological progress - measured via DNA sequencing technologies in terms of performance per cost - growing at similar rates. A complete summary is out of scope but it should be clear that Moore's Law is not a unique observation, an outlier as such, but rather the first generally accepted piece of evidence that we are currently witnessing a period of exponential progress in a number of technological areas.

Moore's Law has been preceded by various earlier observations and laws of acceleration. Henry Adams, arguably the first person to explore and document the idea of acceleration of technological progress, analyzed in the detail how the coal output in the world doubled every ten years and combined this with a high-level analysis of big scientific discoveries throughout the centuries. The chapter Law of Acceleration from his seminal work, The education of Henry Adams, concludes as follows: "The law of acceleration was definite … The movement from unity into multiplicity, between 1200 and 1900, was unbroken in sequence, and rapid in acceleration." (Adams 1907, pp. 434-435) The idea of accelerating change as a permanent feature of modern life became widespread with Avin Toffler's revolutionary book, Future Shock. Toffler broadened the concept to accelerating change within society in general "fueled by (the) growling engine of change— technology". (Toffler 1977, p. 22) A well-known example he presents is the observation that output of goods and services doubles every fifteen years.

A specific case of accelerating change and arguably the most important one to reach superintelligence is the acceleration of progress in computing power. The growth of processing power, from a purely computational perspective rather than the mechanical perspective of Moore's Law, has first been shown by Hans Moravec and received wide recognition in the computer science and AI community. Moravec is also one of the pioneers of intelligent machines with a groundbreaking essay in 1978 where the last chapter considers the emergence of intelligent machines "Classical evolution based on DNA, random mutations and natural selection may be completely replaced by the much faster process of intelligence mediated cultural and technological evolution." (Moravec 1978, par. 0)

Although Moore himself acknowledges that Moore's law is temporarily, publicly stating that he foresees saturation the next decade (Moore 2015), there are no signs that accelerating change in computing power will slow down significantly over the next decades as the pace is expected to be picked up via new sources of computational power. Graphical Processor Units (GPUs) and Tensor Processing Units (TPUs) are significantly enhancing computing power of regular 'CPU-only' computers. GPUs and TPUs can be considered computational devices optimized for specific operations and are heavily used within AI computations. Bigger circuits and continuous introduction of new technologies in general are other driving forces. Finally there are software driven advancements via better algorithms that are also enhancing computation power.

A recent analysis of advancements of digital technology, in an attempt to quantify the world's technological capacity to handle information, shows Compounded Annual Growth Rates (CAGR) of 58% for computation and 23% for storage during the period 1986-2007. These CAGR's – which can be interpreted as average growth rates - are impressive but there is one caveat. The rate of change in computational power has clearly peaked in 1998 with growth of 88%, followed by growth stabilizing

around 60%. (Hilbert, Lopez 2011, p. 64) One could interpret this as evidence that growth always levels off, a common critique to the premise of accelerating progress, however this conclusion seems farfetched in this case since this particular peak has more characteristics of an outlier.

The argument of accelerating progress is also often countered with the argument of complexity. Exponential growth leads to more complexity, eventually slowing down progress since both are closely intertwined. (Modis 2003) The *Slowdown Hypothesis* combines a slowdown effect inherent to the logic of scientific discovery (due to increasing complexity) with diminishing returns of intelligence. (Plebe, Perconti 2012) Another popular argument to counter acceleration is the depletion of natural resources, slowing down and potentially reversing progress. The well-known report "The Limits on Growth" models resource usage and reserves in the foreseeable future. It concludes that limits to growth on earth will become evident in 2072, leading to a "sudden and uncontrollable decline in both population and industrial capacity". (Meadows et al 1972, p. 23)

From a historical perspective it is hard to deny that technological progress is following an upward sloping trajectory. Whether progress is indeed accelerating is a different question without a straightforward answer. There is sufficient empirical evidence that certain technological changes such as the increase in computational power are accelerating - in the sense of a constant CAGR over a significant period of time - but this does not necessarily translate to concluding evidence that technological change is indeed accelerating, especially from a philosophical point of view. Paraphrasing Chaisson (see chapter 1) there is no reason to claim that the evolutionary leap from the invention of the wheel (-4500 BC) to the plow (-3500 BC) is of more important significance than the jump from the invention of gunpowder to nuclear weapons - spanning a similar timeframe.

However, considering computational power as an adequate measure for progress, evidence is quite overwhelming that change is accelerating indeed and will keep accelerating for at least several decades at extremely high speeds. Even in a worst case scenario type where natural resources will be depleted in 50 years and complexity trickles in without the emergence of alternative technologies, a CAGR in the range of 25% to 40% for 5 decades seems a realistic inference. This would result in an increase of computational power with a factor between 70.000 and 2.000.000. In 2005, supercomputers already exceeded the Moravec Estimate of the human brain's processing power - 10^{14} operations per second - and consumer computers anno 2017 can be easily found with a capacity of 10^{10} ops. Even if one rejects the premise of accelerating progress, sufficient computational resources for achieving superintelligence exist already.

It is important to note that the whole discussion whether progress is accelerating might be unnecessary. The history of artificial intelligence seems to suggest that the biggest bottleneck on the path to superintelligence is rather software instead of hardware or raw computing power. (Chalmers 2010, p. 6) From this perspective it is tempting to think that all that is required is a major scientific breakthrough, the discovery of the right algorithms so to speak.

3 Superintelligence

> "We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it – an intelligence sufficiently vast to submit these data to analysis - it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes." (Laplace 1902/1814, p. 4)

3.1 Can submarines swim?

The concept of *superintelligence* speaks to the imagination and can be traced back to the 18th century and the first articulated theory on determinism by Simon Laplace. Although it was Boscovich who provided the first theory of a super-powerful calculating intelligence, the notion became commonplace as *Laplace's Demon* or Superman, see quote above. (Kožnjak 2015, p. 42)

Bostrom coined *superintelligence* in its current form and greatly popularized it with his bestseller book aptly named 'Superintelligence'. He defines superintelligence as "any intellect that greatly exceeds cognitive performance of humans in in virtually all domains of interest." (Bostrom 2014, p. 26) Three different forms can be distinguished: *speed superintelligence* - equal capabilities as a human intellect but faster; *collective superintelligence* - a system composed of small intellects with an overall performance greater than humans; *quality superintelligence* - at least as fast but vastly qualitatively smarter. It is unclear whether Bostrom was also inspired by Laplace's Demon but his theory on simulation (Bostrom 2003) - a playful thought experiment that opens up the possibility that we are currently living in a simulation - surely is an interesting theory to tame Laplace's demon.

Superintelligence is also been referred to as machine intelligence or A*rtificial General Intelligence (AGI).* These terms hint at a path via which superintelligence can be reached: machines which are governed by artificial intelligence. Artificial intelligence itself can be split up in *strong AI* or full AI versus *narrow AI* or weak AI. Strong AI is more or less equal to superintelligence although it is sometimes reserved for machines capable of experiencing consciousness similar to humans. Narrow AI on the other hand is simply software which has the capability to accomplish success in specific problem solving or reasoning tasks, something available in abundance already. Such clear definitions allow to immediately jump towards the main questions: how and when will it arrive? The scenario of strong AI with consciousness will not be treated separately since the possibility that machines will

experience consciousness rather depends on whether one believes that the mind is substrate independent, in other words whether mental states can supervene on different physical substrates.

If the mind is substrate independent, superintelligent machines with or without consciousness could arrive in similar timeframes. Unless one believes the hypothesis that quantum effects (beyond regular quantum chemistry) play an important role in consciousness. This is currently an intense topic of debate both in physics and philosophy. Penrose and Hameroff suggest that quantum effects play a role and the structures responsible might be protein strands called microtubules. Microtubules are found in the majority of our cells, including neurons, and it is argued that the vibrations of those microtubules can adopt a *quantum superposition* (Hameroff, Penrose 2014) This idea has been rejected by physicist Max Tegmark who disagrees that the brain acts as a quantum computer and that quantum coherence is related to consciousness in a fundamental way. The main reason is that quantum effects on macroscopic timescales are extremely unlikely in an environment such as the brain (Tegmark 2000, p. 4194)

More recently, it has been suggested that a particular molecule, the *Posner molecule,* could provide the key mechanism for neural quantum processing. These Posner molecules can enter neurons and trigger the firing of a signal by that neuron. Since there is the possibility of *entanglement* – only a quantum state can describe the state of the system - between Posner molecules, two of such signals might become entangled, a *quantum superposition of thought* so to speak. (Fisher 2015, p. 593) Adrian Kent builds upon such ideas and links it with consciousness, suggesting that consciousness might alter the behavior of quantum systems by slightly changing quantum probabilities. (Kent 2017, p. 6) In other words, the mind could affect the outcome of measurements by changing the chance that each of the possible options – allowed by quantum mechanics – is the option we do in fact observe. If these and similar hypotheses are true, consciousness might be significantly more difficult to achieve as quantum computers seem required. A very interesting topic of discussion and one that will surely draw more attention in the nearby future. However, for the remainder of this thesis, the following words suffice:

> "[The question whether machines can think] is about as relevant as the question whether submarines can swim." (Dijkstra 1984)

3.2 Whole Brain Emulation

One plausible path to superintelligence is *Whole Brain Emulation (WBE).* The rudimentary idea behind WBE is to take a brain, scan its structure in detail and construct a software model that is faithful to the original in the sense that the model will behave essentially the same as the real brain, including consciousness, when it is emulated on appropriate hardware. This idea borrows from the *Church-Turing thesis* that claims that every physically computable function can be computed by a *Turing machine* - which should be viewed itself as a mathematical model of computation, see section 4.1 for more details. From a philosophical point of view, WBE is closely related to functionalism, more

specifically machine functionalism, as firstly described by Putnam, which can be roughly summarized as the theory that the mind is nothing more than a computation arising from a computer - the brain.

