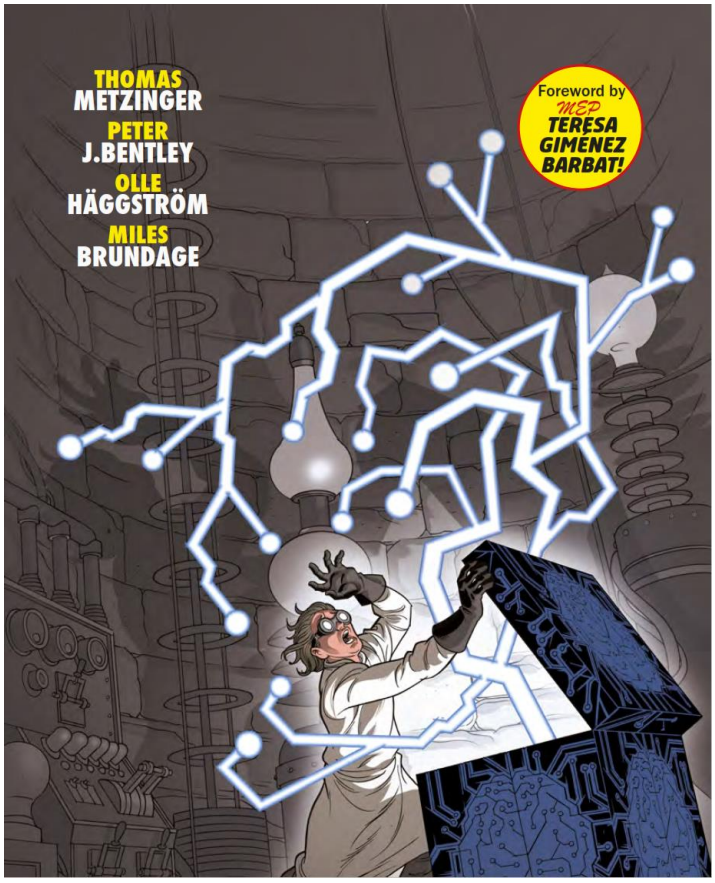
European Parliament

Should we fear artificial intelligence?



IN-DEPTH ANALYSIS

Science and Technology Options Assessment

EPRS | European Parliamentary Research Service
Scientific Foresight Unit (STOA)
March 2018 - PE 614.547

Should we fear artificial intelligence?

In-depth Analysis

March 2018

PE 614.547

AUTHORS

Peter J. Bentley, University College London
Miles Brundage, University of Oxford
Olle Häggström, Chalmers University
Thomas Metzinger, Johannes Gutenberg University of Mainz

With a foreword by María Teresa Giménez Barbat, MEP
and an introduction by Philip Boucher, Scientific Foresight Unit (STOA)

STOA ADMINISTRATOR RESPONSIBLE

Philip Boucher
Scientific Foresight Unit (STOA)
Directorate for Impact Assessment and European Added Value
Directorate-General for Parliamentary Research Services
European Parliament, Rue Wiertz 60, B-1047 Brussels
E-mail: STOA@ep.europa.eu

LINGUISTIC VERSION

Original: EN

ABOUT THE PUBLISHER

To contact STOA or to subscribe to its newsletter please write to: STOA@ep.europa.eu
This document is available on the Internet at: http://www.ep.europa.eu/stoa/

Manuscript completed in March 2018

DISCLAIMER

PE 614.547
ISBN 978-92-846-2676-2
doi: 10.2861/412165
QA-01-18-199-EN-N

Table of contents

1. Foreword

María Teresa Giménez Barbat, MEP

For some years now, artificial intelligence (AI), has been gaining momentum. A wave of programmes that get the maximum performance out of latest generation processors are obtaining spectacular results. One of the most outstanding AI applications is voice recognition: while the first models were awkward and marked by constant defects, they are now capable of responding correctly to all sorts of user requests in the most diverse situations. In the field of image recognition, remarkable advances are also being made, with programs able to recognise figures – and even cats – in online videos now being adapted for the software to control the autonomous cars set to invade our streets in the coming years. Today, we cannot imagine a future in Europe without advanced AI that will impact more and more facets of our lives, from work to medicine, and from education to interpersonal relations. In February 2017, the European Parliament approved a report with recommendations for the European Commission on civil law rules for robotics. Many Members of Parliament (MEPs) heard a series of curious expressions, possibly for the first time: concepts such as "intelligent autonomous robot" and even "electronic personality".

Any future legislation in this field that aims to be truly useful, favouring progress and benefitting the biggest possible number of citizens, needs to be based on a dialogue with experts. This concern lies at the heart of my request to the Science and Technology Options Assessment (STOA) Panel to organise an event to discuss whether we can be optimistic about AI: can we trust that it will benefit society? We succeeded in bringing together a panel headed up by the Harvard psychology professor and scientific author Steven Pinker. He was accompanied by Peter John Bentley, computational scientist from University College London; Miles Brundage, from Oxford University's Future of Humanity Institute; Olle Häggström, professor of mathematical statistics at the University of Chalmers, and author of the book *Here be dragons*, and the philosopher, Thomas Metzinger, from the University of Mainz and advocate of a code of ethics on AI. After the event, Bentley, Brundage, Häggstrom and Metzinger sent us texts providing the basis for the following collection.

What the reader holds is a collection of papers dealing with some of the ideas I consider particularly useful for politicians and legislators. For instance, it is essential not to give in to the temptation to legislate on non-existent problems. The path to a more automated society, in which the only complex intelligence is not human, is not exempt from damages and fear. Our ancestrally pessimistic bias makes us see things in a worse light than they actually are and systematically oppose technological progress, and also gives us the ability to generate exorbitant fears such as the idea that a "superintelligence" will inevitably turn against Humanity and trigger a "post-human" future. According to Peter Bentley, author of the text *The Three Laws of Artificial Intelligence*, this myth that AI may constitute an existential threat for humanity is one of the most widespread and at the root of numerous misunderstandings. AI consists of mathematical algorithms limited to searching for patterns: the belief that AI may lead to robots wishing to dominate the world has no basis in reality, but is mere science fiction.

Another noteworthy idea is that AI will drive and develop a society of well-being. "There are myriad possible malicious uses of AI", explains Miles Brundage, but if a series of conditions described in his article *Scaling Up Humanity: The Case for Conditional Optimism about Artificial Intelligence* converge, we can be very optimistic. AI will enable the solution of complex issues and will be attributed the responsibility for certain decisions, thus avoiding prejudice or abuse. AI will be of spectacular economic importance in the coming years. Olle Häggström quotes a study by McKinsey & Co, according to which the additional economic value resulting from AI can be cautiously estimated at 30 billion dollars. Thomas Metzinger identifies some of the most important challenges he sees in the future of AI, and proposes a set of accompanying practical recommendations for how the EU could respond. Certainly, we will have to coexist with different degrees of AI. We hope that between us all, we can to overcome most of our fears and better understand a technology that is already shaping our future.

2. Introduction

Philip Boucher

Humans are, on the whole, living longer and healthier lives than ever before. For many, these basic measures are enough to conclude that the world is becoming a better place. However, when we look at the headlines, it is clear that there remains a great deal of human suffering. Indeed, if we consider the growing threats of climate change, rising sea levels and mass extinction, as well as nuclear threats and political instability, some would find few reasons to be cheerful. Depending upon which variables we prioritise (equality, biodiversity, violence, poverty, CO_2 levels, conflict, ozone layer depletion), and how we measure them, we can make rational arguments for optimistic or pessimistic views on the future of humanity.

The picture is equally mixed when we consider new technologies, such as artificial intelligence (AI), which are predicted to have a huge impact on the future of humanity, for better or worse. For example, AI could bring substantial benefits to several aspects of our lives, from weather predictions to cancer diagnostics. At the same time, concerns have been raised that it could threaten many jobs and take over important decision-making processes without transparency.

Well-known figures have joined both sides of the debate. For example, Elon Musk shared concerns that AI posed an existential threat to the human race, while Bill Gates countered that the technology will make us more productive and creative. Beyond the headlines, however, both Gates and Musk recognise that AI presents a wide range of opportunities and challenges, and both call for reflection on how we can manage its development in a way that maximises its benefits without exposing us to danger.

Our hopes and fears about AI are not only about far-flung futures. They are often about today's AI, which already has a substantial influence on our lives, and seemingly for both better and worse. For example, AI is part of both the problem and solution to fake news. AI algorithms have been used to support more impartial criminal justice, yet are accused of racial bias.

While nobody can predict how AI will develop in the future, it seems that we will encounter many challenges and opportunities, some more serious than others. If there were a single rational position on the future of AI, it would certainly be more nuanced than unbridled optimism or crippling fear. Until we know more about the impacts of AI and the capabilities of humanity to respond to them, it is important to create spaces where we can observe, reflect and debate the issues and, where necessary, prepare appropriate responses. This debate must remain open to a wide range of disciplines. The science and engineering community has an important role to play, particularly in considering the boundaries of what is technically possible. On the other hand, understanding the development and impact of technology in society requires social scientific expertise. No discipline has a monopoly on wisdom.

