University of Tartu

Institute of Philosophy and Semiotics

Department of Philosophy

Taavi Luik

HOW CAN ARTIFICIAL INTELLIGENCE BE RISKY?

Bachelor's thesis

Supervisor: Mats Volberg

Tartu
2019

Table of Contents

Introduction

Technology under the umbrella term ‘artificial intelligence’ (AI) has reached some new heights in the last decade, resulting in widespread interest and usage. Unfortunately, this is often accompanied with exaggerated claims, conflicting predictions and opinions, sometimes also a lack of understanding over the inner workings of these technologies. What is clear, however, is that we have the ability to create systems that outperform humans and we are eager to put them to use.

The purpose of this thesis is to analyze AI from the perspective of risks. The main goal is to show that already existing AI systems (what are often called ‘narrow AI’) pose serious risks, not merely themselves, but how they are intentionally used by human beings.

I define risks in a relatively straightforward way – a risk is the potential to lose something of value. Risks can range in terms of scope from small to existential, they can be likely or unlikely, predictable or unpredictable, affect groups of people, all of humanity or the planet as a whole. ‘Artificial intelligence’ is defined broadly as intelligence demonstrated by machines. Most current AI systems can be rightly said to lack generality or indeed lack most characteristics of human intelligence, yet this does not mean that they are not powerful or do not pose risks.

The motivation for writing on this topic stems from three facts. Firstly, we define and distinguish ourselves from other species largely through our intelligence, which permits us to substantially change our environment to fit our goals, but also to change our goals themselves. We now have the ability to create systems that potentially match or surpass our abilities. There are many ways how these systems can end up imperfect and it is up to us to reduce the likelihood of this happening. This requires careful acknowledgment of what and how can go wrong.

Secondly, there has been substantial technical progress in AI in the last decade or so, resulting in systems that are capable of autonomously driving vehicles (resulting in less accidents than humans), playing complex games such as Chess, Go and Dota 2 (on a superhuman level), translating or creating text and imitating or synthesizing speech (well enough to pass off as humans). While we might be far from creating artificial general intelligence (AGI), then each of these systems is capable of superhuman performance.

Sometimes, systems are put to use before they have passed scrutiny or are used to further selfish goals (zero or negative sum games in game theory[1]).

Thirdly, there is a lack of consensus among experts whether and to what extent artificial intelligence is even possible and if, then when it might be created. Some authors are very optimistic, some very pessimistic about the possibility of achieving and probable consequences of powerful AI systems. Armstrong et al note that expert views are from a wide range of „AI is impossible“ to „just around the corner“ and anything in between (Armstrong et al, 2016: 30). Such disagreement calls for philosophical reflection and the laying of some foundation which is reasonably certain. While there is much uncertainty, then the existence of some risks are clear.

Discussion about AI requires some discussion about intelligence. Therefore, in the first part, I take up the task of defining and analyzing intelligence and pointing out other related concepts to consider in a setting of risk analysis. The second part explores technology – what have we achieved, why, and whether current trends are likely to continue. Lastly, some abstract and concrete risks are pointed out, followed by noting the existence of conflicts of values and goals, which eventually requires some solution.

1 “A zero-sum game [---] is one in which one player can only be made better off by making the other player worse off.” (Ross 2019)

1 What is intelligence?

The choice of definitions strongly matters because the analysis might miss or introduce certain aspects of the phenomena, leading to misleading conclusions. The concept of 'AI' carries with itself connotations of seriousness that a mere 'algorithm' or 'system' does not. The context in which definitions are used can provide some direction, but ultimately language is a human construct and some arbitrariness and vagueness is inevitable. For the present purposes, some vagueness can even be useful, since some systems might not fall under a rigid definition, yet still pose risks.

The definitions eventually used here are neither descriptive of how humans (the larger public or experts) use these terms, nor is it normative – how they ought to be used generally. There exist many different, often conflicting definitions and it is reasonable to suggest that there are better fitting definitions for any particular context or goal.

1.1 The difficulty of defining intelligence

Intelligence is not easy to define. Tentatively, I use intelligence broadly as referring to a set of mental faculties. Partly, the difficulty stems from the recursive nature of the task – we use intelligence to define intelligence. This creates a problem.

Human intelligence is the product of evolution over long periods of time, evolving under some set of constraints which have shaped it. Human mental faculties exist contingently and it is hard to point out their borders, origin, importance or independence. This means that when defining artificial intelligence, we run the risk of requiring too much human-likeness. A very different environment requires different faculties and it is hard to imagine what is possible or impossible, necessary or redundant, making this to an extent a speculative question. The world is chaotic, no set guarantees success since there is always something that is beyond one's control. The assumption here is that different sets of faculties can be functionally equivalent (or near equivalent), that the same problem can be solved in different ways and different systems can be viable.

Therefore, we must distinguish *human intelligence* from *intelligence as such*. One way of achieving this is to focus on core functionality and ignore implementation details. This allows us to be agnostic towards many interesting and otherwise relevant questions about the nature of consciousness, moral status or whether these systems truly possess understanding. At the same time, implementation details can be relevant, so if they are known, they must eventually be taken into account.

Some helpful concepts exist in the literature. **Narrow AI** are systems that operate intelligently only in a relatively narrow domain such as playing chess or medicine, in contrast to **general AI**, which is at least human level (Franklin 2014: 16), i.e. able to operate in many different domains. It is useful to think of intelligence as a spectrum and multi-dimensional – a system is not either narrow or general, but *more or less* general in *more or less* domains. Nick Bostrom defines **superintelligence** as any intellect that vastly outperforms the best humans in practically every field, leaving open the implementation by which this is achieved. He explains that even if the probability of superintelligence emerging is small, it must still be given serious consideration, simply because of the potential consequences it might have (Bostrom 2003).