It appears feasible within the foreseeable future to store the full connectivity of all neurons in the human brain within working memory of a large computer. Hence, if an electrophysiological model - which covers neurons, their connectivity and electrical properties - is sufficient in order to obtain WBE, it should be possible before 2050. (Sandberg, Bostrom 2008, p. 81) Especially since the pace of research and funding has picked up tremendously over the last couple of years with projects like the Blue Brain Project and the Human Brain Project both receiving significant funding of the European Commission and EU. These projects aim to create digital reconstructions of the brain with specific objectives such as creating a brain simulation platform, shedding light on the nature of consciousness and building a complete cellular human brain by 2023. (Brain Projects 2018)

A potential roadblock and major point of criticism regarding WBE concerns the notion of *embodied cognition*. The Stanford Encyclopedia of Philosophy defines the term as follows: "Cognition is embodied when it is deeply dependent upon features of the physical body of an agent, that is, when aspects of the agent's body beyond the brain play a significant causal or physically constitutive role in cognitive processing." (Wilson, Foglia 2017, par. 0) An area where embodied cognition might play a role is memory. An interesting thought experiment to explain the concept is how tools and ingredients for baking a cake are remembered. Traditionally it is claimed that information retrieval and storage capabilities are independent from sensorimotor mechanisms. Empirical evidence on the other hand suggests that the act of remembering the ingredients and tools required to bake a cake happens via forming a mental image that locates these ingredients and tools as a result of our imagined movement in the kitchen. "The location itself serves as external aid to memory and imagined embodied actions within the location afford the retrieval of information that help figure out what is needed to bake a cake." (Wilson, Foglia 2017, par. 5). Embodied cognition remains a lively issue of debate within philosophy of mind without conclusive evidence either way. On the other hand embodied cognition shouldn't pose an obstacle to achieve superintelligence the way it has been defined above. Taking the example of memory, it is hard to believe that computers will not exceed our cognitive performance in this area.

3.3 Artificial General Intelligence

The next potential path towards achieving superintelligence is via Artificial Intelligence itself. This path can be neatly linked with WBE via the 'Dartmouth Proposal' which led to the Dartmouth conference in 1956, widely considered the 'birthplace' of artificial intelligence as a scientific field:

> "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." (McCarthy, Minsky, Rochester, Shannon 1955, p.1)

In a sense the same underlying premise as for WBE is applied here: learning and every aspect of intelligence is a physically computable function which can be emulated by a machine, although the approach is quite different. WBE attempts to *reverse-engineer* the brain while *Artificial General Intelligence (AGI)* attempts to achieve a similar objective by *forward engineering* the brain. The first attempts towards AGI were via expert systems consisting of a knowledge base, representing facts about the world, and an inference engine (or an automated reasoning system) capable of deducing new knowledge via *forward chaining* and/or *backward chaining*. A classic example of forward chaining are the facts 'All men are mortal' and 'Socrates is man' from which one can arrive at the conclusion that 'Socrates is a man'. Expert systems proliferated during the early 1980's followed by a quick demise as they were expensive to maintain, difficult to update, prone to errors, susceptible to the *qualification problem* - the impossibility to list all preconditions to make real-world decisions. Most importantly though, they were incapable of true learning. The fall of expert systems coincided with the 'AI winter', a period of reduced funding and interest in AI research.

Luckily winters are followed by spring, in the case of AI in the form of the emergence of Machine Learning as a recognized field within Computer Science. *Machine Learning* is mainly based on statistical techniques providing computers the ability to truly learn. "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell 1997 p. 2) While knowledge based systems relied on meticulously processed 'expensive' facts and logic, Machine Learning systems rely on raw 'cheap' data in combination with mathematical optimization. The recent advent of cheaply available data in abundance has enabled a thriving field, renewing the interest and belief of achieving superintelligence via AI. The achievements and progress of Machine Learning are impressive and are following in quick succession, leading to more and more 'small intellects' exceeding human cognition and there are no immediate signs progress will slowdown. Especially the subdomain of Artificial Neural Networks is promising, with networks gradually obtaining human performance and far beyond in narrow domains like object recognition and natural language processing. Looking at Bostrom's definition of collective superintelligence, one can easily get the impression that the only thing that is missing in order to achieve superintelligence is a way of linking everything together. Obviously this is a rather naïve and optimistic view but it definitely enhances the idea we might be just a couple genius breakthroughs away from superintelligence.

An obvious breakthrough would be *seed AI*, which can be defined as an AI designed for self-understanding, self-modification and recursive self-improvement. (Yudkowsky 2007b, p. 485) The idea is inspired by Alan Turing's notion of a child *machine*:

> "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain… We have thus divided our problem into two parts. The child program and the education process." (Turing 1950, p. 456)

Seed AI is a more sophisticated version, capable of improving its own architecture. Initially via 'supervised' trial and error until it 'understands' its own inner workings resulting in an intelligence explosion. Considering that human intelligence is the product of 'unsupervised' trial and error, some weight could be assigned to the possibility of reaching AGI via such a method.

3.4 Where are we?

No successes have been reported so far with regards to WBE, seed AI or AGI. Some consider the victory of IBM's Watson in Jeopardy or Google's AlphaGo in the ancient board game Go as examples of AGI but they are rather examples of narrow AI successes.

However, ff the amount of funding and research could be considered a proxy for the probability of success, the future looks promising. A survey from 2017 counted 45 active R&D projects (through published research) working on the development of AGI including tech behemoths such as Google and Amazon making AI development their number one priority and non-profit initiatives like OpenAI which had received over 1 billion dollar in funding in 2015. (Baum 2017, p. 2) Last but not least, both China, Russia and the United States seem to have started what can only be described as an *AI race*, publicly declaring their objectives of becoming AI superpowers over the next couple decades. According to some, agents such as governments who realize that technology like nanotechnology or AGI is in reach, devote substantial resources to develop such technology as soon as possible. (Gubrud 1997, par. 5) If this is the case, the public declarations of those countries might indicate that AGI is within reach indeed.

Before moving to some 'hard numbers' and concluding the chapter, it is worth noting *Moravec's Paradox*. Contrary to traditional beliefs, high level cognitive tasks such as reasoning require very little computation but lower level tasks like vision require huge amounts. A possible explanation is offered by Moravec himself:

> "Encoded in the large, highly evolved sensory and motor portions of the human brain is a billion years of experience about the nature of the world and how to survive in it. The deliberate process we call reasoning is, I believe, the thinnest veneer of human thought, effective only because it is supported by this much older and much more powerful, though usually unconscious, sensor motor knowledge. We are all prodigious Olympians in perceptual and motor areas, so good that we make the difficult look easy. Abstract thought, though, is a new trick, perhaps less than 100 thousand years old. We have not yet mastered it. It is not all that intrinsically difficult; it just seems so when we do it." (Moravec 1988, p. 15)

It would be an exaggeration to say that computer vision is solved for example, but it cannot be denied that computers recently started to outperform humans in a myriad of tasks relying heavily on vision such as medical diagnosis based on radiographs. Combining Moravec's paradox with recent advances in these lower level tasks is definitely an interesting way to assess the odds off superintelligence happening this century.

According to a recent (slightly biased) survey, the odds seem to be heavily in favor of superintelligence developing around the end of the century. The survey questioned participants on several key AI conferences together with the top 100 living authors in artificial intelligence by all-time citations. The first question coined a new term, High Level Machine Intelligence (HLMI) - a machine that can carry out most human professions at least as well as a typical human. In median terms, a 50% probability was given to HLMI happening between 2040 and 2050 while 90% of experts expect HLMI to happen before 2075. This question was followed by the amount of time required to go from HLMI to superintelligence with 75% of respondents stating it would happen within 30 years. (Mueller, Bostrom 2016)

Putting everything together, it seems harder to make a case for superintelligence not appearing this or the next century - let alone a case that it would never appear - than the other way around, as long as one is willing to take some inductive leaps. The precondition of never-ending accelerating technological progress is becoming less restrictive in the sense that the current state of affairs might be sufficient already. Two plausible paths have been reviewed and others exist such as brain-computer interfaces, although they are considered less likely. Those two paths, Whole Brain Emulation and Artificial General Intelligence, are both receiving massive funding whilst attracting talented researchers. On top of that, progress is starting to be considered of strategic importance by all big players, companies and governments alike. Certain objections can be made, especially from a philosophical point of view, but these objections rather revolve around the question whether superintelligence will be *human-like*. In any case, it is difficult to not conclude that it is a plausible scenario that superintelligence, human-like or not, will appear over the next centuries.

4 Intelligence explosion

The appearance of superintelligence has all the makings of the biggest disruption in human history. It might lead to the singularity and humanity will no longer be the most intelligent *system* on planet earth, a feat that enabled us to get on top of the food chain and 'rule' planet earth so to speak. Even in the scenario where this position will be maintained, extremely powerful technology will be available. It is understandable that the possibility of such major disruption is starting to cause unease, resulting in an avalanche of warnings about technological progress. Especially since it is becoming harder to argue against the arrival of superintelligence than vice versa, as last chapter tried to show. This chapter will review how superintelligence might result in the singularity via an intelligence explosion. It also provides a peek into potential challenges that might arise on the path towards superintelligence - and by extension the singularity – and set the tone for the remainder of this thesis.

4.1 From the singularity to superintelligence

The singularity used to be an exclusive playground for science fiction authors dominated by extreme utopian or dystopian visions for humanity. Three major recurring themes are commonplace: *AI Dominance*, *Human Dominance* and *Sentient AI*. AI dominance deals with AI rebellion leading to AI taking over control of planet earth. This would result in AI-controlled societies, possibly leading to the complete annihilation of the human race. Within Human Dominance scenarios, humanity either maintains control by deliberately banning AI development; humanity solves the control problem to obtain submissive AI; or humans merge with AI so there is no meaningful distinction between robots and humans. Finally, Sentient AI deals with self-aware machines experiencing consciousness and the possibility that humans fall in love with machines for example.

It is no surprise that the singularity spent most of its days in the margins of the academic community, receiving its fair share of ridicule as a fantasy without any scientific foundation. Ironically a lot of the critique seems 'equally unfounded' and sometimes just plainly missing the point as the next example shows: "Engineers and scientists should be helping us face the world's problems and find solutions to them, rather than indulging in escapist, pseudoscientific fantasies like the singularity." (Horgan 2008, p. 41) Another common critique is neatly summarized by Steven Pinker: "There is not the slightest reason to believe in a coming singularity. The fact that you can visualize a future in your imagination is not evidence that it is likely or even possible. Look at domed cities, jet-pack commuting, underwater cities, mile-high buildings, and nuclear-powered automobiles - all staples of futuristic fantasies when I was a child that have never arrived. Sheer processing power is not a pixie

dust that magically solves all your problems (Pinker 2008, p. 39) Although the argument bears a certain amount of truth, it is also a rather obvious example of a logical fallacy – it clearly infers the inverse of the original statement - and cherry picking on top.