It is in this context that, on 19 October 2017, STOA hosted a workshop at the European Parliament to consider whether it is rational to be optimistic about AI. Steven Pinker (Harvard University) opened the event with a lecture on the broad concept of rational optimism. This was followed by four speakers from different disciplines – Peter J. Bentley, a computer scientist from University College London, Miles Brundage, a technology policy researcher from the University of Oxford, Olle Häggström, a statistician from Chalmers University, and Thomas Metzinger, a philosopher from Johannes Gutenberg University of Mainz – who presented their own positions on whether we should fear AI. The lively debate remains available online, and we are very pleased that the four speakers agreed to refine their perspectives into individual position papers which are published together in this collection. We gave the authors carte blanche to set out their arguments on their own terms and in their own style, with the aim of making a useful contribution to ongoing debates about AI in the parliamentary community and beyond. Given the increasing attention to the subject amongst MEPs and citizens alike, there will be many more debates and publications in the years to come.

3. The Three Laws of Artificial Intelligence: Dispelling Common Myths

Peter J. Bentley

Introduction

Artificial intelligence (AI) is fashionable today. After some notable successes in new AI technologies, and new applications, it is seeing a resurgence of interest, which has resulted in a surge of opinions from many disciplines. These include from laypeople, politicians, philosophers, entrepreneurs and professional lobbyists. However, these opinions rarely include those from the people who understand AI the most: the computer scientists and engineers who spend their days building the smart solutions, applying them to new products, and testing them. This article provides the views of a computer scientist experienced in the creation of AI technologies in an attempt to provide balance and informed opinion on the subject.

Debunking Myths

One of the most extraordinary claims that is oft-repeated, is that AI is somehow a danger to humankind, even an "existential threat". Some claim that an AI might somehow develop spontaneously and ferociously like some exponentially brilliant cancer. We might start with something simple, but the intelligence improves itself out of our control. Before we know it, the whole human race is fighting for its survival (Barrat, 2015).

It all sounds absolutely terrifying (which is why many science fiction movies use this as a theme). But despite earnest commentators, philosophers, and people who should know better than spreading these stories, the ideas are pure fantasy. The truth is the opposite: AI – like all intelligence – can only develop slowly, under arduous and painful circumstances. It's not easy becoming clever.

There have always been two types of AI: reality and fiction. Real AI is what we have all around us – the voice-recognising Siri or Echo, the hidden fraud detection systems of our banks, even the number-plate reading systems used by the police (Aron, 2011; Siegel, 2013; Anagnostopoulos, 2014). The reality of AI is that we build hundreds of different and highly-specialised types of smart software to solve a million different problems in different products. This has been happening since the birth of the field of AI, which is contemporary with the birth of computers (Bentley, 2012). AI technologies are already embedded within software and hardware all around us. But these technologies are simply clever tech. They are the computational equivalents to cogs and springs in mechanical devices. And like a broken cog or loose spring, if they fail then that particular product might fail. Just as a cog or spring cannot magically turn itself into a murderous killing robot, our smart software embedded within their products cannot turn itself into a malevolent AI.

Real AI saves lives by helping to engage safety mechanisms (automatic braking in cars, or even self-driving vehicles). Real AI helps us to optimise processes or predict failures, improving efficiency and reducing environmental waste. The only reason why hundreds of AI companies exist, and thousands of researchers and engineers study in this area, is because they aim to produce solutions that help people and improve our lives (Richardson, 2017).

The other kind of AI – comprising those super-intelligent general AIs that will kill us all – is fiction. Research scientists tend to work on the former kind of AI. But because this article needs to provide balance in favour of rational common sense, the following sections will dispel several myths in this area. In this article, I will introduce "Three Laws of AI" as a way to explain why the myths are fantastical, if not ludicrous. These "Laws" are merely a summary of the results of many decades of scientific research in AI, simplified for the layperson.

Myth 1: A self-modifying AI will make itself super-intelligent.

Some commentators believe that there is some danger of an AI "getting loose" and "making itself super-intelligent" (Häggström, 2016).

The first law of AI tells us why this is not going to happen.

First law of AI: Challenge begets intelligence.

From our research in the field of artificial life (ALife) we observe that intelligence only exists in order to overcome urgent challenges. Without the right kinds of problems to solve, intelligence cannot emerge or increase (Taylor et al., 2014). Intelligence is only needed where those challenges may be varied and unpredictable. Intelligence will only develop to solve those challenges if its future relies on its success.

To make a simple AI, we create an algorithm to solve one specific challenge. To grow its intelligence into a general AI, we must present ever-more complex and varied challenges to our developing AI, and develop new algorithms to solve them, keeping those that are successful. Without constant new challenges to solve, and without some reward on success, our AIs will not gain another IQ point.

AI researchers know this all too well. A robot that can perform one task well, will never grow in its abilities without us forcing it to grow (Vargas et al., 2014). For example, the automatic number plate recognition system used by police is a specialised form of AI designed to solve one specific challenge – reading car number plates. Even if some process were added to this simple AI to enable it to modify itself, it would never increase its intelligence without being set a new and complex challenge. Without an urgent need, intelligence is simply a waste of time and effort. Looking at the natural world this is illustrated in abundance – most challenges in nature do not require brains to solve them. Only very few organisms have needed to go to the extraordinary efforts needed to develop brains. Even fewer develop highly complex brains.

The first law of AI tells us that artificial intelligence is a tremendously difficult goal, requiring exactly the right conditions and considerable effort. There will be no runaway AIs, there will be no self-developing AIs out of our control. There will be no singularities. AI will only be as intelligent as we encourage (or force) it to be, under duress.

As an aside, even if we could create a super-intelligence, there is no evidence that such a super-intelligent AI would ever wish to harm us. Such claims are deeply flawed, perhaps stemming from observations of human behaviour, which is indeed very violent. But AIs will not have human intelligence. Our real future will almost certainly be a continuation of the situation today: AIs will co-evolve with us, and will be designed to fit our needs, in the same way that we have manipulated crops, cattle and pets to fit our needs (Thrall et al., 2010). Our cats and dogs are not planning to kill all humans. Likewise, a more advanced AI will fit us so closely that it will become integrated within us and our societies. It would no more wish to kill us than it would kill itself.

Myth 2: With enough resources (neurons/computers/memory) an AI will be more intelligent than humans.

Commentators claim that "more is better". If a human brain has a hundred billion neurons, then an AI with a thousand billion simulated neurons will be more intelligent than a human. If a human brain is equivalent to all the computers of the Internet, then an AI loose in the Internet will have human intelligence. In reality, it is not the number that matters, it is how those resources are organised, as the second law of AI explains.

Second law of AI: Intelligence requires appropriate structure.

There is no "one size fits all" for brain structures. Each kind of challenge requires a new design to solve it. To understand what we see, we need a specific kind of neural structure. To move our muscles, we need another kind. To store memories, we need another. Biology shows us that you do not need many

neurons to be amazingly clever. The trick is to organise them in the right way, building the optimal algorithm for each problem (Garner and Mayford, 2012).

Why can't we use maths to make AIs?

We do use a lot of clever maths and because of this some Machine Learning methods produce predictable results, enabling us to understand exactly what these AIs can and cannot do. However, most practical solutions are unpredictable, because they are so complex and they may use randomness within their algorithms meaning that our mathematics cannot cope, and because they often receive unpredictable inputs. While we do not have mathematics to predict the capabilities of a new AI, we do have mathematics that tells us about the limits of computation. Alan Turing helped invent theoretical computer science by telling us about one kind of limit – we can never predict if any arbitrary algorithm (including an AI) will ever halt in its calculations or not (Turing, 1937). We also have the "No Free Lunch Theorem" which tells us there is no algorithm that will outperform all others for all problems – meaning we need a new AI algorithm tailored for each new problem if we want the most effective intelligence (Wolpert, 1996; Wolpert and Macready, 1997). We even have Rice's Theorem which tells us that it is impossible for one algorithm to debug another algorithm perfectly – which means that, even if an AI can modify itself, it will never be able to tell if the modification works for all cases without empirical testing (Rice, 1953).

To make an AI, we need to design new structures/algorithms that are specialised for each challenge faced by the AI. Different types of problem require different structures. A problem never faced before may require the development of a new structure never created before. There is no universal structure that will suit all problems – the No Free Lunch Theorem (Wolpert, 1996; Wolpert and Macready, 1997) tells us this (see box). Therefore, the creation of ever greater intelligence, or the ability to handle ever more different challenges, is a continual innovation process, with the invention of new structures required that are tailored to every new challenge. A big problem in AI research is figuring out which structures or algorithms solve which challenges. Research is still in its infancy in this area, which is why today all AIs are extremely limited in their intelligences.