1.2 Turing's observation

According to Alan Turing, the question whether machines can think is too meaningless to deserve discussion on its own. Instead, more precise and related questions should be considered. (Oppy, Dowe 2016)

Turing had a specific idea in mind (an 'Imitation Game', now more broadly called the Turing Test) to test the intelligence of machines via their ability to deceive humans (Oppy, Dowe 2016). A system passes the Turing test if a human interrogator, after having written some questions, cannot say whether the responses came from a person or a computer (Russell, Norvig 2016: 2).

It is not clear to what extent is deception an indicator of intelligence (is it part of necessary or sufficient conditions?). Russell and Norvig note that while the Turing test is still relevant many years later, then "The quest for 'artificial flight' succeeded when the Wright brothers and others stopped imitating birds and started using wind tunnels and learning about aerodynamics." (Russell, Norvig 2016: 3). What matters is that we move from vague concepts to more concrete ones. Additionally, while a Turing Test might not be a good test for measuring intelligence, then any deception test (including whether the system is capable of intentionally failing the test) is especially relevant for a risk analysis.

1.3 Intelligence as achieving goals

Shane Legg and Marcus Hutter compile a long list of known informal definitions of intelligence, compare them and come up with a formal definition for universal intelligence, one that is applicable both to humans and machines. They present an informal definition as the following: „Intelligence measures an agent's ability to achieve goals in a wide range of environments", stressing that this contains three essential components: an agent, an

environment and goals (Legg, Hutter 2007: 1,15). We can point out two shortcomings of this definition for the present purposes.

Firstly, "a wide range of environments" hints at generality. While it is reasonable to assume that a more intelligent system might be more capable in a wider range of environments, then lack of generality itself does not mean that the system is not powerful or risky. Fundamentally, narrow and general AI systems are both engaging in the same process. They are acquiring some knowledge that was not manually built in – they are learning.

Secondly, "the ability to achieve" links intelligence to action or agency. Questions can be raised about the nature of the link between intelligence and action in general – can a system consistently behave intelligently without being intelligent? The authors defend this link – in a section replying to John Searle's Chinese Room Argument, they write that if the system lacks understanding yet performs identically to a system that has understanding, then "[...]it is not even clear to us how to define 'understanding' if its presence has no measurable effects". However, if 'understanding' does have a measurable impact, then it is of interest. The authors note that they desire a simple and very general definition and this is easier to achieve if they abstract over the internal workings of the agent – what matters is how well something works. The method of achieving the goal can even be absurdly inefficient. (Legg, Hutter 2007: 41-42)

However, one could bring many modern AI systems as examples which lack any sort of agency (insofar as providing the result of a classification task is not really agency), yet these systems can possess dangers – agency can come from outside, from the operators of these systems. I do not wish to only focus on AI systems themselves as risks, but also see what they could be used for. 'Machine ethics' is the area of research that deals with giving ethical principles to AI systems in order to increase the likelihood of positive social outcomes (Brundage 2016b: 87-88). Miles Brundage writes that „...a 'technical' solution to machine ethics may mean little in a world in which unethical humans exist and have access to advanced technology" (Brundage 2016b: 107).

Based on this discussion, Legg and Hutter's definition seems to be a better fit for AGI rather than AI systems in general.

1.4 Intelligence as learning

Taking the previous criticism into account, we could separate concepts and redefine intelligence as only dealing with the process of learning. *Prima facie*, a system could

possess knowledge in various domains and the ability to learn more, yet have very limited (direct or indirect, i.e. through other agents) agency in the world. Additionally, we ought to explicitly think in terms of predictability, since ultimately this is also to a degree independent of intelligence and agency. If a system is sufficiently predictable – either it is transparent, slow or limited in some ways that it requires human cooperation etc. – it poses limited risks.

In order to justify defining intelligence as learning, we could look at two motivations behind creating AI systems. Firstly, we want these systems to help us do something that we already can do – e.g. we want to automate tasks we find tedious, dirty or dangerous. Yet we also want some systems to produce new knowledge or possess abilities we lack. In both of these cases, these are achieved, at least in the long term, through learning. Some knowledge can be manually built in, but e.g. in order to teach a robot to walk, it must learn it on its own (although through our scaffolding that facilitates the learning).

How is learning achieved, what is knowledge and how is it represented etc – these are otherwise very important details to work out (Russell, Norvig 2016 provides an extensive overview of various aspects of AI research). As long as the system learns, it should be considered intelligent. This trio of concepts – intelligence as learning, agency as (direct or indirect) acting in the world, and predictability, each a multidimensional spectrum – seem to be a better fit for risk analysis than mere intelligence on its own.

1.5 The problem with requiring human-likeness

There is a tendency to measure AI by comparing it to humans. To an extent, this is understandable – human beings indeed possess more faculties than other species and this is a good place to start. However, I wish to argue that this can be unhelpful.

Nick Bostrom introduces two theses related to AI systems. The **orthogonality thesis** states that goals and intelligence are independent variables – an AI system could have any combination of abilities and final goals. The **instrumental convergence thesis** states that superintelligent agents with various final goals will have common intermediary goals due to instrumental reasons. (Bostrom 2014: 105). The orthogonality thesis keeps naive optimism at bay (e.g. see Steven Pinker[2]) by giving a reason to consider very pessimistic scenarios and agents that are completely inhuman. This widens the range of systems to consider, but at the same time, the instrumental convergence thesis shows

2 Pinker argues that seeing AI as dangerous is another exaggerated doomsday scenario, see his essay at: www.theglobeandmail.com/opinion/the-dangers-of-worrying-about-doomsday/article38062215/

the universal value of certain types of goals, meaning that they are more likely to be found in intelligent systems, and as such, narrows the range. Powerful systems have the potential to be more dangerous and looking for systems that have the ability or desire to acquire more power seems a good heuristic.

Bostrom also brings out some sources of advantage for a digital intelligence (Bostrom 2014: 59-60):

- The speed of computational elements – biological neurons operate 7 orders of magnitude slower than modern microprocessors (200 Hz *vs* ~2 GHz).
- Internal communication speed – axons carry signals at the speed of ~120 m/s, while electrical signals move at the speed of light (300 000 000 m/s).
- Storage capacity – the amount of information stored and the speed and accuracy at which it is done can be much greater in digital systems.
- Editability – software is easier to change and improve than hardware, leading to more experiments and improvements.