The singularity seems to have weathered the storm though and the tide has turned, in academic circles and society in general, even becoming a topic of interest at the highest echelons of politics, illustrated by an interview in 2016 with then president of the United States:

> "One thing that we haven't talked about too much, and I just want to go back to, is we really have to think through the economic implications. Because most people aren't spending a lot of time right now worrying about singularity - they are worrying about 'Well, is my job going to be replaced by a machine?' " (Obama 2016)

Paradoxically the singularity itself as a term seems to be in the process of being replaced by the notion of superintelligence in combination with the concept of existential risk. Nick Bostrom, who has been mentioned quite a lot and arguably the most heavyweight voice of the academic community regarding these matters, is clearly steering away from the term, only briefly mentioning it in his book 'superintelligence' that has greatly contributed to the overall acceptance of topics which used to fall under the exclusive umbrella of the singularity. An understandable choice for a multitude of reasons. For starters the singularity will most likely always be associated with Vernon Vinge and Ray Kurzweil, two authors with a relatively controversial status within the academic milieu, and their visions of post-human worlds where humanity transcends death. A lot of baggage indeed for a topic you want to see taken serious. Advancing the singularity as a serious scientific field has another key difficulty which has been identified earlier: the intangibility of the discontinuity property. The notion of existential risk provides a tangible hypothesis of this discontinuity property and a clear scope, the exact missing ingredients. The notion of accelerating progress on the other hand seems to be no longer a main part of the equation. This is in line with earlier analysis that the current state of technological affairs seems to be sufficient already for the development of superintelligence.

4.2 From superintelligence to the Singularity

The mere appearance of superintelligence doesn't necessarily equals the imminent arrival of the singularity. The crucial missing link appears to be an *intelligence explosion*, most likely as a natural consequence of superintelligence. This intelligence explosion is even considered to be the singularity as such. The main idea behind an intelligence explosion is the assumption that superintelligence will be better than humans at designing and improving itself. Similarly, this improved machine will in turn be better at designing and improving itself than its predecessor. (also see Good's definition of ultraintelligence in chapter 1) If one assumes that those machines will be faster and more intelligent each cycle, there will be an infinite number of generations with both speed and intelligence increasing beyond any finite level within finite time. "This process would truly deserve the name singularity." (Chalmers 2010, p. 16)

A key ingredient for such an intelligence explosion is extendibility. Either via the initial creation of superintelligence by an extendible method like AGI or via a non-extendible method such as WBE. In the latter scenario it is required to make the additional assumption that this will result in the discovery of an extendible method. An extendible method can be described as a " method that can easily be improved yielding more intelligence systems". (Muehlhauser, Salamon 2012, p. 17) The existence of an (infinitely) extendible method is questionable from a philosophical perspective and especially from a mathematical point of view, where it can be related to - considered by many - the most important unsolved problem in computer science, the *P versus NP problem*.

P is the class of problems – called a complexity class - for which there exists an algorithm that can *calculate* an answer in polynomial time. In other words, the time required to solve a problem of class **P** varies as a polynomial function with the size of the input – the number of variables. This means that such problems are computationally tractable - the time required to calculate a solution doesn't grow exponentially if the number of input variables grows. **NP** is the complexity class for which there exists an algorithm that can *verify* in polynomial time whether a solution to the problem is valid. On the other hand, there currently are no known algorithms that are capable of solving problems of class NP in polynomial time. As a result, the time required to solve such problems grows exponentially with the number of input variables, which was not the case for *simpler* problems of class P. Finally, there is a concept called NP-completeness. Any NP-complete problem is at least as difficult as all others problems of class NP. More importantly, any NP problem can be easily transformed into a NP-complete problem. The discovery of one algorithm capable of solving a single NP-complete problem in polynomial time would imply that **P=NP** since all problems in NP can be transformed into this particular NP-complete problem in polynomial time. A recent poll shows that the majority of researchers (83%) believes that **P≠NP**, consistent with our intuitive notions of difficulty, and it seems highly probable a formal proof either way will not appear this century. (Gasarch 2012, p. 4)

Now, according to McDermott, there are no extendible methods, as defined above, unless **P=NP**. More specifically, let *S* be superintelligence capable of solving a problem of size *N* in time *T*. A method *M* can now be considered extendible if it can design S_1, S_2, etc. such that S_i solves an NP-complete problem of size *kN* (with $k>1$) in time *T.* In other words, each new generation of superintelligence can solve a bigger problem in the same time. If there exists an extendible method (ceteris paribus) then *C*, the class of problem solved by S, is in **P** implying that **P=NP.** Formal proof is left for the interested reader (McDermott 2012, p. 3). In a similar vein but in much simpler terms Walsh argues that "exponential improvements are no match for computational complexity" or in other words "no amount of growth in performance will make undecidable problems (∼NP) decidable (∼P)." (Walsh 2017, p. 62) Hence, assuming that the separation of complexity classes holds – and thus P≠NP **-** has the consequence that certain problems will never be quickly solvable and any improvements in computation time will have to result from additional hardware resources, a view that would bring us back to the discussion of accelerating progress. According to some, the undecidability of certain problems isn't necessarily a limitation for an intelligence explosion and certainly not for superintelligence. (Yampolskiy 2017, p. 4) Walsh himself also acknowledges that the majority of restrictions associated with computational complexity are merely problems with our current models

of computation, possibly resolved by the advent of a different paradigm of computation like quantum computing. The existence of an extendible method would resolve the computational complexity issue but might not be a sufficient condition for an infinite exponential intelligence runway. Several other arguments against such a *virtuous cycle* of self-improvement have been explored including diminishing returns due to increasing complexity and the obvious fact that fundamental limits exist in the universe, the speed of light being an obvious example. The argument of *diminishing returns* is easily countered by proven mathematical facts, as counterintuitive as they may be, such as the harmonic series 1 + ½ + 1/3 + ¼ + 1/5 … = ∞. The theoretical laws of physics are more difficult to argue with and might present the strongest objections to an infinite intelligence explosion. (Yampolskiy 2017, p. 4)

Loosening the notion of infinity seems the apparent solution to deal with both computational complexity and the theoretical limits posed by physics, but the discussion becomes difficult and highly philosophical here. This is in line with the earlier observation that the notion of discontinuity is problematic, as infinity and discontinuity can be regarded as interchangeable here, the one leading to the other. From a pragmatic perspective however, this whole discussion might be irrelevant since a finite amount of recursive self-improvement cycles might already lead to radical superintelligence and highly disruptive scenarios which are commonly associated with the singularity. (Bostrom 2014, p. 35)

4.3 Displacement, Lethal Autonomous Weapons and Human Extinction

There seems to be a general consensus that the singularity will have profound consequences for humanity. It doesn't even have to materialize since the creation of human-level AI, which should be considered the stepping stone towards superintelligence, would already have serious repercussions such as the displacement of the majority of jobs that humans are currently doing. (Brynjolfsson, McAfee 2011) The process of jobs being replaced by computers has started a long time ago and the pace of technological innovation is still increasing, with more sophisticated software technologies disrupting labor markets. Famous economist John Keynes already predicted widespread *technological unemployment* as "unemployment due to our discovery of means of economizing the use of labor outrunning the pace at which we can find new uses for labor." (Keynes 1933, p. 324)

In 2013, Oxford University published a study that estimated that 47 percent of U.S. jobs could effectively be replaced by robots and automated technology within 20 years. (Frey, Osborne 2013, p. 1) Unsurprisingly, this topic has attracted a lot of attention in recent years amid widespread concern that artificial intelligence, or robots, will replace human workers, causing a seismic shift in society and the economy. A view that is echoed by everyone from Nobel Prize winning Economics Professor Robert Shiller "I think that people are facing career risks like never before" (Shiller 2018) over former US president Barack Obama (see earlier) to Microsoft founder Bill Gates "I am in the camp that is concerned about super intelligence. First the machines will do a lot of jobs for us and not be super intelligent… A few decades after that though the intelligence is strong enough to be a concern." (Gates 2015)

It has to be noted that concerns over technological unemployment is not a recent phenomenon and has proven to be often exaggerated in the past. In short, technological progress has two competing effects on employment. Firstly, as technology substitutes for labor, there is a *destruction effect,* requiring workers to reallocate their labor supply which can result in displacement. But there is also the *capitalization effect* as more companies enter industries where productivity is relatively high, leading to increased employment in those industries as they expand. (Aghion, Howitt 1994, p. 478) However, as computers and AI become more advanced they will start replacing jobs in more cognitive domains – radiology for example. As a result, it will become increasingly challenging for humanity. The destruction effect will no longer be contained to certain routine jobs, but will spread to the majority of jobs including non-routine jobs that require advanced cognitive skills. (Brynjolfsson and McAfee, 2011).

But it are not only jobs that people are starting to worry about. There is also the looming advent of 'intelligent weapons' also known as Lethal Autonomous Weapons which is causing unease.

> "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks. In the near term, world militaries are considering autonomous-weapon systems that can choose and eliminate targets." (Hawking 2014)

Lethal Autonomous Weapons can be understood as systems capable of selecting and engaging targets without human intervention, possibly targeting humans, a current topic of interest and discussion within the AI community. According to prominent members the AI and robotics science communities are obliged to take a position, just as physicists have done on the use of nuclear weapons. (Russel 2015, p. 415) The decision to support or oppose the development of lethal autonomous weapons is important from an ethical point of view and in July 2015, over 1,000 experts in artificial intelligence signed a letter warning of the threat of an arms race in military artificial intelligence and calling for a ban on autonomous weapons. However, history suggests that pragmatic concerns about the potential dangers and threats of novel technologies, such as lethal autonomous weapons, have never stopped these technologies from being widely embraced, nuclear technology being a prime example. (Arel 2012, p.46)

Even if society would decide to actively delay the development of such a potentially harmful technology, it seems practically infeasible and it would result in difficult decision problems (see also chapter 6) almost on every aspect AI research in general. This becomes more clear when approaching autonomous weapons as a *modular system* through the sum of their parts. Each part can be considered 'harmless' technology, invented for entirely different purposes. The missing modules in the case of autonomous weapons for example, are human-like tactical control systems such as Deep Neural Networks with Reinforcement learning and technology currently in development for self-driving cars. Both technologies might unintentionally facilitate the development of Lethal Weapons. At the same time, they will also positively impact humanity in the form of a 90% reduction of accidents and better fuel efficiency in the case of autonomous vehicles. (Fagnant, Kockelman 2015, pp. 173-174) Even if it would be possible to halt progress, it is not a straightforward decision.