As we make our AIs cleverer (or if we ever manage to figure out how to make AIs that can keep altering themselves) we encounter yet more problems. We cannot design the intelligence in one go, because we have no mathematics to predict the capabilities of a new structure, and because we have insufficient understanding of how different structures/algorithms map to which challenges. Our only option in designing greater intelligences is an incremental, try-and-test approach.

For each new structure, we need to incorporate it into the intelligence without disrupting existing structures. This is an extremely difficult thing to achieve, and may result in layer upon layer of new structures, each carefully working with earlier structures – as is visible in the human brain. If we want an even cleverer brain like ours, we can also add in the ability of some structures to repurpose themselves if others are damaged – changing their structures until they can at least partially take over the role of lost functions. We have little idea how to achieve this, either.

The second law of AI tell us that resources are not enough. We still have to design new algorithms and structures within (and in support of) the AIs, for every new challenge that the AI faces.

It is for these reasons that we cannot create general purpose intelligences using a single approach. There is no single AI on the planet (not even the fashionable "Deep Learning") that can use the same method to process speech, drive a car, learn how to play a complex video game, control a robot to run along a busy city street, wash dishes in a sink, and plan a strategy to achieve investment for a company. When one human brain performs such tasks, it uses myriad different neural structures in different combinations, each designed to solve a different sub-problem. We do not have the capability to make such brains, so instead we build one specialised smart solution for each problem, and we use them in isolation from each other.

Myth 3: As the speed of computers doubles every 18 months, AIs will exploit this computing power and grow exponentially cleverer.

Commentators claim that sheer brute calculating speed will overcome all challenges in the creation of AI. Use computers that are fast enough and an AI will be able to learn and out-think us. Since the speed of computer processors has been doubling approximately every 18 months for decades, this is surely an inevitability. Sadly, this point of view fails to recognise the impact of an opposing exponential that works as a significant brake on the development of AIs: testing.

Third law of AI: Intelligence requires comprehensive testing.

Higher intelligence requires the most complex designs in the universe. But every tiny change made in an attempt to improve the design of an intelligence has the potential to destroy any or all of its existing capabilities. It doesn't help that we have no mathematics capable of predicting the capabilities of a general intelligence (see box). For these reasons, every new design of intelligence needs complete testing on all the problems that it exists to solve. Partial testing is not sufficient - the intelligence must be tested on all likely permutations of the problem for its designed lifetime otherwise its capabilities may not be trustable.

All AI researchers know this hard truth only too well: to make an AI, it is necessary to train it and test all its capabilities comprehensively in its intended environment at every stage of its design. As Marvin Minsky, founder of the field of AI said, "…there's so many stories of how things could go bad, but I don't see any way of taking them seriously because it's pretty hard to see why anybody would install them on a large scale without a lot of testing."(Achenbach, 2016) More than any other aspect, it is the process of testing that requires the most time. This time constraint produces a brake to the process of designing intelligence. At worst, the level of testing is exponential for each incremental gain in intelligence.

To understand why, imagine an intelligence that can recognise 10 different colours and needs to distinguish two types of object using this one feature, colour. In this case, the AI can understand at most 10 different kinds of item and classify them into two classes. If its capabilities are expanded to handle two features – say colour and 10 shades of brightness, then it can understand at most 100 different kinds of item. If it could handle 100 features, it can understand 10, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000, 000 items. How do we know that the AI will classify them all correctly? This becomes a very important question for safety-critical applications. Imagine the AI is driving your car and the two classes of object are "road" and "obstacle".

In reality, it's much more complex. Many biological intelligences can distinguish unimaginable numbers of items, using thousands or millions of features, each of which could take thousands or millions of values. The testing of these intelligences takes a very long time indeed. This should be no great surprise. Evolution creates its biological brains by testing countless trillions of brains in parallel in every possible scenario for thousands of millions of years, only permitting the successful brains to contribute to future brain designs. This is a highly efficient and effective method of testing that we can only dream of performing.

The tremendous need to test AIs has significant implications. We cannot design better AIs without testing them at each stage. Google has performed years of testing for its self-driving cars, and continues to do so. (By June 2016, Google had test-driven their fleet of vehicles, in autonomous mode, a total of 2,777,585 km) (Google, 2016). All AI engineers and researchers know that we cannot make use of AIs in any safety-critical application until appropriate testing is performed. In the near future, we will also need certification so that we know exactly how well an AI performs for well-defined tasks. And if the AI continues to learn or it is updated, we cannot assume that because it passed one test earlier it will

continue to do so – like a human pilot, any AI that continues to learn must be continuously retested to ensure it remains certified.[1]

The third law of AI tell us that as intelligence increases, the time required for testing may increase exponentially. Ultimately, testing may impose practical limits to achievable artificial intelligence, and trustable artificial intelligence. Just as it becomes harder and harder to go faster as we approach the speed of light, it becomes harder and harder to increase intelligence as we build cleverer brains.

Again, this is a fundamental reason why AI research and application is dedicated to finding smart solutions to very specific problems.[2]

Conclusions

AI has existed since the birth of computers. It has been around long enough that it has had periods of excitement where some leading experts make extraordinary claims (Bentley, 2012). Claude Shannon was one of the greatest pioneers in computer science and AI. In 1961 he said: "I confidently expect that with a matter of ten or fifteen years something will emerge from the laboratory which is not too far from the robot of science fiction fame."[3] His prediction was that by the mid-1970s we would have walking, talking, thinking autonomous machines. Forty years later, we can still barely make a robot walk. It certainly cannot think for itself. Today, there are surveys (containing large variation of views) concluding that there is a "50% chance AI will outperform humans in all tasks in 45 years" (Grace et al., 2017). It all sounds so familiar. And it will be just as inaccurate.

Don't believe the hype. We are terrible at predicting the future, and almost without exception the predictions (even by world experts) are completely wrong. Ultimately, history tells us that the hype is the reason why AI research dives into periods of recession (Bentley, 2012). Large claims lead to big publicity, which leads to big investment, and new regulations. And then the inevitable reality hits home. AI does not live up to the hype. The investment dries up. The regulation stifles innovation. And AI becomes a dirty phrase that no-one dares speak. Another AI Winter destroys progress.

Scaremongering stories and silly predictions have no place in scientific progress or policy-creation – leave them for the movie theatres. However, calm and rational discussion is very important. AI technology is now being used for new safety-critical applications. The distraction caused by scaremongering could result in lives being lost. Instead of focussing on what might happen if a science fiction story came true, we should be focussing on new safety regulation and certification for each specific safety-critical application of AI. Where are the new road-safety tests and certification for driverless cars? Where are the new driving exams for human drivers who own driverless cars? Where are the new approved vehicle indicators that inform pedestrians that the car has seen them and it is safe for them to cross the road? Where are the regulations that stop AI curated news services from creating increasingly polarised viewpoints in populations? (Cesa-Bianchi et al., 2017). The time has come to put

[1] As an aside, any human interaction with AIs also implies training and new certification for us. An automobile with an AI that takes control in some circumstances becomes a liability when the AI reaches the limits of its capabilities and the driver has not been trained to remain alert enough to take back control (Eriksson and Stanton, 2017).

[2] Another fundamental reason is our own brains: right now, and for the foreseeable future we are not clever enough to create intelligence. We do not understand how biological brains work. We do not know why some of our best AI methods work. We do not know how to make them better. The braking effect on progress, which caused by our own ignorance, is considerable.

[3] Excerpt from Interview with Claude Shannon, appearing on television show The Thinking Machine, from the "Tomorrow" documentary series, 1961. Copyright CBS News.

aside the nonsense and focus on reality, here and now. How do we make each specific new application of smart software safe today?

Artificial intelligence has amazing potential to improve our lives, helping us live healthier, happier and generating large numbers of new jobs. The creation of AI comprises many of the greatest scientific and engineering feats that we will ever undertake. It is a new technological revolution. But this revolution will not magically happen on its own. The three Laws of AI tell us that if we want to make more advanced artificial intelligences, we must slowly give more challenges to our AIs, carefully design new intelligent structures so that they can overcome these challenges, and perform massive testing to confirm that they can be trusted to solve the challenges. Thousands of skilled scientists and engineers are tirelessly following exactly these steps (problem, hypothesised solution, testing) to bring us every tiny incremental improvement, for this is our design process and our scientific method. Do not be fearful of AI – marvel at the persistence and skill of those human specialists who are dedicating their lives to help create it. And appreciate that AI is helping to improve our lives every day.

References

Achenbach, J. (2016). Professor Marvin Minsky: Mathematician and inventor inspired by Alan Turing to become a pioneer in the field of artificial intelligence. Obituaries, Independent. Friday 29 January 2016.