This could be supplemented with an analysis of human limitations that AI systems need not have. Biological constraints limit us to a narrow comfortable environment, we need to rest and sleep, we have a strong desire of self-preservation. AI systems can and often are different, especially if they lack a physical body. Of course, self-preservation is one of those aspects that are likely to be found in persistent systems, yet this does not have to be as strong or as static. Also, lacking human limitations or weaknesses does not mean that AI systems have none. However, the known advantages of digital intelligences already constitute a major advantage.

Computer scientist Richard S. Sutton has noted that in the last 70 years of machine learning, computational and statistical methods keep winning over methods that are based on human knowledge. Moore's law[3] has created massive computation which has been put to good use. AI researchers have often built knowledge into their agents, and while this has helped in the short term, then progress is eventually achieved by opposing approaches that are based on scaling computation. He writes that "The two methods that seem to scale arbitrarily in this way are search and learning" and "We want AI agents that can discover like we can, not which contain what we have discovered." (Sutton 2019)

This means that the best AI systems are often not like us, yet outperform us, providing empirical evidence that human-centrism can be unhelpful and unnecessary.

3 Moore's law refers to the observation that the number of transistors in a microchip double roughly every two years. Sutton calls a generalization of Moore's law as the ultimate reason behind the success: the "continued exponentially falling cost per unit of computation" (Sutton 2019).

1.6 The possibility of general AI

The 'no-free-lunch' theorem states that general superintelligence is a free lunch and there are no free lunches. Colloquially, these theorems are stated as: „Algorithms are successful only when they are 'tuned' to their domain; there are no universal learning algorithms." (Danks 2014: 158)

Even if true, this does not rule out potential AI-s outperforming human beings, nor does it rule out AI as such (Armstrong et al 2016: 37). Defining intelligence as continual learning or adapting seems to provide a solution. Humans are to an extent general intelligences, but this is not static – we achieve generality through adapting to the goal and the environment, we modify our mindset, acquire relevant tools and knowledge etc. Systems can adapt to their environment and goals, making the previously impossible or hard become possible or easy – these systems can be general enough to be successful and pose threats.

1.7 The problem with requiring generality

So, while generality can (to an extent) be achievable, it can still be unnecessary and even misleading, since AI systems need not exist in isolation.

Consider how easy it is for most people in a developed country to improvise an explosive device – the information can be found with a simple Internet search and ingredients can be bought without questions from construction and electronics shops. It is fortunate that most people do not consider actually creating such devices, but it is possible nevertheless, and basically only requires access to the Internet, some money, and the ability to read and follow basic instructions.

Now, imagine a person with the same goal, but living in isolation. This person might not even have the knowledge of explosive chemical reactions, let alone the ability to create the necessary materials and assemble them together. We have created an environment where many tasks can be delegated to others and the free market takes care of the details. The agent itself need not even acknowledge who they are using or the amount of work that went into the result.

AI systems, if they do not exist in isolation, are in a similar position. They too do not have to possess general intelligence or even superhuman intelligence in a narrow domain in order to achieve very impactful goals. If they have access to the Internet, they have access to a myriad of free resources and potential helping hands to achieve their goals, as long as they pass off as humans or find curious, naive or like-minded people – or even

other AI systems. Most details can remain hidden behind multiple layers of abstraction and many goals can become achievable.

One could object that the mere possibility of something happening is not enough to cause concern – there are terrorists who drive vehicles to crowds, killing many, yet we do not consider banning vehicles. The orthogonality and instrumental convergence theses are relevant here – an AI system might discover and do things that humans do not, it might find weaknesses that are not visible or exploitable for humans.

The example of improvising an explosive device is imperfect as it is mostly a physical process and most AI systems are not physical agents. However, we have created and are moving even more towards a world that is digital (hence potentially accessible to digital AI systems), where there are a lot of specialized agents and where a marketplace exists (i.e. tasks can be delegated). A single system does not have to be general. Generality could be achieved collectively, similarly how humans have limited knowledge and abilities, but collectively we can achieve very complex goals.

2 AI Technology

Frankish and Ramsey note that philosophers cannot ignore the actual achievements AI research (Frankish, Ramsey 2014: 1, 3). This is especially relevant today since promising new ideas and state-of-the-art results are presented in some area every few months. I do not wish to provide an extensive overview, but merely point out that powerful AI systems are not something to only expect from the future. Superintelligence, according to Bostrom, might come suddenly (Bostrom 2003), but powerful and potentially dangerous AI technology is already here.

I also wish to present some problems surrounding the predictions and evaluations of AI technology, discuss the possibility that the last decade of progress is due to the picking of low hanging fruits, explain some reasons behind this progress and analyze a line of argumentation that points out the limits of AI systems.

2.1 Difficulties of predicting or evaluating AI

There are three problems I wish to point out – experts and intuition are unreliable, goalposts tend to be moved *post hoc* and there exist some double standards.

Armstrong et al present tools to analyze, judge and improve AI predictions. The key lessons that they learned are that experts are overconfident, it is important and possible to derive testable predictions, model-based predictions are superior to those based on judgment and that it is very difficult to assess the reliability of predictors (Armstrong et al, 2016: 32). They conclude that there are theoretical and practical reasons to claim that timeline predictions are completely untrustworthy (Ibid, 46). The authors analyze the predictions made in the Dartmouth conference and those made by Hubert Dreyfus, John Searle, Ray Kurzweil and Steve Omohundro. The only consistent message in all predictions were that all predictors were overconfident in their verdicts. The authors suggest future predictors to learn from this – that they make their assumptions explicit, their models clear, predictions testable and uncertainty greater. (Ibid, 64)

There is often a negative reaction towards AI achievements. Pamela McCorduck writes (McCorduck 2004: 204):

> "..it's part of the history of the field of artificial intelligence that every time somebody figured out how to make a computer do something—play good checkers, solve simple but relatively informal problems—there was a chorus of critics to say, but that's not thinking."