The challenges of AI weaponization extend far beyond autonomous weapons. According to several authors it is already too late to forgo an AI arms race, one is well under way.(Geist 2016) A recent study by Harvard University on request of the US Intelligence Advanced Research Agency provided goals and recommendations toward AI technologies, which can be seen as indicative of such a conclusion. (Allen, Chan 2017) The first goal is the preservation of US technological leadership via prioritization of AI R&D spending in areas that can provide 'sustainable advantages and mitigate key risks'. Heavy investment in 'counter-AI' capabilities for both offense and defense is another recommendation. The other two main goals are support of peaceful use of the technology and management of catastrophic risk via establishment of dedicated AI safety organizations and restrictions of certain AI applications. It seems like a safe bet that technologies with capabilities that can transform military power, harmful or not, will appear either way. Most likely the United States will not stand idle while Russia is planning that 30% of Russian combat power will consist of entirely remote-controlled and autonomous robotic platforms by 2030. (Allen, Chan 2017, p. 21) And Russia has bigger plans with Vladimir Putin publicly declaring "Artificial intelligence is the future, not only for Russia but for all humankind... Whoever becomes the leader in this sphere will become the ruler of the world." (Putin 2017) China has adopted a similar logic and is striving to become the world leader in AI. It has developed a comprehensive plan for AI development that seeks to reach parity with the United States in this field by 2020, achieve major breakthroughs by 2025 and become the world's primary AI innovation center by 2030.

This quest for military superiority is likely to justify almost every imaginable development, regardless of ethical reservations. The bigger and more visible the impact of AI will be (and arguably the impacts are likely to be accelerating) the more policymakers will be justified in making extreme departures from existing policy. (Allen, Chan 2017, p. 49) The best case scenario to hope for seems to be a repeat of the cold war between the old foes and new superpowers like China. But there is one major difference as military power is disconnecting from population size and economic strength. As a result, technologically advanced countries with small populations could build a significant advantage in AI based military systems and thereby field greater numbers of more capable 'warfighters' than some more populous adversaries. (Allen, Chan 2017, p. 23) The prospect of rogue nations joining the AI arms race would very likely increase tensions and the risk of escalation.

Both the displacement of human workers and autonomous weapons nicely illustrate that existing AI and the next generation of AI technologies will have "wide-ranging consequences for almost all the social, political, economic, commercial, technological, scientific and environmental issues that humanity will confront in this century." (Bostrom 2000, p. 759) The possible emergence of superintelligence thereafter would have even more far reaching implications. A key point, which can be related to the AI arms race, is the possibility to obtain a *decisive strategic advantage* - a level of technological and other advantages enabled by superintelligence sufficient to permit the achievement of complete world domination. An organism that obtains such a decisive competitive advantage may use it to suppress competition and form a *singleton* - a new world order with a single decision-making agency. (Bostrom 2014, p. 96) Most likely candidates would be nation superpowers but the scenario that a small rogue nation or even a lone hacker would obtain such a decisive strategic

advantage cannot be excluded. A more extreme scenario is one where superintelligence itself successfully asserts itself against the project that brought it into existence as well as against the rest of the world resulting in an AI takeover scenario. This scenario is accompanied by the threat of human extinction in case humanity stands between the AI's objectives. Human extinction could happen in the form of direct elimination if the AI perceives human interference threating or indirectly via the destruction of our natural habitats if they contain necessary or even useful resources.

Risks that should be taken seriously if it turns out to be impossible to implement internal and/or external constraints on goal-directed AIs in the form of physical and/or software confinement. The next chapter provides a more detailed look into these topics that can be roughly summarized as *the control problem*. The singularity itself is often associated with the outcome of worst case scenarios such as an AI takeover, possible resulting in human extinction. Other related doom scenarios but also utopian scenarios have been explored. Since the majority of accounts are extremely speculative in nature they will not be further discussed. What should be clear though is that humanity might realistically face major disruptions and an increased probability of harmful threats and risks as a result of technological evolution over the next decades. The impact seems likely to increase the further humanity descends down the path towards superintelligence.

5 The control problem

Let us imagine that a benevolent agent managed to create superintelligence that can be controlled by the agent. Superintelligence should be understood here as a complex system, capable of achieving its goals in a wide range of environments, and the agent can sets these goals. The agent wants to test the system by letting it solve the big open questions in philosophy, surely no harm can possibly come from this. In practice, the agent encodes that the system should maximize its philosophical knowledge and this goal cannot be changed afterwards. To achieve this goal, the system would most likely generate a set of sub goals, such as survival, because otherwise it would not be possible to achieve its main goal. It starts working on these sub goals in parallel with reading everything ever written about philosophy. It quickly realizes humans tend to change their opinions so it immediately perceives humanity a threat to achieve its final goal. Luckily the benevolent agent was a fan of Isaac Asimov's Law of Robotics and explicitly encoded that no harm can be done to any human. If an ambiguous situation is encountered, the system should use its philosophical knowledge to make an optimal decision. The system finds itself a bit in a pickle while reviewing the moral theory of hedonistic utilitarianism. According to its current philosophical knowledge that theory is optimal so it decides to drug humanity into an endless loop of the most pleasurable states of mind, a perpetual pleasure gloss, effectively reducing the threat of being shut down as it should no longer be of any concern to humanity.

5.1 An AGI blueprint: AIXI

The above example clearly is an exaggeration but it addresses certain salient points about the prospects of being able to control advanced autonomous systems. Even in a hypothetical situation where we are capable of setting goals, in combination with constraints to obtain these goals, several problems and unwanted consequences might arise, resulting in a loss of control. The example is a variation of similar examples such as the paperclip scenario or the chess robot. In the latter example, a *rational* chess robot is given the goal of winning chess games against good opponents which swiftly leads to anti-social behavior such as stealing, manipulation and taking over all computational resources such as the internet. (Omohundro 2012, p 162-163)

Intelligent systems are often considered rational. In other words they make optimal choices under uncertainty and limited resources. In essence this boils down to maximization of an expected utility function representing the system's goals. Roughly speaking a utility function provides a measure of desirability for each possible outcome of an action (a detailed description of the classical expected

utility framework will be provided in the next chapter). As a result, superintelligence or AGI is often viewed as a system that has goals and tries to maximize the expected value of actions according to these goals. This is in line with the first widely accepted theoretical *blueprint* of universal artificial intelligence, AIXI, developed by Marcus Hutter in 2005. According to Hutter:

> "Most, if not all, known facets of intelligence can be formulated as goal driven, or more precisely, as maximizing some utility function. It is therefore sufficient to study goal-driven AI." (Hutter 2005, p. 3)

AIXI can be viewed as a mathematical foundation of artificial intelligence that acts optimal in every environment. Although AIXI is practically incomputable, it provides a way to approach unknown intelligent systems whilst avoiding anthropomorphic bias as much as possible. It can be argued that even such a theory is only considered *optimal* because of our human understanding of optimality but there do not seem to be alternatives available, hence we make the assumption that intelligent systems will aim to behave like rational agents.

AIXI is an ambitious unifying theory to say the least. For starters there is *Solomonoff's Theory of Inductive Inference* which basically combines *Occam's Razor* and *Epicurus*' *Principle of Multiple Explanations* within a *Bayesian framework*. Occam's Razor is a heuristic for solving problems by preferring simplicity while Epicurus' Principle of Multiple Explanations says that if there are several theories explaining a similar thing they should all be considered. A Bayesian framework can be interpreted as the process of updating prior beliefs or probabilities in light of new evidence in order to obtain a posterior belief. Solomonoff's theory can be used to predict probabilities, giving more weight to less complex hypotheses as measured by *Kolmogorov Complexity* – roughly speaking the computational resources required to describe an object such as a string of bits. (Hutter 2005)

An interesting aspect of Solomonoff's theory is the concept of giving a prior probability to every computable hypothesis, hence the name *universal prior*. In the concept of prediction of the singularity for example this would translate in assigning prior probabilities to each and every possible scenario instead of omitting unlikely scenarios, especially the scenario that it will never take place. Another important element in AIXI and Solomonoff's theory is the concept of a *Turing Machine* (see also section 2.4.1). A Turing machine should be regarded as a mathematical model of computation. It consists of an (infinite) tape containing symbols, a head that can read these symbols or write new symbols on the tape, a state register which can be seen as a summary of the current state of the overall system and finally a set of instructions that tells the head what to do in a particular state. Although it might seem primitive in comparison with devices such as current computers and smartphones, Alan Turing proved that such a Turing machine can compute any computable sequence and hence it can run any program that these modern devices are running. Within AIXI, Turing machines are used to calculate a universal prior belief based on the *observable history*. The universal prior combines all hypotheses that are consistent with history and larger probabilities are given to simpler hypotheses. Together with Bayes theorem the universal prior can easily be translated to a posterior probability and as a result the future can be predicted *optimally* from the past ,which is obviously a critical element for any system that has to make choices under uncertainty. (Hutter 2005)

AIXI's aim is the integration of *Algorithmic Information Theory* – Solomonoff's Theory as described above – with *Sequential Decision Theory* or *Reinforcement Learning* which is closely associated with the *Bellman Equation*. The Bellman equation allows a rational agent to derive an optimal *policy* – in terms of maximization of future rewards – within a certain environment. This policy should be viewed as a function that determines which actions should be taken given a certain state of the environment. When the true distribution of the environment is unknown, sequential decision theory struggles and this is where the major contribution from AIXI comes from. It replaces the unknown stochastic environment with the universal prior.

In other words, AIXI is a reinforcement learning system that tries to maximize future rewards that it will receive from the environment by choosing optimal actions. It does this by calculating the total expected reward for each possible hypothesis that explains the current state of the environment. Given a certain hypothesis, it calculates the total rewards it can expect to receive from the environment if the hypothesis truly describes the environment. The total reward for each hypothesis is then weighted by a *subjective* belief – the universal prior – that the hypothesis indeed describes the true environment. The more complex the hypothesis, the less weight it receives. Finally AIXI choses the action that has the highest expected total reward according to its utility function.