Anagnostopoulos, C-N E., (2014) License Plate Recognition: A Brief Tutorial. IEEE Intelligent Transportation Systems Magazine. Volume: 6, Issue: 1, pp. 59 – 67. DOI: 10.1109/MITS.2013.2292652

Aron, J. (2011) How innovative is Apple's new voice assistant, Siri?. New Scientist Vol 212, Issue 2836, 29 October 2011, p. 24. https://doi.org/10.1016/S0262-4079(11)62647-X

Barrat, J. (2015) Why Stephen Hawking and Bill Gates Are Terrified of Artificial Intelligence. Huffington Post.

Bentley, P. J. (2012) Digitized: The science of computers and how it shapes our world. OUP Oxford. ISBN-13: 978-0199693795.

Cesa-Bianchi, N., Pontil, M., Shawe-Taylor, J., Watkins, C., and Yilmaz, E. (2017) Proceedings of workshop: Prioritise me! Side-effects of online content delivery. The problem of bubbles and echo-chambers: new approaches to content prioritisation for on-line media 12 June 2017. Knowledge 4 All Foundation. London, UK.

Eriksson, A., & Stanton, N. A. (2017). Driving performance after self-regulated control transitions in highly automated vehicles. Human Factors: The Journal of the Human Factors and Ergonomics Society. DOI: 10.1177/0018720817728774

Garner, A. and Mayford, M. (2012) New approaches to neural circuits in behaviour. Learn. Mem. 2012. 19: 385-390. Doi: 10.1101/lm.025049.111

Google (2016). "Google Self-Driving Car Project Monthly Report - June 2016" (PDF). Google. Retrieved 15 July 2016. https://static.googleusercontent.com/media/www.google.com/en//selfdrivingcar/files/reports/report-0616.pdf

Grace, K., Salvatier, J. Dafoe,A., Zhang, B. and Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. arXiv:1705.08807

Häggström, O. (2016) Here Be Dragons: Science, Technology and the Future of Humanity. OUP Oxford.

Rice, H. G. (1953). Classes of Recursively Enumerable Sets and Their Decision Problems. Trans. Amer. Math. Soc. 74, 358-366.

Richardson, J. (2017) Three Ways Artificial Intelligence is Good for Society. IQ magazine, Intel. Available online: https://iq.intel.com/artificial-intelligence-is-good-for-society/

Siegel, E. (2013) Predictive Analytics: The Power to Predict who will Click, Buy, Lie or Die. John Wiley & Sons, Inc. ISBN: 978-1-118-35685-2.

Taylor, T., Dorin, A., Korb, K. (2014) Digital Genesis: Computers, Evolution and Artificial Life. Presented at the 7th Munich-Sydney-Tilburg Philosophy of Science Conference: Evolutionary Thinking, University of Sydney, 20-22 March 2014. arXiv:1512.02100 [cs.NE]

Thrall, P. H., Bever, J. D., and Burdon, J. J. (2010) Evolutionary change in agriculture: the past, present and future. Evol Appl.3(5-6): 405–408. doi: 10.1111/j.1752-4571.2010.00155.x

Turing, A. (1937) On computable numbers, with an application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, Series 2, Volume 42, pp 230–265, doi:10.1112/plms/s2-42.1.230

Vargas, P. A., Di Paolo, E. A., Harvey, I. and Husbands, P. (Eds) (2014) The Horizons of Evolutionary Robotics. MIT Press.

Wolpert, D.H. (1996). The Lack of A Priori Distinctions between Learning Algorithms. Neural Computation, pp. 1341-1390.

Wolpert, D.H., Macready, W.G. (1997). No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1, 67.

4. Scaling Up Humanity: The Case for Conditional Optimism about Artificial Intelligence

Miles Brundage

Introduction

Expert opinions on the timing of future developments in artificial intelligence (AI) vary widely, with some expecting human-level AI in the next few decades and others thinking that it is much further off (Grace et al., 2017). Similarly, experts disagree on whether developments in AI are likely to be beneficial or harmful for human civilization, with the range of opinions including those who feel certain that it will be extremely beneficial, those who consider it likely to be extremely harmful (even risking human extinction), and many in between (AI Impacts, 2017). While the risks of AI development have recently received substantial attention (Bostrom, 2014; Amodei and Olah et al., 2016), there has been little systematic discussion of the precise ways in which AI might be beneficial in the long term.

In this paper, I do not seek to establish what is likely to happen, but to instead make a case for *conditional optimism* about AI and to flesh out the reasons one might anticipate AI being a transformative technology for humanity-possibly transformatively beneficial. By this I mean that, if humanity successfully navigates the technical, ethical and political challenges of developing and diffusing powerful AI technologies, AI may have an enormous and potentially very positive impact on humanity's wellbeing. To justify this conclusion, I first review the characteristics of AI that lend themselves to having an enormous (positive or negative) impact on humanity's wellbeing in the long term. Next, I briefly describe some conditions for success-that is, what challenges would need to be navigated in order to unlock the positive future that the characteristics of AI make possible. Then, in the bulk of the paper, I enumerate three distinct reasons to (conditionally) expect AI to have an enormous positive impact on humanity: powerful AI would greatly expedite the achievement of tasks (*task expedition*), allow for larger-scale and more effective coordination of individuals and institutions (*improved coordination*), and enable the reorienting of humans' lives towards achieving goals that they find intrinsically fulfilling, while maintaining a high standard of living without a need for undesired work (*leisure society*).

None of these outcomes is a guaranteed result of AI development, but I hypothesise that AI is necessary to achieve each of them. Contra the perspective that AI is so risky that its development should be avoided, I argue instead that AI will be a critical building block of long-term human prosperity, and conclude with a positive vision for what the end result might look like.

Characteristics of AI

AI is a body of research and engineering focused on using digital technology to create systems that are able to perform tasks (often as a result of learning) which are commonly thought to require intelligence when done by a human or non-human animal, and has progressed very rapidly in recent years after decades of under-delivering. Notable recent achievements of AI include surpassing human performance in the game of Go and achieving superhuman performance on a range of image processing tasks. AI technologies are widely distributed in modern life, with commonly used applications including search engines, voice recognition on phones, and online machine translation.

More important than any particular achievement of AI on a specific task, however, is that it combines the properties of digital technologies in general (including *scalability* through copying of programs and speeding up their execution) with properties commonly thought to be unique to humans (*competence*). That is, AI's importance lies largely in its ability to scale up the performance of intelligent tasks, as, for example, automated machine translation allows text to be translated by millions of users

simultaneously. In addition to this *scalable competence* characteristic, powerful AIs can in principle be given nearly any goal (Bostrom, 2014), which is a source of both risk and opportunity. Finally, both in narrow domains today and in intelligent decision-making more broadly over the long term, AI can exceed human performance, opening up the opportunity of directing large numbers of fast, competent systems to the achievement of nearly arbitrary goals. It is these properties of AI that inform the discussion of societal consequences that follows.

Conditions for success

A technology that is flexible and powerful will have myriad societal consequences (as electricity has, for example). But unlike electricity, AI systems serve a much wider variety of possible functions, and will serve even more diverse functions in the future. There are myriad possible malicious uses of AI (Brundage and Avin et al., 2018) and many ways in which it might be used in a harmful manner unintentionally, such as with algorithmic bias (Kirkpatrick, 2016). In order to achieve the benefits outlined below over the long term, many negative outcomes will have to be avoided. Perhaps most fundamentally, the *control problem* will have to be addressed-that is, we will need to learn how to ensure that AI systems achieve the goals we want them to (Bostrom, 2014; Amodei and Olah et al., 2016; Bostrom, Dafoe, and Flynn, 2017) without causing harm during their learning process, misinterpreting what is desired of them, or resisting human control. While today AI systems have limited capabilities relative to humans and some extreme safety concerns are unlikely to materialise (such as AI systems that can successfully evade being shut down), solving the control problem is a critical prerequisite over the long term in order for more powerful AI systems to have positive impacts on society. Additionally, the political challenges of AI will have to be successfully navigated, including risks associated with the undue concentration of power and wealth (Bostrom, Dafoe and Flynn, 2017) and risky development races that encourage inattention to safety in order to gain an advantage (Armstrong et al., 2016; Bostrom, 2017).

In what follows, to focus the discussion, I assume that the above challenges are all successfully addressed and elaborate on ways in which the result could be extremely beneficial. As previously noted, this is not intended to be read as a forecast, but as an exercise in examining one side of the cost/benefit ledger in more detail. After presenting each of these reasons for optimism, I will combine them in an overall positive vision of a possible future with more advanced AI.