A similar point is made by Müller: “Intelligence is whatever machines haven’t done yet”. Successful things are re-branded (e.g. to ‘machine learning’) and what is left for AI are problems that are deemed impossible and other long-term visions (Müller 2016: 3). To be fair, sometimes goalposts are indeed placed poorly or the critics might be different people. The problem seems to stem from vague definitions and the lack of acknowledgment that different systems could be functionally equivalent, leading to the possibility of *post hoc* rationalizations. Intellectual honesty calls for explicit definitions that are adhered to and to acknowledge that there might be many different intelligent systems. As Armstrong et al state, the predictions we make can and must be falsifiable and be based on models, not judgment.

Lastly, there seem to exist some double standards when evaluating AI. Humans are known to exhibit many types of biases and limitations, continuously learn throughout life, are taught knowledge accumulated over thousands of years, can interact with the world to test their hypotheses (i.e. are in some sense active learners). AI systems tend to be passive learners (the data is chosen and presented by humans), are usually taught from zero and often using noisy and otherwise imperfect data. In some sense, it should be no surprise that many AI systems make errors that might seem obvious or comical. A human would as well, if they were taught in a similar manner, under similar limitations.

2.2 ‘AI Winter’

It can be seen that there are exaggerated claims, hype, around AI technology today. This high level of optimism has happened before, and once it has died down, periods of ‘AI winters’ have occurred. The relevant question to ask here is whether the current ‘AI summer’ is temporary, can we expect another winter period coming? To answer this, I start with a short overview of the Dartmouth conference.

The Dartmouth conference was a conference in 1956 intended to bring together experts for two months in the hope of making significant advances in problems related to machine intelligence (Armstrong et al, 2016: 46). Although they proposed simple models and outlined further improvements to previously successful problems, there was a fundamental problem – they assumed that AI was a similar problem to those that they were used to solving, leading to overconfidence (Ibid, 48). Their success could have been due to picking low-hanging fruits and further progress could have been much harder. The conclusion is that all the tasks mentioned in creating AI were much harder to accomplish

than they thought. Programming concepts into computers that seem simple to humans can be very difficult, we must avoid anthropomorphising AI. (Ibid, 48-49)

Franklin writes that this Good Old-Fashioned AI (GOFAI) type approach was unsuccessful and symbolic AI went into a 'winter' period (Franklin 2014: 21), which ended in the beginning of the second decade of the 21st century (Ibid, 28). According to Armstrong et al, Hubert Dreyfus pointed out an important pattern in AI research: initial success is followed by big claims, which are then followed by unexpected difficulties and disappointment. Dreyfus highlighted the inherent ambiguity present in human language and syntax, claiming that computers could not deal with them. (Armstrong et al, 2016: 49-50)

Müller states that today many classical AI problems are solved and viewed as trivial. This is largely due to improved resources – processing speed and the ability and availability of large data sets. (Müller 2016: 3). An AI winter could occur again, given the amount of hype present in the media and because progress is tied to funding, which is a political matter. At the same time, it is very clear that useful and powerful AI systems can be created and improved upon, making a long winter unlikely. The potential of another AI winter should not dismiss causes for concern, as existing technology is already powerful, as will be shown in the next section.

2.3 State of the art systems

In this section, I wish to present some examples of narrow AI that have been produced in the last few years.

OpenAI, a non-profit organization, has created a model called **GPT2**, which is able to generate coherent and authentic looking text matching a given prompt. It was trained with text from 8 million web pages and was given the simple objective of predicting the next word, given the words that have occurred before in a given text. It is able to produce text of unprecedented quality, especially about topics for which there is enough data (e.g. Lord of the Rings, Miley Cyrus), however it sometimes fails at world modeling (e.g. writing about fire happening under water), repeats text or switches topics unnaturally. The authors have decided to not publish the full model due to concerns about potential malicious applications (Radford et al 2019). GPT2 is a prime example of what the quantity of data and hardware can achieve.

Google has created **Duplex**, an AI assistant able to synthesize speech and schedule appointments over the phone for the customer (Leviathan 2018). Adobe **VoCo** can imitate

the speech of real people (Jin et al 2017). It is also possible to generate fake videos with deep machine learning methods. These 'deepfakes' "...are challenging for both face recognition systems and existing detection methods, and the further development of face swapping technology will make it even more so." (Korshunov, Marcel 2018: 1).

We are moving beyond using AI merely for the dull, dangerous or dirty jobs. The above systems are all relatively narrow in their domain, yet achieve superhuman performance. At the same time, not all details are known – companies tend to flaunt with achievements when possible. It is actually in the interest of everybody if some of these results are cherry-picked and better systems are not yet available, giving us more time to explore risks and prepare in other ways.

2.4 Reasons behind success and future outlooks

Some reasons can be pointed out which have helped us achieve this level of success. As stated earlier, Richard Sutton has stated that Moore's law and not building domain knowledge into the system are important (Sutton 2019).

While some agnosticism was claimed earlier, then one strategy has so far been proven to be productive and is likely to continue – the imitation of biology and especially the brain. At a first glance, this makes sense, as the brain is the most complex structure we know of and it has produced intelligence. At the same time, as stated earlier, too eager comparisons to humans can be unproductive. Connectionism and neural networks have become integral in the study of intelligence and cognition (Sun 2014: 109). Sun explains connectionism as (Sun 2014: 108-109):

> "... a way of capturing and understanding the mechanisms and processes of cognition through building models using networks of simple, neuron-like processing elements (units), each of which performs simple numerical computations."

Recent breakthroughs in machine translation, speech synthesis and recognition, computer vision have been achieved using these tools. Some algorithms are provided with the capacity for memory or attention (important characteristics of human intelligence), leading to drastic improvements in some research areas. These are examples of characteristics that are sufficiently abstract or high level, meaning that some human comparisons and inspirations can be useful.