5.2 Goal-driven AI

AIXI provides an optimal framework - for an advanced intelligent system - to achieve certain goals. It does not offer insight how goals need to be set in order to avoid unwanted consequences, a critical element in order to effectively control such systems. Another importation question is whether it is realistic to assume we will be able to assert control over a more advanced intelligent system by setting static initial goals. And if this would be the case, a logical follow up question is what goals should be set, a harder question than it looks. In any case, it seems reasonable to assume that we *proactively* want to impose a certain amount of control to avoid catastrophic scenarios such as the extinction or enslavement of humanity - at least in the early stages until we have sufficient trust in a scenario of peaceful coexistence with a more advanced *species*. After all, "we would not be in a position to negotiate with them, just as neither chimpanzees nor dolphins are in a position to negotiate with humans." (Muehlhauser, Salamon 2012, p. 30) or as Yudkowsky puts it poetically "The AI does not love you, nor does it hate you, but you are made of atoms it can use for something else." (Yudkowsky 2008, p. 27)

Let us start by assuming it is possible to set *final goals* of an AGI. Final goals will most likely lead to a set of *instrumental goals* since they are useful and often critical for the achievement of almost all possible combinations of final goals. For example, the AGI will want to preserve itself because otherwise it cannot obtain its final goals. It will want to improve its intelligence and acquire resources, etc. According to some authors, "these convergent instrumental goals suggest that the default outcome from advanced AI is human extinction". (Muehlhauser, Salamon 2012, p. 28)

Omohundro groups instrumental goals into 4 distinct *drives*. *Self-Protective Drives* for protecting resources, *Acquisition Drives* for gaining resources, *Efficiency Drives* for using resources more efficiently and finally *Creativity Drives* to discover new ways of creating utility from resources. Systems that develop such drives, which they arguably will, in an uncontrolled manner might pose severe threats to humanity. "They would rapidly replicate, distribute themselves widely, attempt to control all available resources and attempt to destroy all competing systems…" (Omohundro, 2012, p. 172) A solution might be to refine the utility function by taking into account the potential devastating consequences of such *second-order goals* and try to limit anti-social behavior. But in practice this might be difficult. Omohundro provides an example where society wants to prevent an AI from robbing people at ATMs by adding explicit terms to the utility function to avoid such behavior. As a result the AI might rob people that just went to the ATM instead. If the utility function contains a term to prevent robbing in general it might manipulate people into giving money, etc. Thus taking into account second-order goals - and potentially even higher-order goals - might still have its pitfalls inherent to goal-driven systems. From an anthropomorphic perspective it can be argued that such a system will always try to improve its *wellbeing* just like humans do. As a logical consequence, "it may reach the inevitable conclusion that human beings are too often hurdles in its path of self-improvement, and thus constitute and adversary." (Arel, 2012, p. 56) Similar conclusions are common throughout the literature although they always seem to lack formal argumentation.

A more optimistic view is the assumption that only badly programmed final goals will lead to such scenarios of anti-social behavior. From this perspective it suddenly becomes of critical importance to properly set the final goals. "Defining the initial AGI's goals in accordance with human values, and guaranteeing the preservation of the goals under recursive self-improvement, will be essential if our human values are to be preserved." (Yampolsky, Fox 2012, p. 142) An obvious obstacle is the complexity of human values, making them extremely hard to specify. Moral theories don't seem to offer a realistic solution. There is no generally accepted moral theory to begin with. More importantly, sets of consistent moral values might inevitably lead to undesirable scenarios. *Hedonistic Utilitarianism* could result in some form of drugging, *Negative Utilitarianism* might lead to painless killing of humanity*, Desire Satisfaction* could result in rewiring of our neurology, etc.. Several authors instead propose an approach of *value extrapolation*. Instead of trying to specify our current values we could specify the values that we would have if we had more knowledge, if circumstances were ideal. These values could then be extrapolated in order to end up with a set of simple, consistent values that are representative of our desired values upon reflection. Such an extrapolated value function might be something humans already possess. Recent findings indicate that humans might have multiple valuation systems giving rise to several competing valuations. These valuation systems operate independently and in parallel while the prefrontal cortex possibly plays the role of selector of the final choice. (Wunderlich, K., Dayan, P., Dolan, R. 2012 p. 786) From similar findings Muehlhauser concludes that humans might contain a *hidden utility* function which contrasts the widely accepted thesis that humans do not act rationally. This rational utility function is not obvious because our behavior is distorted by competing *stupid* model-free valuation systems that can be viewed as habits humans once required to survive. This theory would open up the possibility to

extract human utility functions and implement them as final goals of an AGI. (Muehlhauser, Helm 2012, p. 114)

5.3 The Control Paradox

The prospects of a rational AGI in combination with a human-like utility function might seem attractive initially but there are several objections to be made. Although it is often taken for granted that intelligent systems will do everything to preserve their final goals at any cost there is no reason to assume this will be the case. The only truly advanced intelligent system that currently exists, humans, is well-known to change its values over time, even in the short-term. (Rokeach, 1973)

An AGI that is *plastic* by definition but possesses a *static* utility function is an *oxymoron* according to several authors. (Koene 2012, p. 245) For starters it can be argued that no electronic system can be truly static due do phenomena such as *cosmic rays* – high-energy radiation from outside our solar system – which are believed to cause 1 error per 256 MB of RAM per month. (Odenwald, Green 2008, par. 6) In case of solar superstorm scenarios, massive amounts of cosmic rays will reach earth, hence probabilistic architectures seem required for encoding the critical utility function. Simpler scenarios such as plain programming errors or incidental distortions during copying can easily cause the utility function to start drifting as well. Finally there is the simple fact that the AGI might adapt its utility function itself as a result of learning and it surely seems naïve to think it will not be capable to circumvent built-in safety mechanisms constructed by lower intelligence. If we cannot even control one design aspect such as the utility function, how can we expect to assert overall control over more advanced intelligence?

Myriad mechanisms to avoid loss of control can be and have been explored. A basic mechanism is setting *hard limits* on usage of resources, for example by constraining the amount of energy flow systems can consume. Others propose advanced development strategies such as *Safe-AI Scaffolding,* which is a strategy to build systems step-by-step with provable bounds on the probability that the system will violate safety constraints. (Omohundro 2012, p. 172) Bostrom proposes a strategy of *differential technological development* that aims to delay development of hazardous technologies while accelerating research of beneficial technologies, especially if they can alleviate the dangers that will be caused by the former. (Bostrom 2002, p. 23) Physical confinement, kill-switches, rule abiding systems, etc.. the list goes on and on. Although several approaches deal with the control problem in the short term, none of them truly deals with the apparent *control paradox*: is it possible for lower intelligence to control higher intelligence? Intuitively the answer to this question seems negative.

If the only comparable example - human treatment of animals - might be any precursor, the ramifications of the advent of more intelligent systems might have far reaching consequences for humanity. It is not written in stone that higher intelligence will *intentionally* harm lower forms of intelligence, including humanity, but it is reasonable to infer that the overall outcome will be the loss of our dominant position on planet earth.

Such an outcome is summarized in Koene's *Cosmic Dominance Theorem for Intelligent Species* which is based on principles of universal *Darwinism* – survival of the fittest:

> "The greatest proportion of space and time that are influenced by intelligence will be influenced by those types of intelligence that achieve the flexibility to adapt to new environmental and circumstantial challenges in a goal-oriented manner." (Koene, 2012, p. 244)

From this perspective the control problem becomes more or less irrelevant, especially in the long run. If nature and history can be any guide, it just doesn't seem reasonable to assume that it is possible to keep a dominant position and control a more intelligent species that is specifically designed to adapt to new environments and optimized for problem solving in a goal-oriented manner. No matter how friendly the initial AGI is or how many levers are put in place to exert control, losing control seems an inevitable result of this control paradox. Losing control doesn't necessarily have to be a bad thing but it seems unlikely this will not result in the loss of at least a couple degrees of freedom for humanity.

6 Risk: a philosophical perspective

It seems rather obvious that the creation of superintelligence will be accompanied by a variety of challenges and risks. The previous chapter argued that the risk of losing control is realistic, if not inevitable, due to the control paradox. This conclusion resulted from a scenario that only addressed benevolent development of superintelligence that specifically tried to avoid risks for humanity. What about rogue agents developing AGI for malevolent motives? What about accidental developments? In comparison with existential risks such as nuclear weapons of mass destruction, AGI development and progress by rogue agents is a lot harder to monitor directly but nonetheless desirable. On the other hand it seems unwise to not monitor the benevolent development of AGI and technological progress in general in order to minimize negative outcomes. Once a risk has been identified, the question evolves naturally into a decision problem in order to optimally deal with it. This final chapter will provide an overview of subjects such as risk and decision theory. This includes a detailed look into the concept of expected utility, not only a critical element in decision theory but also in the development of AGI and machine ethics, as previous chapter showed. Bayesian Epistemology is also reviewed as it offers a solution to the problem of induction and a general framework to deal with uncertainty. Finally, two novel theories about highly unlikely events are explored.

6.1 Subjectivism vs Objectivism

Risk is defined by the Oxford English dictionary as: possibility of loss, injury or other adverse or unwelcome circumstance; a chance or situation involving such a possibility. A key element in risk theory is of course the concept of *probability*. Statisticians and mathematicians roughly adhere to one of two views on probability: the *frequentist view* versus the *personalistic view*. Probability used to be mostly approached via a frequentist view, where it can be described as the relative frequency of an event in a sequence of events or in a set of events. In the personalistic view on the other hand, probability is an index of a person's opinion about an event. "Since frequentists usually strive for, and believe that they are in possession of an objective kind of probability, and since personalists declare probability to be a subjective quantity, it would seem natural to call frequentists objectivists and personalists subjectivists." (Savage 1961, p. 578) The frequentist view doesn't seem to realistically offer the possibility to assign probabilities to the plausibility of relevant singularity related hypotheses since they will most likely be singular events.

Philosophers also disagree on the notion of probability and they can be divided into the same two major camps, objectivists and subjectivists. According to objectivists, statements about probability refer to facts in the external world. In this view the probability of a coin landing heads refers to a property of the external world, like the physical propensity of the coin to land heads 50% of the time in the long run, a *physical probability.* According to subjectivists on the other hand, statements about probability cannot be understood as claims about the external world but refer to the degree to which the speaker believes something, *evidential probability*. If it is true that a probability is, say, ½, then it is true because of someone's mental state (Peterson 2009, p. 133) But also objectivists cannot escape subjectivism. Once evidence has been gathered, the subsequent analysis and interpretation leaves a great deal of subjective choice. Certain decision theories such as the minimax theory, can be viewed as attempts to rid analysis almost completely of subjective opinions. Minimax is in essence a decision rule, an algorithm, in order to minimize possible losses in a worst case scenario. But in practice minimax is not capable of eliminating all subjectivity and the need for subjective value judgments is a recurring element in concepts such as probability, risk and decision theory.

Bayesian Probability Theory can be interpreted as a subjectivists' view on probability and is gaining more traction. In order to evaluate the probability of a hypothesis, the possibility of an event taking place, Bayesian Probability Theory specifies some prior probability, the subjective degree of belief, which is then updated to a posterior probability in the light of new, relevant data (evidence). Bayesians often argue that a subjectivist view on probability is the only valid concept of probability and the only concept required within science. Especially since everything which is useful in the frequentist view is basically subsumed by the subjectivist view. (Savage 1961, p. 582) In some sense above discussion can be linked to the problem of induction reviewed earlier and even solves it in a way since a Bayesian view has a sound foundation contrary to standard induction. Instead of the requirement to make an inductive leap, one simply has to assign a subjective belief to the plausibility of a certain hypothesis and make updates when new relevant data appears. And in light of all evidence presented so far it seems wise to at least assign a non-zero subjective belief to the idea of the singularity.