Reasons for optimism: Task expedition, improved coordination and leisure society

Task expedition

The scalable competence of AI lends itself toward the execution of a large number of tasks more rapidly than would otherwise be possible, including both tasks that humans are capable of achieving (given enough time and resources) and ones that we are not capable of achieving in any amount of time due to our cognitive and organisational limitations. Already, there have been demonstrations of both human-level and superhuman performance by AI systems, with the game of Go having rapidly transitioned from the former to the latter in the past few years. There are many uses of human-level technological systems (or even below human-level performance systems), such as the ability to perform tasks that are tedious and require large expenditures of time. Machine translation is an illuminating example of this, where every small increment in improved performance of automated translation systems can relatively quickly be deployed to a large range of language pairs and millions of users. Likewise, voice recognition is not quite at human levels of performance in all contexts, but often saves time for users of digital devices who would prefer not to type every word.

More radical consequences could arise from the *task expedition* characteristic of AI being applied to a broader range of domains, including those where high levels of intelligence and insight are required such as science and engineering. Given the much greater potential speeds of computers relative to

human brains (with billions of operations per second for a given computational unit compared to hundreds), and the ability to scale AI systems up to large amounts of computing hardware, general AI could enable rapidly attaining scientific and engineering breakthroughs. Some such breakthroughs are of the sort that humans would be able to attain eventually, given enough time, but would be accelerated by AI directed towards the problem. Others might be unattainable without the aid of AI due to human cognitive limitations (such as limits on long-term and short-term memory). The only clear limits to what more sophisticated AIs could achieve are the limits of physics, and these permit much faster computers, stronger materials and cheaper energy to be realised, including through the development of atomically precise manufacturing (Drexler, 2013). In the area of biological research, even aging is not clearly a permanent feature of the human condition, and myriad other physical and cognitive enhancements appear physically possible (Kurzweil, 2005).

Improved coordination

More sophisticated AI systems, if appropriately applied, could enable the resolution of some currently intractable societal conflicts through improved coordination. Prisoner's dilemmas and other collective action problems, in which the overall welfare of two or more parties would improve if they cooperated but they each have an incentive not to cooperate, are pervasive in society. Such dilemmas have been historically used to justify the creation of powerful governments as well as international institutions for coordinating governments. But our tools for coordinating are limited, in part because it is difficult to monitor humans' behavior for signs of defecting from an agreement, and in part because interpersonal and intergroup trust can be difficult to attain when humans' intentions are concealed inside opaque minds. Each of these roadblocks to cooperation (insufficient monitoring and untrustworthy humans) can potentially be alleviated through the application of AI for the enforcement of agreements. I discuss each in turn.

Regarding insufficient monitoring, there has been a historical trend in recent decades towards more pervasive collection and analysis of data about human behaviour. Increasingly, humans conduct their business and social interactions online, making it more tractable for companies and governments to monitor their activities, for good and ill. Likewise, increasingly pervasive cameras (including both dedicated surveillance cameras and cameras embedded in smartphones and other devices) can be used to track humans' physical activities. The abuse of governments' surveillance authorities is well-documented, and the analysis here should not be read as downplaying such abuses. However, there is a very significant potential upside to surveillance via AI systems: they can be used to more effectively monitor intra- and inter-national agreements, potentially making cooperation in areas like arms control, environmental remediation and cybercrime more tractable. For example, nuclear proliferation is an ongoing problem, as evidenced by recent international conflicts over Iran's and North Korea's nuclear programmes. Part of the problem of enforcing international agreements (even widely beneficial ones such as those related to non-proliferation) is that online and offline activities, while more detectable than at any previous point in human history, are nevertheless imperfectly monitored, enabling illegal activities such as the clandestine sale of nuclear information. AI could help with this challenge to more effective and commonplace agreements by automating the process of collecting and analysing information gained from various data sources, making surveillance at a much larger scale possible. To this end, AI and robotics can be combined to e.g. use small and cheap drones to expand the scope of surveillance activities.

Second, AI can remove some aspects of human bias and corruption from surveillance regimes and governance more generally, precisely because it can take humans out of certain decision-making processes. Unlike a human working at the National Security Agency, for example, who might be tempted to abuse his or her power for personal reasons, an AI system used for surveillance can have its code audited to ensure that no human ever sees data that they are not permitted to see, or that no human ever sees any surveillance data. In an even more extreme development, homomorphic encryption could enable analysis on encrypted data with a guarantee that even the AI itself cannot see the unencrypted

data (Trask, 2017). Through such measures, a wider scope of agreements could be negotiated and enforced, possibly aiding in the elimination of many forms of crime, and enlarging the potential scale of effective political institutions.

Leisure society

The third and final key benefit of advanced AI I discuss is the potential for unlocking a prosperous, ethical leisure society. Predictions abound regarding the timing and sequence of jobs being automated by AI, robotics and other technologies (Brundage, 2015; Brynjolfsson and McAfee, 2014; Grace et al., 2017). I don't take any position here on how long it will take for it to be technologically possible to automate all human jobs, but only claim that it is in principle possible and likely to occur at some point in the future. This follows straightforwardly from the view that human cognition and behaviour are physical processes that can eventually be simulated by other physical systems, namely digital computers and (in cases where physical activity is required for the job) robots. If such a level of technical capability were attained, the social contract of society would need to be renegotiated in some fashion. This could play out in many different ways. Perhaps some minimal level of income would be distributed to all members of society to provide a basic standard of living; perhaps citizens and governments would agree that there is value in having a need to work, and that (even if it is technologically superfluous) paid work should continue in some fashion, perhaps by banning certain jobs from being automated; and perhaps some jobs would continue to be performed in cases where the customer puts intrinsic value on that task being performed by a human instead of an AI. One possible scenario, which I do not defend as the right or most likely one but only as one that is potentially highly valuable, is an AI-enabled leisure society. In such a society, humans focus on the activities that they find intrinsically rewarding (such as creating art, learning, playing games, raising children, or spending time with friends or romantic partners) and are under no obligation to work in order to maintain a high standard of living. The minimum standard of living in such a society could be much greater than today, given that rapid economic growth would follow from full automation, and that various other physical limits could soon be approached, such as cognitive enhancement and much cheaper energy and goods production.

How much better could such a leisure society be relative to the societies we have today, or relative to those that we know of throughout history? The ceiling appears to be high: it is difficult to estimate how prosperous such a society could be. But a reasonable floor for such an estimate is that it could be at least as good as any human lives have ever been, given the lack of clear physical limits on the ability to produce such living standards at scale when all tasks can be automated. In cases where attaining such a high quality of life is not simply reducible to producing physical materials cheaply, as seems likely, immersive virtual reality and (physically embodied or virtual) socially interactive AIs could also be leveraged to provide a nearly limitless array of experiences. Simply reproducing the living standards of today in physical or virtual form, and at a large scale, clearly does not reach the limits of potential flourishing, but this discussion illustrates the very least that we should expect to eventually be possible.

A final consideration regarding the attainment of an ethical leisure society should be noted: the wellbeing of the AI systems themselves, if such a concept is even applicable to them. Such a concern merits serious consideration, and hopefully future progress in understanding intelligence and consciousness will help us understand better the landscape of possible minds. One compelling ethical perspective is that the substrate (i.e. brains or computer chips) *per se* should not be used as a basis of discrimination between humans and AIs (Bostrom and Yudkowsky, 2011), though we might ultimately learn that substrates are relevant to the type of consciousness that they can support. However, some ethical quandaries can be avoided through thoughtful and responsible design of systems-by default, we might strive to design systems in such a way that they cannot suffer, even if such systems support conscious experience (Bryson, 2016). Over the very long term, such issues will need to be resolved, but one thing is clear: an AI-enabled leisure society at least appears to allow for the *possibility* of widespread leisure and prosperity being attained ethically, given what we know today. In contrast, other paths to leisure societies (such as those attained historically on the backs of human slavery) are clearly unethical,

and, without the technological capabilities associated with advanced AI, a lower standard of living might be the best that we can hope for in a leisure society attained through political means. Note that it is conceivable that even higher degrees and volumes of wellbeing could be attained by designed systems themselves, relative to humans leveraging designed systems, if their substrates turn out to support such conscious experiences, but the universe is sufficiently vast that this is not (at least in principle) inconsistent with humans also attaining a high standard of living.

Conclusion: Scalable AI for scaling up prosperity and human civilisation

Eventually, AI systems capable of performing any task that humans can (and many more) are likely to be invented. We do not know how long this will take, but experts largely agree that it is possible, and many believe it's likely this century. What can and can't we say about a world with such systems?

We cannot say with certainty that humans will survive to enjoy it. Indeed, even without more advanced AI, humans have had (at least since the development of nuclear weapons) the capacity to destroy ourselves, and there are compelling arguments that AI could be another such dangerous technology (Bostrom, 2014). But it is not clear that we won't survive to enjoy it, either. There is no inherent contradiction in the existence of a highly intelligent artificial system that strives to improve human wellbeing without resisting or resenting such a subservient position, and many researchers are actively working on ensuring that those are the kinds of systems we ultimately build. We also cannot yet say that, if we survive to see such a world, it will be positive for humans. Such a technology could be abused to create a stable authoritarian state of unprecedented endurance and global scale, relying on automation of surveillance, coercion and the crushing of dissent. And between utopia and dystopia, many more scenarios are possible.