Most AI systems are not created with the hope of achieving powerful AGI, but rather to improve the execution of some narrow task. These AI systems are tools, rather than potential persons or animals. Humans desire progress and better tools, better AI

technology is just a form of this desire. Given the massive infrastructure that exists that helps to train, learn, compete, provide or order AI-related services, along with the general desire for progress, it is hard to see this machinery of innovation significantly slowing down anytime soon. While the examples given in the previous section implement neural networks in some way, it is entirely likely that connectionism itself might be insufficient to create powerful AGI and many approaches must be combined (Goertzel 2016 provides a discussion on this topic).

2.5 Limitations

One type of criticism of AI systems consists of comparing them to humans and showing what they cannot yet do or showing how they achieved their results through some unfair advantages. David Danks writes (Danks 2014: 161):

> „Insight and creativity are often held up as a central feature of human learning, if not the central feature. Our learning seems to depend at times on crucial intuitive leaps that we do not seem to be able to explain or predict."

Lake et al (2016) analyze what seems to be an important difference between humans and machines – humans have built in software in the form of intuitive physics and intuitive psychology and humans construct causal models of the world that enable us to learn from mistakes and hypothetical situations. AI systems in video games tend to be better than humans because of their speed and multi-tasking ability, instead of employing novel strategy. A recent result of Google DeepMind's AI in the strategic video game Starcraft 2 was able to beat top professional players, yet closer inspection revealed that it relied on its superhuman speed (Pietikäinen 2018).

This line of criticism can be responded to in two ways. Firstly, as previously argued, AI systems can achieve success and pose risks even if they are different or lacking in some respects. A system might indeed be considered very limited and different in its abilities, yet still achieve its goals (or bring about unintended consequences). If a property is deemed crucial, yet equivalent or near equivalent results are achieved without this property, then perhaps it is not as necessary as thought. This criticism is more against the creation of human-like AGI.

Secondly, there are many research groups working on introducing various human faculties to AI systems – e.g. learning from mistakes (Andrychowicz et al 2018), learning from imagination or dreams (Ha, Schmidhuber 2018), curiosity and internal motivation (Burda et al 2018), one-shot learning (Finn et al 2017). The crucial point is that there exists

an industry demand, but also academic and personal curiosity that drives this type of research. A previously intimately yet vaguely known aspect of human intelligence is defined in a computationally friendly way and experimentation with some clever insights and powerful hardware often leads to some form of improvement of existing results.

The previous discussion tries to turn the argument for pessimism into an argument for optimism. If we assume that some double standards exist (i.e. given what is provided to AI systems and what they can actually achieve, they are doing well) and claim also that AI systems lack some crucial characteristics, yet still outperform humans, then one can only imagine the power these systems could have if these limitations are overcome.

3 Risks

Risks can and should be evaluated systematically, along the axes of source (humans vs AI), degree (from trivial to existential), target (from individuals to the planet as a whole), intent, but also their likelihood of occurrence and the availability of solutions. However, a proper systematic analysis is beyond the scope of this thesis. Instead, I will point out some inherent and some concrete risks with already existing technologies.

This discussion builds on the choice of definitions and technological examples. The main goal is to take the spotlight away from the cluster of superintelligence, AGI, the AI of science fiction, that seem to be a discreet phenomena achieved at some point in the future, and instead show how current technologies are already powerful enough to potentially lead to very bad consequences. The most important question to ask is for the source of risks – is it AI itself, or is it humans?

3.1 AI and unintended consequences

Amodei et al list and analyze five practical research problems related to machine learning systems. The authors believe that AI technologies are likely overwhelmingly beneficial for humanity, but serious thought must be given to potential challenges and risks. Accidents are defined as unintended or harmful behavior that may emerge when errors are committed (Amodei et al, 2016: 1). The authors note that much of the existing discussion has highlighted extreme scenarios, but this might lead to unnecessarily speculative discussions that lack precision. They instead frame accident risk in terms of practical issues with modern machine learning techniques. (Ibid, 2)

They analyze negative unintended side effects, reward hacking (achieving the reward technically, but not in the way it was intended), scalable oversight (can AI find a way to do the right thing without assistance?), safe exploration (we want AI to explore potential solutions, but some of these lead to bad consequences), robustness to distributional shift (how to make AI robust enough to handle the real world as well as it handled the training environment?) (Amodei et al, 2016: 3). The authors argue that problems related to AI systems are not only possible, but likely – depending to a large extent on the people behind constructing the agent.

On a related note, we need measurements to judge performance, but there is always a risk that the measurement becomes a goal in itself. Marilyn Strathern has said "When a measure becomes a target, it ceases to be a good measure." (Strathern 1997: 308). Using AI

systems to measure something might increase our certainty and trust in the measures, especially if there is hype involved. Measure that are poorly chosen or become goals in themselves can also lead to poor consequences.

As a concrete example, consider social media that has algorithms that choose what information the user sees on their news/content feed. These algorithms act as the user's attention, choosing what the user might be interested in. When given just a single goal of keeping the user coming back, they can find very questionable strategies, e.g. keep the user stuck in an information bubble by providing it information it wants to see. This information can be biased or outright false – there is an abundance of misleading ('clickbait') or outright false (fake) news. There are plenty of anecdotal stories about how watching a couple of educational, otherwise innocent videos about Nazi Germany will soon lead to suggestions of videos about wild conspiracy theories. These theories act like viruses of the mind, keeping the user coming back. Surely, we would all be better off if the algorithm cared for more than mere presence on the site?

3.2 Humans and intended consequences

In this section, I wish to give two examples of AI use that lead to intended consequences for the operators of the system, at a potential cost for others.

The first is regarding job automation. On the one hand, automation and the accompanied backlash has been happening for a long time (e.g. Luddites). There has been backlash and fears, yet eventually people have adapted and new jobs have been created. One could make the inductive argument that this will likely continue. There are two problems, however, with this argument. Firstly, if powerful AGI systems are possible and eventually reached, then they can potentially do *everything* a human could do, potentially on a superhuman level. Humans might become unemployable from the point of view of profitability. However, this will probably not happen anytime soon and we collectively have enough time to prepare. The second problem is that corporations currently have no legal or moral obligation to create jobs. If automation yields better profits, why bother employing humans? If the wave of automation comes sufficiently suddenly, then many social problems could arise in a short amount of time, even if even eventually new jobs are created – timing matters.