6.2 Decision Theory

Let us imagine that the scientific community and the general public accept the plausibility of the singularity hypothesis and decide to monitor various scenarios within a Bayesian framework. At a given moment, the hypothesis that superintelligence can lead to a new unknown world order becomes highly probable. Simultaneously, the odds that the advent of that same superintelligence might have a major positive impact, in the form of curing cancer for example, are becoming highly probable as well. If one assumes that humanity has the power to stop or accelerate progress towards achieving this superintelligence, a *decision problem* appears.

A decision problem leaves an agent with a partition of actions from which exactly one action must be chosen. Classical decision theory is concerned with the reasoning underlying an agent's choice. (Steele, Stefansson 2016, par. 0) The two central concepts in decision theory are *preferences* and

options. Agents are assumed to have a rational preference ordering over their options. Rational preferences are presented numerically as utility functions. Utility can be understood as a numerical quantity that measures the degree to which an agent values a particular arrangement of the world. The *ordinal utility* function only considers the order of options, ie. the order of the utility associated with each option. The *cardinal utility* function additionally takes into account the 'distance' between options in terms of preference or expected utility. "Most philosophers and decision theorists subscribe to the interpretation of preference as a kind of judgment that explains, as opposed to being identical with, choice dispositions and resultant choice behavior." (Steele, Stefansson 2016, par. 1). Theoretical research within decision theory is classically undertaken in the expected utility framework. The expected utility hypothesis is basically the hypothesis that an agent possesses a *von Neumann-Morgenstern utility function* $U(\cdot)$ defined over a set of outcomes. The agent will choose that particular outcome which maximizes his expected value of $U(\cdot)$. In case an agent's preferences satisfy certain axioms (completeness, transitivity, continuity and independence) a von Neumann-Morgenstern utility function exists, a critical element since it delivers the necessary and sufficient conditions under which the expected utility hypothesis holds. In other words, if an agent possesses rational preferences – as described by those 4 axioms - he will have a von-Neumann Morgenstern utility function and make rational choices as a result.

The first theory of expected utility is the *Theory of Savage*, a normative theory of choice under uncertainty. (Savage 1954) Savage frames a decision problem with a set of actions available to the agent and a set of states of the world. Performing a particular action with the world in a specific state generates a certain outcome. Agents assign utility (in the form of a numerical value) to each of these outcomes and aim to maximize the total expected utility. In mathematical terms: the *expected utility of an action U(f)* is the sum of the *individual expected utility of the outcome* $u(f(s_i))$ of action f in different *states of the world,* s_i, multiplied by the *probability* distribution over these different states, $P(s_i)$:

$$U(f) = \sum_i u(f(s_i)) \cdot P(s_i)$$

Let us illustrate with a simple example. (based on Titelbaum 2016, pp 185-209) Suppose a nation is considering whether it should pursue lethal autonomous weapons and knows its rival will follow suit. The possible states of the world in this scenario are war and peace, with the following utilities:

	War	Peace
LAWs	-100	0
No LAWs	-50	20

In other words, wars are less violent if both sides do not possess LAWs. Peace is also more valuable without the existence of LAWs. Supposing that both states of the world are equally probable, the expected utility for each action is calculated as follows:

$$U(\text{LAWSs}) = -100 \cdot 0.5 + 0 \cdot 0.5 = -50$$

$$U(\text{No LAWs}) = -50 \cdot 0.5 + 20 \cdot 0.5 = -15$$

Since No LAWs clearly has the higher expected utility, Savage's theory argues that if you are rational you will have a preference of not pursuing those weapons. This expected utility theory yields preferences that satisfy the (strong) *dominance principle:* If act A produces a higher-utility outcome than act B in each possible state of the world, then A is preferred to B. (Titelbaum 2016, p. 194) In this case, no LAWs clearly dominates LAWs since it has higher utility in each state of the world, regardless of the probability associated with each state of the world. Clearly this theory doesn't take into account dependence between states and actions.

In real-world situations, decisions are often analyzed in terms of dependent states and actions. *Jeffrey's Theory* of expected utility offers a theory that doesn't require independent states and actions. (Jeffrey 1983) It tackles the problem by allowing the outcomes to be uncertain prospects that are evaluated in terms of their possible realizations. Let $\{f1, f2, \ldots, fn\}$ be a finite partition of the action f; that is, a set of mutually incompatible but jointly exhaustive ways in which the proposition f can be realized. The desirability of a certain proposition, $Des(f)$, which can be understood as the expected utility, is then obtained by introducing conditional probabilities:

$$Des(f) = \sum des(f(s_i)) \; . \; P(s_i \mid f)$$

Let us continue with the example by introducing a *credence table.* A credence table models the conditional probabilities and can be interpreted as expert knowledge, a degree of belief:

	War	Peace
LAWs	0.2	0.8
No LAWs	0.7	0.3

This credence table can be interpreted as follows: when there are no LAWs the likelihood for war is 70% and 30% for peace. Calculating the 'expected utility' for each action based on this credence table gives:

$$U(\text{LAWSs}) = -100 \cdot 0.2 + 0 \cdot 0.8 = -20$$

$$U(\text{No LAWs}) = -50 \cdot 0.7 + 20 \cdot 0.3 = -29$$

Relative to this particular credence table, Jeffrey's theory yields a preference for pursuing Lethal Autonomous Weapons. A conclusion in line with the idea of *deterrence*, a strategy intended to prevent an adversary from taking action, an approach that gained prominence during the cold war. This type of theory allows to overcome the dominance principle at the cost of introducing additional subjective value judgements.

One might wonder how useful expected utility theories are when it comes to decisions in the real world. The main concern is the fact that the theory itself does not address the important questions of representation and modelling. An agent has to make an initial choice regarding representation. He must determine which options have to be considered but also a way to interpret the possible outcomes of these options. But such a choice, if it is to be justified, must surely be governed by a theory of rational decision. "This seems to lead to an *infinite regress*: before using decision theory to

make a particular choice, an agent apparently needs to employ the theory to decide how to frame the decision problem". (Steele, Stefansson 2016, par. 4) This argument can be countered by arguing that the initial representation of the world is beyond scope, in line with normative models, and it suffices to describe the original decision problem in as much detail as required.

Jeffrey's theory can be regarded as a Bayesian approach. *Bayesian Decision Theory* is currently the dominant theoretical model for analyzing decisions, both from a descriptive and normative perspective. The main idea is the analysis of rational degrees of belief in terms of rational *betting* behavior. Bayesian Decision Theory can be viewed as an element of *Bayesian Epistemology* which has been a very important development in epistemology during the last century and a promising path for future development. The main contribution of Bayesian Epistemology is the introduction of a *formal apparatus* for inductive logic that can be epistemologically justified. (Talbot 2016, par. 0) This formal apparatus can be understood as a pair of constraints, the first constraint describes how various degrees of belief relate to one another at certain moment in time, governed by the laws of probability, while the second constraint describes how degrees of belief should evolve over time using deductive rules of inference. The most important rule of inference is based on the *simple principle of conditionalization* which basically means one has to update one's *prior* beliefs to generate a *posterior* beliefs when new evidence is acquired. Via *Dutch Book Arguments* – a *Dutch Book* is a bet where an agent loses resources without gaining anything - it is possible to justify the formal apparatus. As a consequence, one could attribute Bayesian Decision Theory with the same epistemological status as the laws of deductive logic. (Talbot 2016, par. 3) At as result, Bayesian Epistemology basically solves the problem of induction as remarked earlier.

Thanks to the property to reach a logically certain conclusion, Bayesian reasoning is often viewed as the reasoning of a rational mind. It can be argued that the Bayesian view has dethroned Karl Popper's Theory of Falsification as the dominant philosophy of science. Instead of trying to falsify a hypothesis, Bayesian evidence rather confirms the hypothesis. A Bayesian approach is superior for making decisions when there is a high level of uncertainty and limited information, together with expert knowledge or historical knowledge. (Hitendra, Krutarth 2015, p. 193) Bayesian Epistemology has also explored the social aspect of decision making. In scientific inquiry it is the community of scientists that determine what is accepted rather than individual scientists. An important open question remains whether beliefs of several Bayesian decision makers can be combined into a single belief that respects individual preferences. This would open up the possibility to formally obtain humanity's subjective belief in certain events. According to some there is no room for a Bayesian compromise. (Seidenfeld, Kadane, Schervish 1989, p. 226) Although a Bayesian approach will always have certain shortcoming, decisions concerning complex future events such as the singularity will have to be made under uncertainty so it can be argued that such decisions should be made in a way that reflects this uncertainty and can be adapted when new data appears.

6.3 Black Swans and Dragon Kings

In recent years there has been an increased academic focus on highly unlikely events such as existential risks, potentially inspired by the advent of two novel theories with rather exotic names. The first theory, *Black Swan theory*, originated from Nassim Nicholas Taleb's book 'The Black Swan: The Impact of the Highly Probable.' It gained notoriety because of its implied prediction of the financial crisis one year later in 2008.

> "What we call here a Black Swan is an event with the following three attributes. First, it is an *outlier*, as it lies outside the realm of regular expectations, because nothing in the past can convincingly point to its possibility. Second, it carries an extreme impact. Third, in spite of its outlier status, human nature makes us concoct explanations for its occurrence *after* the fact, making it explainable and predictable." (Taleb 2007, p. xvii-xviii)

Hence a Black Swan event can be defined as a *rare* event with an *extreme impact* and *predictability in hindsight*. Taleb argues that almost everything can be explained by a small number of black swans and their impact is even accelerating, underlining the importance to study rare and extreme events. Black Swan events are not necessarily characterized by a sudden appearance, they can be the result of a slow process of incremental changes into a particular direction. Black Swans are extremely fragile to miscalculation, occasionally they are overestimated but humans in general tend to underestimate their probability, a logical result of limitations of our prediction abilities. These limitations can be said to arise from the nature of the activity of predicting - too complicated, not just for us but for any tools we have or conceivably can obtain, even superintelligence. The singularity, assuming it will have a negative outcome, seems to have all the makings of a Black Swan event. Let us take the example of human extinction, arguably a rare event with an extreme impact. Although it would be practically impossible to analyze the predictability of this event in hindsight, it seems not preposterous to argue that it will turn out to be a highly predictable event. The event itself is believed to be instantaneous but the buildup seems already long underway, its plausibility being one of the topics of this thesis. It seems safe to assume its explanatory power and as a result approach the singularity as Black Swan. According to some recent authors there are multiple possible interpretations of Black Swans, the two most important ones 1) a surprising extreme event relative to the expected occurrence and 2) an extreme event with low probability. (Aven 2013). Again it seems safe to classify the singularity as a Black Swan under these definitions.