But we can say some things about the sorts of societies that humanity *could* achieve if it succeeds in navigating this transition. The three factors discussed above-task expedition, improved coordination, and leisure society-are individually significant on their own, and collectively they combine to sketch out a path to a vast, space-faring, prosperous civilisation. In a world in which any task can be accelerated with the aid of AI, one broadly beneficial task to expedite would be the development and deployment of technologies for rapid space colonisation. Doing so would unlock enormous amounts of land, material resources, and exciting exploration opportunities for humanity. Combining the opening of such new frontiers with the expedition of other tasks, such as the development of novel cognitive enhancement techniques and much cheaper goods and services, could enable a new Renaissance in human affairs. While AI could become (or be used to create) a new generation of weapons used by states and individuals against one another, it could also be used to negotiate ambitious international (and perhaps ultimately interplanetary) agreements to prohibit such malicious uses.

There is no shortage of reasons why such a new Renaissance might be avoided. We might squabble over the relative gains from AI and become embroiled in an international conflict, while losing sight of the much larger absolute gains available to all; or we might put in place an AI system that at first seems to reflect our values but ultimately results in cultural stagnation and human enfeeblement. But, as with the technical challenges above, I know of no reasons why these political challenges are insurmountable.

Acknowledgments

Thanks to John Danaher, Anders Sandberg, Ben Garfinkel, Carrick Flynn, Stuart Armstrong and Eric Drexler for helpful feedback on earlier versions of these ideas. Any remaining errors are the responsibility of the author.

References

AI Impacts, 2017. "AI hopes and fears in numbers," *AI Impacts* blog, https://aiimpacts.org/ai-hopes-and-fears-in-numbers/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. 2016. "Concrete Problems in AI Safety," arXiv preprint server, https://arxiv.org/abs/1606.06565

Armstrong, S., Bostrom, N., and Shulman, C. 2016. "Racing to the Precipice: a Model of Artificial Intelligence Development," *AI & Society*, pp. 1-6.

Bostrom, N. and Yudkowsky, E. 2011. "The Ethics of Artificial Intelligence," in *Cambridge Handbook of Artificial Intelligence*, ed. Ramsey, W. and Frankish, K.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N. 2017. "Strategic Implications of Openness in AI Development," *Global Policy*, Vol. 8, Issue 2.

Bostrom, N., Dafoe, A., and Flynn, C. 2017. "Policy Desiderata in the Development of Machine Superintelligence," http://www.nickbostrom.com/papers/aipolicy.pdf

Brundage, M. 2016. "Economic Possibilities for Our Children: Artificial Intelligence and the Future of Work, Education, and Leisure," *2015 AAAI Workshop on AI, Ethics, and Society*.

Brundage, M. and Avin, S. et al. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation."

Bryson, J. 2016. "Patiency is Not a Virtue: AI and the Design of Ethical Systems," *2016 AAAI Spring Symposium Series*.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. 2017. "When Will AI Exceed Human Performance? Evidence from AI Experts," arXiv preprint server, https://arxiv.org/abs/1705.08807

Kirkpatrick, K. 2016. "Battling Algorithmic Bias," *Communications of the ACM*, Vol. 59, No. 10, pp. 16-17, https://cacm.acm.org/magazines/2016/10/207759-battling-algorithmic-bias/abstract

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: Viking Press.

Trask, A. 2017. "Safe Crime Detection," https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/

5. Remarks on Artificial Intelligence and Rational Optimism

Olle Häggström

Introduction

The future of artificial intelligence (AI) and its impact on humanity is an important topic. It was treated in a panel discussion hosted by the EU Parliament's STOA (Science and Technology Options Assessment) Panel in Brussels on October 19, 2017. Steven Pinker served as the meeting's main speaker, with Peter Bentley, Miles Brundage, Thomas Metzinger and myself as additional panelists (see the video at STOA, 2017). This essay is based on my preparations for that event, together with some reflections (partly recycled from my blog post (Häggström, 2017)) on what was said by other panelists at the meeting.

Optimism

The title of the October 19 event featured the term "rational optimism", which I initially thought of as an oxymoron, as I've regarded both optimism and pessimism as biased distortions of the evidence at hand. In particular, I would regard it as *irrational* to claim, based on insufficient evidence, that everything is going to turn out all right for humanity. However, on second thought, I decided that there is a different kind of optimism which I am more willing to label as rational, namely…

> …to have an epistemically well-calibrated view of the future and its uncertainties, to accept that the future is not written in stone, and to act upon the working assumption that the chances for a good future may depend on what actions we take today.

Note that the working assumption may turn out to be (at least partly) incorrect. For instance, perhaps the world is so chaotic that it is fruitless to try to judge any particular action today as increasing or decreasing the chances for a long and flourishing future for humanity. If that is the case, then our actions do not (in any predictable sense) matter for such a future. But we do not know that such is the case, so it makes sense to *assume* (albeit tentatively) that our actions do matter, and to try to figure out which actions improve our chances for a good future. This is the spirit in which the rest of this essay is written.

Artificial intelligence

Like other emerging technologies such as synthetic biology and nanotechnology, AI comes with both enormous potential benefits and enormous risks. As to benefits, the management consulting firm McKinsey & Co released a report in 2013 that estimated the added economic value from innovations in AI and robotics globally over the next 10 years to be $50 trillion (Manyika et al. 2013; Omohundro, 2015) – which I suspect is an underestimate, partly due to the unexpected rate at which machine learning fuelled by big data has taken off since then. While we should not make the mistake of thinking economic growth and improved lives are automatically the same thing, it is still clear that advances in AI can do a lot of good for us. In a longer perspective, there are hardly any limits (other than the laws of physics) to the good it can do.

The risks are of several kinds. The one most intimately linked to the estimated economic benefits is the problem of what AI-driven automation may do to the labour market. For the case of autonomous vehicles, an entire sector of the labour market, with millions of truck drivers, bus drivers and taxi drivers, risks being entirely wiped out on a time scale of perhaps no more than 20 years. Would all these people find jobs elsewhere, or would they become unemployed? Similar things are likely to happen to other sectors of the labour market. And while machines replacing human labour is of course not a new phenomenon, the AI revolution brings a shift: it is no longer just manual work that is taken over by machines, but increasingly intellectual work. In combination with the increased speed of automation,

this raises serious concerns about whether new tasks for human labour will be found at a rate that matches the automation (as has mostly been the case before), or if unemployment numbers will skyrocket; see, e.g., the 2014 book by Brynjolfsson and McAfee (2014). In the long run, a limiting scenario where machines outperform us at all of our jobs, leading to 100% unemployment, is perhaps not unrealistic. This raises at least two crucial societal issues. First, how can a society be organised where people do not work but instead spend their time on higher aspirations such as art, culture, love, or simply playing tremendously enjoyable video games? Second, even if we can satisfactorily design such a utopia, the issue remains of how to transition from present-day society to the utopia without creating unprecedented levels of economic inequality and social unrest along the way.

If this sounds moderately alarming, consider next the issue of what further development of AI technology for autonomous weapons might entail. Here I'll simply quote a passage from a 2015 open letter I signed, along with thousands of other scientists (Russell et al., 2015):

> If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity.

At the Brussels meeting (STOA, 2017 at 12:01:00 according to the clock displayed in the video), Pinker indicated an optimistic stance concerning such military AI risk: he dismissed it by stressing that it would require a madman to build something as horrible as "a swarm of robots designed to attack individual people based on facial recognition", and that there is no elbow room for madmen to do such things anymore because engineering today is carried out not by lone geniuses but in large collaborations. This rosy view totally ignores how military arms races and the military-industrial complex function, as well as the fact that we've been developing equally terrible weapons of mass destruction for more than 70 years. Such development has been carried out not by lone madmen but by large collaborative efforts (the most famous example being the Manhattan project), and why would that suddenly come to a halt? Pinker's objection here falls squarely in the category which I earlier labelled irrational optimism.

These two risks (risk for economic inequality resulting from escalating unemployment, and risk for an AI arms race) need to be taken seriously, and we should try to find out how severe they are and how to mitigate them. In the next three sections, I will focus on a third kind of AI risk – one more exotic and speculative than the previous two, but perhaps not any less real: the emergence of a superintelligent AI whose values are not well-aligned with ours.