The second usage is to do with disinformation – false information intended to deceive. The creators of powerful narrow AI systems that generate speech, text or video might view the creation and spread of disinformation as an unintended use of their

systems, but the operators might not. Recent reveals of Russia interfering into US elections have shown how much can be achieved with organized online efforts. One does not need high quality deepfakes, which might be expensive to produce and could potentially be quickly debunked, when a huge quantity of lower quality content can achieve much more. Using AI-driven data analytics to find when, where and what to say, coupled with AI to produce the content itself, is a dangerous combination.

To an extent, unintended consequences seem easier to eliminate or deal with – systems can be thoroughly tested, in simulations and limited real environments. It can also be decided to refrain from the use of a system if the potential dangers are too big or unavoidable. Humans, on the other hand, are harder to control or change. The majority of the world might agree that autonomous weapons should not be pursued, but some major powers ignore this call[4].

3.3 Learning is inherently risky

We want AI systems precisely for their ability to learn, but learning is inherently risky. This is because agents might have knowledge that is shared or private. There can be knowledge about knowledge, meta-knowledge – agent A knows that B knows (or does not know). An intelligent system can (presumably, by definition) learn and potentially arrive at private knowledge. This is knowledge that it can use to its advantage (or its operators). This in itself is already a potential source of risk for whomever this knowledge is used against. However, if the system also has the ability to deceive, it can retain private access and create a false belief in others (potentially with the meta-knowledge of others having a false belief), giving even further power to AI or its operators. Another way to understand the inherent riskiness is to think in terms of known unknowns and unknown unknowns. Learning can increase the amount of unknown unknowns about the system, making it more unpredictable.

The previous paragraph was a discussion about the power and risks associated with private knowledge, but shared knowledge can also be risky (by definition), if it takes away some dearly held belief or value. To illustrate this, consider the relationship between parents and children. If parents teach their children to think independently and be interested in learning, then it is very likely that they will eventually develop some conflicts over factual matters and values. Of course, not every idea a child produces is correct and

4 "Russia, Australia, Israel, United Kingdom and United States [...] are investing significant funds and effort into developing weapons systems with decreasing human control over the critical functions of selecting and engaging targets." Source: stopkillerrobots.org/2019/03/minority-of-states-delay-effort-to-ban-killer-robots/

not every belief a parent has is wrong, but there can be some stubbornness, both in holding or introducing beliefs. Whether changing beliefs and values is a risk that is to be avoided is a different matter.

3.4 Game theory and shared responsibility

One unfortunate aspect of technology in general is that in the long term we are better off when we create systems that benefit everyone and we implement them carefully. However, financial motives might (and indeed have) lead to the relaxation of standards. The recent Boeing 737 Max disasters are good examples of situations where money trumps safety. The accidents seem to have occurred due to the fact that the pilot had no way of overriding a system, which was not developed and scrutinized properly in order to cut down costs[5]. All systems are imperfect or limited, with strengths and weaknesses, and must ultimately be overrideable.

Bostrom analyses how the first inventor of a truly powerful AI system might get a decisive strategic advantage over its competition (Bostrom 2014: chapter 5). This is true also for narrow AI that gives a competitive advantage in a narrow domain for individuals, corporations or countries. AI technology can be very powerful and goal-neutral and hence be a useful instrument for any agent.

Finding and avoiding these 'prisoner dilemma' type situations is important. There is one occurring between humans in the creation of AI technology, but another one takes place between humans and powerful AI systems – it is useful if they require or prefer human cooperation and mutually beneficial outcomes. As noted in the previous section, sometimes humans are hard to persuade. China is a major international power that is heavily experimenting and exploring the potential of AI systems. They also have a history of lack of interest in protecting human rights. It is difficult to apply pressures to China, when states are economically dependent and might face negative consequences. There are no simple solutions.

Miles Brundage analyses research practices and the responsibilities around it, stressing the need for researchers themselves to bear some responsibility (Brundage 2016a: 543). This seems reasonable since shifting blame is easy and a likely solution shares responsibility between the researchers, the companies marketing and selling and eventually the users themselves. Perhaps the bigger conclusion to draw is that since

5 Gregory Travis provides some insight in an opinion piece: https://spectrum.ieee.org/aerospace/aviation/how-the-boeing-737-max-disaster-looks-to-a-software-developer

everyone has a stake in the consequences of AI technology, everyone should bear some (proportional) responsibility.

3.5 The fundamental problem

One fundamental source of problems is the existence of conflicts among humans. We believe and value different things, sometimes complete opposites. *Prima facie*, we can say that some goals and values are fundamental, others are instrumental in achieving some other goal or value. While one could argue against having some instrumental goal on objective grounds (i.e. it does not achieve what it purports to achieve), then resolving disagreements over fundamental values is much more difficult.

To make matters worse, the speed of technological progress is accelerating and has been doing so for a while. The speed of our capacity as individuals, corporations or states to keep up is also accelerating, but at a slower pace. If this continues, then problems accumulate. The concept of Singularity, an intelligence explosion, is the extreme example of this phenomena, but undesirable consequences can happen much earlier.

Any type of recognition and elimination of bottlenecks is welcomed. One example of movement in the right direction is Distill, introducing itself as "A modern medium for presenting research", enabling interactivity in machine learning articles[6]. They are also publishing articles which deal with increasing the transparency of neural networks. Many institutions, especially the education and legislative systems, require heavy modernization in order to keep up.

3.6 A potential partial solution

As stated in the previous section, some goals and values are instrumental. When powerful technology is achieved that enables the attaining of some goal, then we have the chance to take a step back and have a broader look – maybe the same technology renders the current goals or values redundant?