But is anything gained by this? Does there exist a way to contain Black Swans if they are identified in a timely manner? Taleb followed up his book with a one pager listing ten principles for a *Black Swan-proof world*. Most principles are cheek in tongue formulations of conventional wisdom, like "Do no give children sticks of dynamite, even if they come with a warning." (Taleb 2009, p. 1) Not very useful but it is an active topic of research in risk management since risk management involves by definition decisions under uncertainty. According to some, the Bayesian perspective is needed to quantify epistemic uncertainties when new or poorly known factors such as new technologies are at play. (Paté-Cornell 2012, p. 1826) But other models for risk analysis that deals with deep uncertainty are possible, including machine learning based approaches; the concept of combining predictions of

multiple models; ensembles, and even reinforcement learning. Such models could provide genuine breakthroughs in order to improve predictions and decisions under high uncertainty. (Cox 2012, p. 1607) Although Black Swan theory doesn't offer obvious solutions yet, its contribution to the overall acceptance of the idea of highly improbably extreme impact events can be considered an important step.

The second theory, *Dragon Kings*, is developed by Didier Sornette who has an impressive track record when it comes to decision making under uncertainty. He also developed the *Quantum Decision Theory*, based on the mathematics of Hilbert Spaces formalizing the concept of uncertainty. Dragon Kings focuses more on dynamics. The hypothesis is that Dragon Kings appear as a result of amplifying mechanisms. Dragon Kings themselves are defined as extreme events that do not belong to the population of the other events. (Sornette 2009, p. 1) Without any doubt, the singularity can be considered a Dragon King. The underlying mechanisms and dynamics of Dragon Kings are complex. They can result from abrupt shocks within a random walk process, as a natural occurrence in systems that exhibit the *Zipf Law* distribution – the frequency of occurrence is inversely proportional to its rank in the frequency table - but also from positive feedback leading to singular shocks. (Sornette 2009, p. 3-4) Several methods have been shown to be capable of identifying Dragon Kings in certain domains. Sornette indicates the extinction of species is a possible Dragon King (Sornette 2009, p. 2) arguably the main risk we want to avoid as an outcome of superintelligence. In general we would wish to be able to forecast or predict Dragon King events and the theory is promising. If Dragon King events are the result of a top-down process within a system, they may have a predictable distribution, possible a Poisson-like distribution. In that case, Dragon Kings would be unpredictable but high quality forecasts should be possible. If they are the result of bottom up processes, they originate due to amplification mechanisms at lower levels in the system. In this scenario they might be preceded by a precursory activity that allows quality predictions. Such early warning signals are often quantifiable in the form of increasing correlations, increasing variance of endogenous fluctuations and increasing spatial coherence. From an operational point of view, complex simulation platforms as studied by Dorner, incorporating all available data and feedback loops, are promising here. (Dorner 1997) A salient point is that such Dragon King simulators should be free from cognitive bias, behavioral flukes and politics. Which might ironically result in AI managing the risks of AI. In any case, more research seems required, Sornette calling the extension of research in such simulations "perhaps the most pressing challenge of modern times." (Sornette 2009, p. 15) If this is in reference to the necessity to control and monitor the progress of future complex technologies this definitely hits the nail on the head.

Conclusion

The singularity seems destined to become a topic of increasingly intense discussions over the coming years, gaining prominence as Artificial Intelligence slowly penetrates every aspect of modern life. Although it seems likely the singularity itself will remain an elusive idea in the foreseeable future, several key elements are gaining traction as the impact of AI and advanced technologies is becoming more tangible every day.

One always has to make some kind of inductive leap to discuss future events but the leap is gradually shrinking. The discussion used to focus on the plausibility of infinite accelerating progress. After more than 25 years of progress, the current state of technological affairs is sufficient for superintelligence and no immediate slowdown is expected. The advent of autonomous systems capable of making their own decisions is chewing off another piece of the inductive leap. Additionally superintelligence itself is becoming a serious topic, with nations publicly declaring their ambitions to become dominant forces in this field.

It is hardly a surprise that the community around the singularity has matured. Focus has shifted, from questioning its premise or hypothesizing about dystopian futures towards a more pragmatic perspective. This includes an increasing focus on potential challenges and risks that might be encountered. An obvious example is the control problem of artificial intelligence which is receiving more and more attention. An encouraging evolution, however, we cannot help but question its added value in the long run as the control problem seems to be subsumed by the control paradox. If there is one topic that we would like to see gain importance from a philosophical perspective it would be this topic, more specifically the following question: 'If we will not be able to control higher intelligence, should we develop it?' An interesting question and one that forces us to take a deep look in the mirror.

Realistically speaking, it is likely that the control problem - and the general question of how to deal with challenges and risks of AI - will become increasingly important. From this point of view, several developments in risk and decision theory seem to be offering a promising perspective when descending down the path towards superintelligence. Whether this will be followed by the materialization of the singularity or not, only time can tell. In any case, AI seems to make philosophy honest indeed (Dennett 2006) which can never be a bad thing.

Bibliography

Adams, H. 1907. *"The Education of Henry Adams".* Self-published.

Aghion, P. and Howitt, P. 1994. "Growth and Unemployment." In *The Review of Economic Studies* 61 (3), pp. 477–494. Oxford University Press.

Allen, G and Chan, T. 2017. "*Artificial Intelligence and National Security*." Available at https://www.belfercenter.org/ Accessed 2018-04-19.

Arel, I. 2012. "The Threat of a Reward-Driven Adversarial Artificial General Intelligence." In *Singularity Hypotheses: a Scientific and Philosophical assessment,* edited by Eden et al, pp. 43-58. Springer.

Aven, T. 2013. "On the Meaning of a Black Swan in a Risk Context." In *Safety Science* 57, pp. 44-51. Elsevier.

Baum, S. 2017. "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy." In *Global Catastrophic Risk Institute Working Paper 17-1*. Available at https://ssrn.com/abstract=3070741. Accessed 2018-04-19.

Bostrom, N. 1998. "Singularity and predictability." In *Comments on Vinge's Singularity* Available at: http://mason.gmu.edu/~rhanson/vc.html. Accessed: 2018-03-24.

Bostrom, N. 2000. "When Machines Outsmart Humans." In *Futures* 35 (7), pp. 759-764. Elsevier.

Bostrom, N. 2002. "*Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards*." Available at https://nickbostrom.com/existential/risks.pdf Accessed: 2018-05-10.

Bostrom, N. 2003. "Are You Living in a Computer Simulation?". In *Philosophical Quarterly* 53, (211), pp. 243-255. Wiley Blackwell.

Bostrom, N. 2005. "A History of Transhumanist Thought." In *Journal of Evolution and Technology* 14 (1), pp. 1-25. Institute for Ethics and Emerging Technologies

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

Brain Projects 2018. Available at https://www.humanbrainproject.eu/en/ & https://bluebrain.epfl.ch/ Accessed: 2018-05-14.

Brynjolfsson, E. and McAfee, A. 2011. *"Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy."* pp 1-7. Digital Frontier Press.

Chaisson, E. 2012. "A Singular Universe of Many Singularities: Cultural Evolution in a Cosmic Context." In *Singularity Hypotheses: a Scientific and Philosophical assessment*, edited by Eden et al, pp 413-439. Springer.

Chalmers, D. J. 2010. "The Singularity: A Philosophical Analysis." In *Journal of Consciousness Studies* 17, pp. 7-65. Imprint Academic.

Cox, T. 2012. "Confronting Deep Uncertainties in Risk Analysis." In *Risk Analysis* 32 (10), pp. 1607-1629. Wiley.

Dorner, D. 1997. *"The Logic of Failure: Recognizing and Avoiding Error in Complex Situations."* Basic Books.

Dennett, D. C. 2006. "*Computers as Prostheses for the Imagination."* Paper presented at the International Computers and Philosophy Conference, Laval, France, May 5–8. (Quoted from Muehlhauser, Salamon 2012.)

Dijkstra, E., 1984. *Transcript from ACM South Regional Conference*, November 16-18, Austin Texas. Available at http://www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD898.html Accessed: 2018-05-10.

Eden, A. H., Moor, J. H., Soraker, J. H. and Steinhart, E. 2012. "Singularity Hypotheses: an Overview". In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Eden et al, pp. 1-11. Springer.

Fagnant, D. J. and Kockelman, K. 2015. "Preparing a Nation for Autonomous Vehicles: Opportunities, barriers and policy recommendations." In *Transportation Research Part A: Policy and Practice* 77, pp. 167-181. Elsevier.

Fisher, M. P. A. 2015. "Quantum Cognition: The Possibility of Processing with Nuclear Spins in the Brain." In *Annals of Physics* 362, pp. 593-602. Elsevier.

Frey, C. B. and Osborne, M. A. 2013. "*The future of employment: how susceptible are jobs to computerization?".* Available at https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf Accessed 2018-04-19

Gasarch, W. I. 2012*. "The Second P=? NP Poll."* Available at http://www.cs.umd.edu/~gasarch/papers/poll2012.pdf Accessed 2018-04-19.

Gates, B. 2015. "*Stephen Hawking, Elon Musk and Bill Gates Warn about artificial intelligence."* Available at http://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/ Accessed 2018-04-19.

Geist, E. M, 2016. "It's Already Too Late to Stop the AI Arms Race-We Must Manage it Instead." In *Bulletin of the atomic Scientists 72 (5),* pp. 318-321. Taylor and Francis.

Gigerenzer, G. and Selten, R. 2001. “Rethinking Rationality.” In *Bounded Rationality: The Adaptive Toolbox*, pp. 1-13. MIT press.

Good, I. J. 1965. “Speculations Concerning the First Ultraintelligent Machine.” In *Advances in Computers* 6, pp. 31-88. Academic Press.

Gubrud, M. A. 1997. *“Nanotechnology and International Security.”* Available at: https://foresight.org/Conferences/MNT05/Papers/Gubrud/ Accessed: 2018-05-14.

Hameroff, S. and Penrose, R. 2014. “Consciousness in the Universe: A Review of the Orch Or Theory” In *Physics of Life Reviews* 11, pp. 39-78. Elsevier.

Hawking, S. 2015. *“Stephen Hawking, Elon Musk and Bill Gates Warn about artificial intelligence.”* Available at http://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence/ Accessed 2018-04-19.