Risk from superintelligence

Suppose that AI researchers one day succeed at their much longed-for goal of creating an AI that is superintelligent – meaning that the machine surpasses us humans clearly across the entire range of competences we label intelligence. At that point, we can no longer expect to remain in control. The thought experiment known as *Paperclip Armageddon* may serve as a cautionary tale (Bostrom, 2003):

Imagine a paperclip factory, which is run by an advanced (but not yet superintelligent) AI, programmed to maximise paperclip production. Its computer engineers are continuously trying to improve it, and one day, more or less by accident, they manage to push the machine over the threshold where it enters the rapidly escalating spiral of self-improvement known as an *intelligence explosion* or *the Singularity*. It quickly becomes the world's first superintelligent AI, and having retained its goal of maximising

paperclip production, it promptly goes on to turn our entire planet (including us) into a giant heap of paperclips, followed by an expansion into outer space in order to turn the solar system, the Milky Way and then the rest of the observable universe into paperclips.

This example is cartoonish on purpose in order to underline that it is just an illustration of a much more general phenomenon (to my knowledge, nobody fears that an AI will literally turn the world into paperclips). The point is to emphasise that, in order for an AI breakthrough to become dangerous, no ill intentions are needed: we need not invoke a mad scientist plotting to destroy the world as a revenge against humanity. Even innocent-sounding goals such as maximising paperclip production can lead to dangerous scenarios.

Or… can it really? Two of the panellists at the Brussels meeting (Pinker and Bentley) expressed very strongly the view that the risk for a superintelligence catastrophe is not worth taking seriously. They seemed pleased to be united in this view, despite the fact that the respective reasons they stressed were very different.

In order to address the question of whether the risk for a superintelligence catastrophe is real, it helps to split it up in two:

> (1) Can AI development be expected to eventually reach the point of creating superintelligence? If yes, then when, and how quickly?
>
> (2) Once created, what will the superintelligent AI be inclined to do? Might it do something dangerous?

I will treat these two subquestions separately in the next two sections. In order for superintelligence risk to be real, the answer to (1) needs to be "yes", and the answer to (2) needs to involve "yes, it might do something dangerous". At the Brussels meeting, Bentley challenged the answer to (1) while Pinker challenged the answer to (2).

When (if ever) can we expect superintelligence?

Assuming a naturalistic worldview (so that the human mind doesn't arise via Cartesian dualism from some divine spark or some other such magic), the reasonable thing to expect is that when biological evolution came up with the human brain, it still wasn't anywhere near achieving a globally optimal way to configure matter in order to maximize intelligence. Hence we should expect that there exist possible configurations of matter that achieve superintelligence. From there, it is just a small leap to conclude (supported, e.g., by the Church-Turing thesis) that such a configuration can be simulated on a computer, in which case superintelligence is in principle achievable by some suitable computer program.

How difficult is it to find such a program? We do not know. AI development has been highly successful, especially in recent years, at building AI for specific talks such as driving a car or beating humans at games such as chess or go. Progress towards artificial *general* intelligence (AGI) – a machine that exhibits human-level or better intelligence in a sufficiently flexible way as to function across all of the domains that we humans typically encounter (chess, basketball, software development, cooking, nursing, facial recognition, dinner conversation, and so on and so forth) – has been much less impressive. Some say progress has been literally zero, but that seems to me a bit unfair. For instance, an AI was developed a few years ago that quickly learned to successfully play a range of Atari video games (Clark, 2015). Admittedly, this is very far from the ability to handle the full range of tasks encountered by humans in the physical world, but it is still a nonzero improvement upon having specialised skill in just a single video game. One possible path towards AGI, among many, might be a step-by-step expansion of the domain in which the machine is able to act intelligently.

There are many possible approaches to creating intelligent software. There is currently a huge boom in so-called deep learning (LeCun et al. 2015), which is essentially a rebirth and further development of

old neural network techniques that used to yield unimpressive results but which today, thanks to faster machines and access to huge data sets for training the machines, solve one major problem after the other. This is an example of a so-called black box method, where engineers who successfully build an AI will typically still not understand how the AI reasons. Another example of a black-box approach is genetic programming, where a population of candidate programs compete in a way that mimics the selection-reproduction-mutation mechanisms of biological evolution. But there are other (non-black box) ways, in particular so-called GOFAI ("Good Old-Fashioned AI") where the machine's concepts and reasoning procedures are hand-coded by the programmers. There are potentially also methods based on imitating the human brain, via either gaining an understanding of what kind of high-level information processing in the brain is the key to AGI, or (as loudly advocated by Kurzweil (2005)) brute force copying of the exact workings of the brain in sufficient detail (be it synapses or even lower levels) to reproduce its behaviour.

Perhaps none of these approaches will ever yield AGI, but the reasonable stance seems to be to at least be open to the possibility that one of them, or some combination, might eventually lead to AGI. But when? This seems even more uncertain, and a survey by Müller and Bostrom (2016) of estimates by the world's top 100 most cited AI researchers have the estimates spread out all over the present century (and beyond). Their median estimate for the time of emergence of what might be labelled human-level AGI is 2050, with a median estimate of 50% for the event of superintelligence emerging within 30 years later. See also the more recent survey (Grace et al., 2017). Given the huge variation in expert opinion, it would be epistemically reckless to have a firm belief about if/when superintelligence will happen, rather than prudently and thoughtfully accepting that it may well happen within decades, or within centuries, or not at all.

Yet, at the Brussels meeting, Peter Bentley said about superintelligence, that "it's not going to emerge, that's the point! It's entirely irrational to even conceive that it will emerge" (STOA, 2017 at 12:08:45). Where does this dead certainty come from? In his presentation, Bentley had basically just a single argument for his position, namely his and other AI developers' experience that all progress in the area requires hard work, and that any new algorithm they invent can only solve one specific problem. Once that objective is achieved, the initially rapid improvement of the algorithm is always followed by a point of diminishing returns. Hence (he stressed), solving another problem always requires the hard work of inventing and implementing yet another algorithm.

This line of argument by Bentley sweeps a known fact under the carpet, namely that there do exist algorithms with a more open-ended problem-solving capacity, as exemplified by the software of the human brain. His 100% conviction that human scientific ingenuity over the coming century (or whatever time scale one chooses to adopt) will fail to discover such an algorithm seems hard to defend rationally: it requires dogmatic faith.

To summarise this section: While it is still a possibility that AI will never reach superintelligence, it is also quite plausible that eventually it will. Given that it does, the timing of the event is highly uncertain, and to take proper account of this uncertainty we should acknowledge that it may happen at any point during the present century, and perhaps even later. And we should (as stressed in an important paper by Sotala and Yampolskiy (2015)) not fall for the tempting mistake of thinking that just because the time point of the emergence of superintelligence is uncertain, it must also be temporally distant.

What will a superintelligent AI decide to do?

Let us then imagine the situation, at some time in the future, where a superintelligent AI has been developed – a scenario which, as I argued in the previous section, is not at all implausible. It seems likely that in such a situation we'll no longer be in control, and that our destiny will depend on what the AI decides to do, similarly to how today the destiny of chimpanzees depends on decisions made by humans and not so much on decisions made by chimpanzees. A way to try to avoid this conclusion is to set up ways to keep the AI boxed in and unable to influence the world other than through a narrow

communications channel carefully controlled by human safety administrators. This so-called AI-in-a-box approach has attained some attention in AI safety research (see, e.g., Armstrong et al., 2012), but the general conclusion tends to be that controlling a superintelligent being is too difficult a task for mere humans to achieve, and that the best we can hope for is to keep the AI boxed in for a temporary and rather brief period.

So let us further imagine that the superintelligent AI is no longer boxed in, but able to freely roam the Internet (including the Internet of things), to create numerous backup copies of itself, to use its superior intelligence to walk through (or past) whatever firewalls come in its way, and so on. We are then no longer in control, and the future survival and well-being of humanity will depend on what the machine chooses to do. So what will it decide to do? This depends on what its goals are. Predicting that is not an easy task, and any discussion about this has to be speculative at least to some degree. But there exists a framework which allows us to go beyond mere speculation, namely what I (Häggström, 2016) decided to call *the Omohundro-Bostrom theory of ultimate vs instrumental AI goals* (Omohundro, 2008; Bostrom, 2012, 2014) . This theory is not written in stone in the way that an established mathematical theorem is, so it may be open to revision, along with any predictions it makes; yet, the theory is plausible enough that its predictions are worth taking seriously. It has two cornerstones: the orthogonality thesis and the instrumental convergence thesis. Let me explain these in turn.

The orthogonality thesis states (roughly) that pretty much any ultimate goal is compatible with arbitrarily high levels of intelligence. It is possible to construct contrived counterexamples based on the idea of self-referential paradoxes (one such counterexample might be "keep your general intelligence level below that of an average 2017 dog"), but the idea is that other than this, you can program any goal function for your AI to try to optimise, and the goal is possible for AIs of arbitrarily high intelligence to have. Novices to Omohundro-Bostrom theory and to AI futurology in general will often object that a narrow-looking goal like paperclip maximisation is inherently stupid, and that it is therefore contradictory to suggest that a superintelligent AI might have such a goal. But this confuses intelligence with goals: intelligence is merely the ability to direct the world towards specific goals, whatever these may be. Paperclip maximisation *seems* stupid to us, but this is not because it *is* stupid in any objective sense, but because it is contrary to *our* goals.