The concept of reward hacking also applies to humans. A business that creates artificial demand for products through aggressive marketing and has AI systems at its disposal could ask that perhaps the same tools could be used to find what customers really lack or desire. A journalist could use AI systems to produce better clickbait content, but also better quality content that reaches its intended audience. It is in our abilities and interests that we collectively foster the latter decisions.

6 More info can be found at https://distill.pub/about/

Conclusion

My aim with this thesis was to analyze artificial intelligence (AI) from the perspective of risks. I defined risks as the potential of losing something of value and AI as intelligence demonstrated by machines.

Much discussion revolved around the definition of intelligence. I noted the importance and difficulty of this task. It is important, because 'AI' carries with it some connotations of seriousness that an 'algorithm' does not. Additionally, misleading conclusions might be reached. The difficulty stems partly from the fact that we use intelligence to define intelligence, making it easy to be too human-centric. I stressed the importance of distinguishing human intelligence from intelligence as such and introduced Alan Turing's idea that when evaluating whether a machine can think should rely on other related concepts. While Turing's own solution (a Turing Test, a test of the ability to deceive) might not be an adequate test of intelligence, is it still useful to measure risks.

I presented and analyzed a definition of universal intelligence by Legg and Hutter, which states that „Intelligence measures an agent's ability to achieve goals in a wide range of environments" (Legg, Hutter, 2016: 12). This definition hints at agency and generality, making it a better fit for artificial general intelligence (AGI) instead of AI systems in general. I continued with arguing for the necessity of separating agency from intelligence, and defined intelligence as the ability to learn. I also stressed the centrality of the notion of predictability. This was followed by a discussion about the potential differences between humans and AI systems, using the ideas of Nick Bostrom. AI systems need not be human-like and Richard Sutton has argued that this requirement has even proven to be harmful in the long run.

Briefly, I considered the impossibility of creating general AI, since the 'no-free-lunch theorem' states that there are no universal learning algorithms. I stated that, similarly to humans, AI systems could be dynamic and adapt to the task at hand. While some generality is possible, I stressed that this requirement can be misleading, since AI systems need not exist in isolation. We live in a world where tasks can be delegated to other agents, where information is shared and accessible via digital channels.

The second part started with some observations regarding predicting and evaluating AI technology. Armstrong et al (2016) note that expert judgments can be very unreliable if they are based on intuition instead of models. Sometimes, AI technology has to deal with

moving goalposts – there are always critics to point out what AI can not yet do. There also seem to exist some double standards. To an extent, it is no surprise that some AI systems perform poorly – humans would as well, when trained under similar limitations.

The concept of AI winters was briefly explored. They are periods of pessimism after short periods of progress and optimism. It was concluded, that another AI winter is possible, but unlikely to last long. Some recent technological achievements were presented, among them GPT2, a text-generating algorithm, followed by a short discussion over some reasons why the current 'AI summer' has been made possible. These include the existence of high amounts of data and computational power, the strategy of imitating biology and the brain, and the persistence of the motivation to create better tools, AI systems being one of them.

One type of criticism against AI systems was presented – AI is often compared to humans and it is pointed out what it cannot do well, or achieves something with some unfair advantage. This was replied to in two parts. Firstly, AI systems can be powerful and dangerous even while being different from humans, even when they are narrow. Secondly, many researchers are exploring and introducing some core characteristics of human intelligence to AI systems, with some success.

The last section built on the given definitions and technological examples. I noted that the first and most important question to ask is about the source of risks – is it AI itself or is it humans? Briefly, unintended consequences of using AI systems were discussed with the work of Amodei et al (2016). I also noted that sometimes measurements become targets and with that become bad measurements, leading to bad consequences.

I also briefly discussed intended consequences – job automation and disinformation being examples that favor some agents (or groups) at the expense of others. This was followed by a short observation why learning itself can be considered risky – it potentially introduces private knowledge which can be exploited. I employed some concepts from game theory, arguing that there exist prisoner dilemma type situations which we must collectively avoid. Shared responsibility is one partial solution to potential AI risks, since everyone potentially has something to lose.

Lastly, I presented a fundamental fact of the human condition – the existence of conflicts in values and goals. These conflicts exist in a world where the progress of technology is accelerating faster than our collective or individual abilities to keep up. Our institutions require modernization, otherwise problems keep accumulating.

As a potential partial solution, I noted that since some goals and values are instrumental, then achieving powerful AI systems enables us to take a step back and potentially re-assess them. Perhaps some of them are redundant, and we can replace negative or zero sum games with positive ones.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. „Concrete problems in AI safety“. 2016. arXiv:1606.06565 [cs.AI]. Accessed January 2018

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., Zaremba, W. „Hindsight Experience Replay“. arXiv:1707.01495 [cs.LG]. Accessed January 2018

Armstrong, S., Sotala, K., ÓhÉigeartaigh, S. S. “The errors, insights and lessons of famous AI predictions – and what they mean for the future”. From *Risks of Artificial Intelligence*. Edited by Müller, V. C. CRC Press: 2016. Pp 34-63

Bostrom, N. „Ethical Issues in Advanced Artificial Intelligence“. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, ed. I.

Bostrom, N. “Superintelligence – Paths, Dangers, Strategies”. Oxford University Press: 2014

Bringsjord, S., Govindarajulu, N. S. "Artificial Intelligence". The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), edited by Zalta, E. N. plato.stanford.edu/archives/fall2018/entries/artificial-intelligence. Accessed January 2018

Brundage, M. „Artificial Intelligence and Responsible Innovation“. From *Fundamental Issues in Artificial Intelligence*. Edited by Müller, V. C. Springer: 2016a. Pp 543-554

Brundage, M. “Limitations and Risks of Machine Ethics”. From *Risks of Artificial Intelligence*. Edited by Müller, V. C. CRC Press: 2016b. Pp 87-114

Burda, Y., Edwards, H., Storkey, A., Klimov, O. „Exploration by Random Network Distillation“. 2018. arXiv:1810.12894 [cs.LG]. Accessed January 2018.