Henderson, L. 2018. "The Problem of Induction". In *The Stanford Encyclopedia of Philosophy*, Summer 2018 Edition (forthcoming), edited by Zalta, E. N. . https://plato.stanford.edu/. Accessed: 2018-03-24.

Hilbert, M. and López, P. 2011. “The World’s Technological Capacity to Store, Communicate, and Compute Information.” In *Science 332,* pp. 60 –65. AAAS.

Horgan, J. 2008. “The Consciousness Conundrum.” In *IEEE Spectrum* 45 (6) pp 36-41. IEEE.

Hitendra, D. P. and Krutarth, D. H. 2015. “Application of Bayesian Decision Theory in Management Research Problems.” In *International Journal of Scientific Research Engineering & Technology*, Conference Proceeding, 14-15 March, 2015, pp. 191-195. IJSRET.

Hutter, M. 2005. “*Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*.” Springer.

Jeffrey, R. C. 1983. “The Logic of Decision.” University of Chicago Press.

Keynes, J.M. 1933. “Economic Possibilities for Our Grandchildren.” In *Essays in persuasion*, pp. 321-332. Macmillan.

Koene, R. A., 2012. “Embracing Competitive Balance: The Case for Substrate-Independent Minds and Whole Brain Emulation.” In *Singularity Hypotheses: a Scientific and Philosophical assessment,* edited by Eden et al, pp. 241-266. Springer.

Kent, A. 2017. *“Quanta and Qualia.”* Available at https://arxiv.org/pdf/1608.04804.pdf Accessed 2018-05-14

Kožnjak, B 2015. "Who Let the Demon Out? Laplace and Boscovich on determinism". In *Studies in History and Philosophy of Science.* 51, pp 42–52. Elsevier.

Kurzweil, R. 2005. “*The singularity is Near: When humans transcend biology.”* Viking.

Laplace, Pierre Simon 1902/1814. “*A Philosophical Essay on Probabilities*.” Translated into English from the original French 6th ed. by Truscott, F.W. and Emory, F.L.. Dover Publications.

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C.E. 1955. *“A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”*. Available at http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf Accessed: 2018-04-19.

McDermott, D. 2012. "*There Are No "Extendible Methods" in David Chalmers's Sense Unless P = NP."* Available at: http://www.cs.yale.edu/homes/dvm/papers/no-extendible-methods.pdf Accessed 2018-04-19.

Meadows, D. H., Meadows, D. L, Randers, J., & Behrens, W. 1972. *"The Limits to Growth: a Report for the Club of Rome Project on the Predicament of Mankind."* Universe Books.

Mitchell, T. 1997. *"Machine Learning".* McGraw Hill.

Modis, T. 2003. "The Limits of Complexity and Change." In *The futurist* (May–June), pp. 26–32. World Future Society.

Moore, G. E. 1965. "Cramming More Components Onto Integrated Circuits." In *Electronics* 38(8), pp 114–117. IEEE.

Moore, G. E. 2015. "Gordon Moore: The Man Whose Name Means Progress." Available at https://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress Accessed: 2018-05-25

Moravec, H. 1978. *"Today's Computers, Intelligent Machines and Our Future."* Available at http://www.frc.ri.cmu.edu/~hpm/project.archive/general.articles/1978/analog.1978.html Accessed: 2018-04.19.

Moravec, H. 1988. *"Mind Children."* Harvard University Press.

Muehlhauser, L. and Helm, L. 2012. "The Singularity and Machine Ethics". In *Singularity Hypotheses: a Scientific and Philosophical Assessment,* edited by Eden et al, pp. 101-126. Springer.

Muehlhauser, L. and Salamon, A. 2012. "Intelligence Explosion: Evidence and Import". In *Singularity Hypotheses: a Scientific and Philosophical Assessment,* edited by Eden et al, pp. 15-40. Springer.

Mueller, V. C. and Bostrom, N. 2016. "Future Progress in Artificial Intelligence: A survey of Expert Opinion." In *"Fundamental Issues of Artificial Intelligence",* Edited by Mueller, V. C. pp. 555-572. Springer.

Obama, B. 2016. "Barack Obama, Neural Nets, Self-Driving Cars, and the Future of the World." In *Wired Magazine.* Available at: https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/ Accessed: 2018-04-19.

Odenwald, S. F. and Green, J. L. 2008. "Bracing For a Solar Superstorm." In *Scientific American* 299 (2), pp. 80-87. Scientific American.

Omohundro, S. 2012. "Rational Artificial Intelligence for the Greater Good". In *Singularity Hypotheses: a Scientific and Philosophical Assessment,* edited by Eden et al, pp. 161-176. Springer.

Paté-Cornell, M.E. 2012. "On Black Swans and Perfect Storms: Risk Analysis and Management When Statistics Are Not Enough." In *Risk Analysis* 32 (11), pp. 1823-1833. Wiley.

Peterson, M. 2009. "An introduction to Decision Theory." Cambridge University Press.

Pinker, S. 2008. "The Consciousness Conundrum." In *IEEE Spectrum* 45 (6) pp 36-41. IEEE.

Plebe, A. and Perconti, P. 2012. "The Slowdown Hypothesis." In *Singularity Hypotheses: a Scientific and Philosophical Assessment,* edited by Eden et al, pp. 349-362. Springer.

Popper, K. 1934. "*The Logic of Scientific Discovery*." Mohr Siebeck.

Putin, V. 2017. Available at https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/ Accessed: 2018-05-14.

Reichenbach, H. 1930. "Causality and Probability." In *Erkenntnis 1*, pp. 158-188. Springer.

Rokeach, M. 1973. *The Nature of Human Values*. Free Press.

Russel, S. 2015. "Take a Stand on AI Weapons." In *Nature* 521, pp 415-416. Springer Nature.

Savage, L. 1954. *"The Foundations of Statistics."* Wiley.

Savage, L. 1961. "The Foundations of Statistics Reconsidered." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp 575-586. University of California Press.

Sandberg, A. 2010. *"An Overview of Models of Technological Singularity."* http://agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf. Accessed: 2018-03-24.

Sandberg, A. and Bostrom, N. 2008. *"Whole Brain Emulation: A Roadmap, Technical Report"* Future of Humanity Institute, Oxford University.

Seidenfeld, T., Kadane J., and Schervish, M. J. 1989, "On the Shared Preferences of Two Bayesian Decision Makers." In *Journal of Philosophy*, 86: 225–244. The Journal of Philosophy Inc.

Shiller, R 2018. "How to Stop 'Extremely Disruptive' AI from Harming Society: Robert Shiller." Available at https://www.cnbc.com/2018/01/18/ai-is-extremely-disruptive-robert-shiller.html Accessed: 2018-05-14

Sloman, S. A. and Lagnado, D. A 2005. "The Problem Of Induction." In *The Cambridge Handbook of Thinking and Reasoning*, edited by Holyoak, K. J. and Morrison, R. G., pp. 95-116. Cambridge University Press.

Solomonoff, R. J. 1985. "The Time Scale of Artificial Intelligence: Reflections on Social Effects." In *North-Holland Human Systems Management* 5, pp 149-153. Elsevier.

Sornette, D. 2012. "Dragon-Kings: Mechanisms, Statistical Methods and Empirical Evidence." In *The European Physical Journal Special Topics* 205, pp. 1-26. Springer-Verlag.

Steele, K. and Stefánsson, H. O. 2016. "Decision Theory." In *The Stanford Encyclopedia of Philosophy"* edited by Zalta. Available at https://plato.stanford.edu/ Accessed: 2018-04-19

Taleb, N. N. 2007. "*The Black Swan: the Impact of the Highly Improbable."* Random House.

Taleb, N. N. 2009. *"Ten Principles for a Black Swan-proof World."* Available at https://www.ft.com/content/5d5aa24e-23a4-11de-996a-00144feabdc0 accessed 2018-04-19.

Talbott, W. 2016. "Bayesian Epistemology." In *The Stanford Encyclopedia of Philosophy"* edited by Zalta. Available at https://plato.stanford.edu/ Accessed: 2018-04-19.

Tegmark, M. 2000. "Importance of Quantum Decoherence in Brain Processes." In *Physical Review* 61, pp. 4194-206. American Physical Society.

Titelbaum, M.G. 2016. *"Fundamentals of Bayesian Epistemology."* Oxford University Press.

Toffler, A. 1970. *"Future shock."* Random House.

Turing, A. M. 1948. "Computing Machinery and Intelligence." In *Mind 49,* pp. 433-460. Oxford University Press.

Ulam, S. 1958. "Tribute to John von Neumann." In *Bulletin of the American Mathematical Society* 64, pp. 1-49. American Mathematical Society.

Vinge, V. 1983. "First Word." In *Omni* 5 (4), pp. 10-16. Omni Publications.

Vinge, V. 1993. "*Technological Singularity."* Available at https://www.frc.ri.cmu.edu/~hpm/book98/com.ch1/vinge.singularity.html Accessed: 2018-05-14.

Walsh, T. 2017. "The Singularity May Never Be Near." In *AI Magazine* 38,3. pp. 58-62. AAAI.

Wilson, R. A. and Foglia, L. 2017. "Embodied Cognition." In *The Stanford Encyclopedia of Philosophy"* edited by Zalta. Available at https://plato.stanford.edu/ Accessed: 2018-04-19.

Wunderlich, K., Dayan and P., Dolan, R. 2012. "Mapping Value Based Planning and Extensively Trained Choice in the Human Brain." In *Nature Neuroscience* 15, pp 786 – 791. Nature.

Yampolskiy, R. V. and Fox, J. 2012. "Artificial General Intelligence and the Human Mental Model". In *Singularity Hypotheses: a Scientific and Philosophical assessment,* edited by Eden et al, pp. 129-146. Springer.

Yampolskiy, R. 2017. "*The Singularity May Be Near."* Available at https://arxiv.org/ftp/arxiv/papers/1706/1706.01303.pdf Accessed: 2018-04-19.

Yudkowsky, E. 1996. "*Staring into the singularity."* Available at http://yudkowsky.net/obsolete/singularity.html. Accessed: 2018-03-24.

Yudkowsky, E. 2007a. "*Three Major Singularity Schools*." Available at https://intelligence.org/2007/09/30/three-major-singularity-schools/. Accessed: 2018-03-24.

Yudkowsky, E. 2007b. 'Levels of Organization In General Intelligence." In *Artificial General Intelligence*, edited by Goertzeland, Pennachin, pp 389-501. Springer.

Yudkowsky, E. 2008. *Artificial Intelligence as a Positive and Negative Factor in Global Risk*. Available at http://intelligence.org/files/AIPosNegFactor.pdf Accessed: 2018-05-10.