Danks, D. “Learning”. From *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K., Ramsey, W. M. Cambridge University Press: 2014. pp 151-167

Frankish, K., Ramsey, W. M. “Introduction”. From *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K., Ramsey, W. M. Cambridge University Press: 2014. pp

1-11

Franklin, S. „History, motivations, and core themes“. From *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K., Ramsey, W. M. Cambridge University Press: 2014. pp 15-33

Finn, C., Yu, T., Zhang, T., Abbeel, P., Levine, S. “One-Shot Visual Imitation Learning via Meta-Learning”, 2017. arXiv:1709.04905 [cs.LG]. Accessed May 2019

Goertzel, T. “Path to More General Artificial Intelligence”. From *Risks of Artificial Intelligence*. Edited by Müller, V. C. CRC Press: 2016. Pp 69-86

Ha, D., Schmidhuber, J. „World Models“. 2018. arXiv:1803.10122 [cs.LG]. Accessed January 2018

Jin, Z., Mysore, G. J., Diverdi, S., Lu, J., Finkelstein A. “VoCo: Text-based Insertion and Replacement in Audio Narration”. 2017. *ACM Transactions on Graphics* 36(4): Article 96

Korshunov, P., Marcel, S. “DeepFakes: a New Threat to Face Recognition? Assessment and Detection”. 2018. arXiv:1812.08685 [cs.CV]. Accessed May 2019

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., Gershman, S. J. „Building Machines That Learn and Think Like People“. 2016. arXiv:1604.00289 [cs.AI]. Accessed January 2018

Legg, S., Hutter, M. „Universal Intelligence: A Definition of Machine Intelligence“. 2007. arXiv:0712.3329 [cs.AI]. Accessed January 2018

Leviathan, Y. “Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone”, Google AI Blog, May 8, 2018, ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html. Accessed May 2019

McCorduck, P. Machines Who Think, Second edition. Natick, MA: A. K. Peters, Ltd: 2004

Müller, V. C. „New Developments in the Philosophy of AI“. From *Fundamental Issues in Artificial Intelligence*. Edited by Müller, V. C. Springer: 2016. pp 1-4.

Oppy, G., Dowe, D. "The Turing Test", *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). Edited by Zalta, E. N. plato.stanford.edu/archives/spr2019/entries/turing-test/. Accessed May 2019

Pietikäinen, A. „An Analysis On How Deepmind’s Starcraft 2 AI’s Superhuman Speed is

Probably a Band-Aid Fix For The Limitations of Imitation Learning“. medium.com/@aleksipietikinen/an-analysis-on-how-deepminds-starcraft-2-ai-s-superhuman-speed-could-be-a-band-aid-fix-for-the-1702fb8344d6. 2018. Accessed January 2018

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., Sutskever, I. “Better Language Models and Their Implications”, *OpenAI*, 14 February 2019, openai.com/blog/better-language-models/. Accessed May 2019

Ross, D. "Game Theory", *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edited by Zalta, E. N. plato.stanford.edu/archives/spr2019/entries/game-theory. Accessed May 2019

Russell, J. S., Norvig, P. Artificial Intelligence – A Modern Approach, Third Edition. Pearson Education Limited: 2016

Strathern, M. “‘Improving ratings’: audit in the British University system.” *European Review*, Volume 5, Issue 3, pp 305-321

Sun, R. “Connectionism and neural networks”. From *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K., Ramsey, W. M. Cambridge University Press: 2014. pp 108-127.

Resümee: Kuidas tehisintellekt võib olla riskantne?

Käesolevas töös uurin ma tehisintellektiga (TI) seotud riske – kuidas saaks TI viia selleni, et me kaotame midagi, mida me väärtustame? Ma rõhutan, et TI defineerimisel ei tohi me olla liiga inimkesksed ega nõuda üldise intelligentsuse olemasolu. Ka inimesest väga erinevad ja väga kitsad tehissüsteemid võivad olla piisavalt võimsad, et kujutada endast ohtu. Mitmed ohud tulenevad TI töö ettekavatsemata tagajärgedest, ent mitmed aktuaalsed ohud tulenevad TI kui tööriista kasutamisest mänguteoreetilistes null või negatiivse summaga mängudes. Ma rõhutan, et tehisintellekti ei tuleks käsitleda abstraktse tulevikunähtusena, vaid juba praegu eksisteerivana ja analüüsi vajavana. Inimeste väärtused ja eesmärgid on tihti omavahel konfliktis, ning see fakt vajab mingit lahendust, sest tehnoloogia areng kiireneb, võimaldades saavutada järjest kergemini erinevaid eesmärke, ent meie võime muutustega kohaneda ei jõua sellega sammu pidada. Ühe võimaliku osalise lahendusena konfliktide olemasolule võib TI pakkuda meile võimalust taashinnata meie instrumentaalseid eesmärke ja väärtusi – on võimalik, et see, mida me seni soovisime saavutada, pole enam relevantne.

Abstract: How Can Artificial Intelligence Be Risky?

In this thesis, I research risks associated with artificial intelligence (AI) – how could AI lead to us losing something we value? I stress that when defining AI, we cannot be too human-centric or require the existence of general intelligence. Narrow AI systems that are very different from humans can be powerful enough to pose risks. Many risks originate from unintended consequences, yet many actual risks come from using AI as a tool in zero or negative sum games, to use concepts of game theory. I stress that AI should not be treated so much as an abstract phenomena of the future, but as an already existing phenomena that requires analysis. The values and goals of humans are often in conflict and this requires a solution, since the progress of technology is accelerating, enabling more different goals to be achieved, and we are often unable to keep up with this pace. AI can provide a partial solution to the existence of instrumental conflicts by enabling us to reconsider them – it is possible that what we have desired so far is no longer relevant.

Non-exclusive licence to reproduce thesis and make thesis public

I, Taavi Luik

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, “How Can Artificial Intelligence Be Risky?”, supervised by Mats Volberg.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons’ intellectual property rights or rights arising from the personal data protection legislation.

Taavi Luik
02/05/2019