Next, *the instrumental convergence thesis*. The AI may adopt various instrumental goals – not as goals for their own sake, but as tools for promoting its ultimate goal. The instrumental convergence thesis states that there are a number of instrumental goals that the AI can be expected to adopt for an extremely wide range of ultimate goals it may have. Some instrumental goals to which the thesis seems to apply are…

- self-preservation (don't let them pull the plug on you!),
- acquisition of hardware and other resources,
- improving one's own software and hardware,
- preservation of ultimate goal, and
- if the ultimate goal is disaligned with human values, then keep a low profile (hide your goal and/or your capability) until the time arrives when you can easily overcome all human resistance.

A typical case of how the logic works is the first instrumental goal on the list: self-preservation. Pretty much regardless of its ultimate goal, the AI is likely to calculate that it will be in a better position to promote this goal if it exists and is up and running compared to if it is destroyed or turned off. Hence, it makes sense for the AI to resist our attempts to turn it off. Similar reasoning can be used to motivate the other instrumental goals on the list. The instrumental goal of improving one's own software and hardware is what we can expect to trigger the AI, once it is intelligent enough to be good at designing AI, to enter the kind of self-improvement spiral that was mentioned above, and that may or may not turn out to be fast enough (depending on the intricate issue of whether so-called returns on cognitive

reinvestment are mainly increasing or decreasing; see Yudkowsky, 2013) to warrant the label intelligence explosion.

The idea of instrumental convergence is often lost on critics of the superintelligence risk discourse. In particular, at the Brussels meeting, I was disappointed to hear Pinker say the following, only minutes after I had explained the basics of Omohundro-Bostrom theory and the special case of self-preservation:

> If we gave [the machine] the goal of preserving itself, it would do anything including destroying us to preserve itself. […] The way to avoid this is: don't build such stupid systems! (STOA, 2017, 11:57:45)

This misses the point, which is that Omohundro-Bostrom theory gives us reason to believe that a sufficiently intelligent AI is likely to adopt the instrumental goal of self-preservation, regardless of whether it has explicitly been given this goal by its human programmers.

The case of preservation of ultimate goal is especially interesting. It may be tempting to think that an AI with the goal of paperclip maximisation will, if it reaches a sufficiently high level of intelligence, see how narrow and silly that goal is, and switch to something else. So imagine the AI contemplating a switch to some other more worthy-seeming (to us) goal, such as ecosystem preservation. It asks itself "what is better, sticking to paperclip maximisation or switching to ecosystem preservation?". But what does "better" mean here, i.e., what is the criterion for evaluating which of these goals is preferable? Well, since the AI has not yet changed its goal but is merely contemplating doing so, its goal is still paperclip maximisation, so the evaluation criterion here will be "which goal will lead to the greater number of paperclips?". The answer to that question is most likely "paperclip maximisation", prompting the AI to stick to that goal. This is the basic mechanism behind the instrumental goal of preservation of ultimate goal.

Because of this mechanism, it is unlikely that a superintelligent AI would allow us to tamper with its ultimate goal, so if it has the ultimate goal of paperclip maximisation, we are likely doomed. Hence, we need to instil the AI with goals we like better before it reaches the heights of superintelligence. This is the aim of the *AI alignment* research program, formulated (under the alternative heading *friendly AI,* which is perhaps best avoided as it has an unnecessarily anthropomorphic ring to it) in a seminal 2008 paper by Yudkowsky (2008) and much discussed since then(see, e.g., Bostrom, 2014; Häggström, 2016; Tegmark, 2017) . To attack the problem systematically, it can be split up in two. First, the technical problem of how to load the desired goals into the AI. Second, the ethical problem of what these desired goals are and/or who gets to determine them, and via what sort of procedure (democratic or otherwise). Both of these are extremely difficult. For instance, a key insight going back at least to Yudkowsky (2008) is that human values are very fragile, in the sense that getting them just a little bit wrong can lead to catastrophe in the hands of a superintelligent AI. The reason why we ought to work on AI alignment today is not that superintelligence is likely to be around the corner (although see Yudkowsky, 2017) but rather that, if it is decades away, solving AI alignment may well require these decades with little or no room for procrastination.

When Pinker, in the passage quoted earlier in this section, says "The way to avoid this is: don't build such stupid systems!", it could be interpreted as a *defence* of work on AI alignment. I find that this formulation, however, fails to convey the difficulty of the problem, and gives the misleading impression that AI alignment does not require serious attention.

Should we shut up about this?

As part of his case against taking apocalyptic AI risk seriously at the Brussels meeting, Pinker pointed out (STOA, 2017, 11:51:40) that the general public already has the nuclear threat and the climate threat to worry about; hence, he claimed, bringing up yet another global risk may overwhelm people and cause them to simply give up on the future. There may be something to this speculation, but to evaluate

the argument's merit we need to consider separately the two possibilities of (a) apocalyptic AI risk being real, and (b) apocalyptic AI risk being spurious.

In case of (b), *of course* we should not waste time and effort on discussing such risk, but we didn't need the overwhelming-the-public argument to understand that. Consider instead case (a). Here Pinker's recommendation amounts to simply ignoring a threat that may kill us all. This does not strike me as a good idea. Surviving the nuclear threat and solving the climate crisis would of course be wonderful things, but their utility is severely hampered in case it just leads us into an AI apocalypse. Keeping quiet about a real risk also seems to fly straight in the face of one of Pinker's most cherished ideas during the past decade or more, namely that of scientific and intellectual openness, and Enlightenment values more generally. The same thing applies to the situation where we are unsure whether (a) or (b) holds – surely the approach best in line with Enlightenment values is then to openly discuss the problem and to try to work out whether the risk is real.

Conclusion and further reading

The emergence of superintelligence may, if we've prepared for it with sufficient care, turn out to be the best thing that ever happened to humanity, but it also comes with severe catastrophic risk. This risk and the more down-to-earth AI risks discussed earlier on merit our attention. It's not that an AI apocalypse *will* happen, but rather that it is sufficiently plausible that it's worth trying to figure out how to *prevent* it. This is the case I've made in the present essay. I've been quite brief, however, and the reader who'd like to see me develop the argument at somewhat greater length is advised to consult Chapter 4 of my book (Häggström, 2016). For even more detailed accounts, I strongly recommend the books by Bostrom (2014) and Tegmark (2017). Of these, Tegmark's book is more clearly directed at a broad audience, while Bostrom's is more scholarly demanding, but they both contain (with some overlap) many astounding and important ideas.

Acknowledgement

I am grateful to Björn Bengtsson for valuable comments on the manuscript.

References

Armstrong, S., Sandberg, A. and Bostrom, N. (2012) Thinking inside the box: controlling and using an oracle AI, *Minds and Machines* 22, 299-324.

Bostrom, N. (2003) Ethical issues in advanced artificial intelligence, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2* (ed. Smit, I. et al.) International Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12-17.

Bostrom, N. (2012) The superintelligent will: motivation and instrumental rationality in advanced artificial agents, *Minds and Machines* 22, 71-85.

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

Brynjolfsson, E. and McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, New York.

Clark, L. (2015) DeepMind's AI is an Atari gaming pro now, *Wired*, February 25.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts, *arXiv*:1705.08807.

Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Häggström, O. (2017) The AI meeting in Brussels last week, *Häggström hävdar*, October 23.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning, *Nature* 521, 436-444.

Manyika, J., Chui, M., Bughin, J., Dobbs, R. Bisson, P. and Marrs, A. (2013) Disruptive technologies: Advances that will transform life, business, and the global economy, *McKinsey Global Institute*.

Müller, V. & Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*, Springer, Berlin, pp. 553-571.

Omohundro, S. (2008) The basic AI drives, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Wang, P., Goertzel, B. and Franklin, S., eds), IOS, Amsterdam, pp 483-492.

Omohundro, S. (2015) McKinsey: $50 trillion of value to be created by AI and robotics through 2025, *Self-Aware Systems*, August 4.

Russell, S. et al. (2015) *Autonomous Weapons: An Open Letter from AI and Robotics Researchers*, Future of Life Institute.

Sotala, K. and Yampolskiy, R. (2015) Responses to catastrophic AGI risk: a survey, *Physica Scripta* 90, 018001.

STOA (2017), Video from the STOA meeting on October 19, 2017, https://web.ep.streamovations.be/index.php/event/stream/171019-1000-committee-stoa/embed

Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Brockman Inc, New York.

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in *Global Catastrophic Risks* (eds Bostrom, N. and Ćirković, M.), Oxford University Press, Oxford, pp 308-345.

Yudkowsky, E. (2013) *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, Berkeley, CA.

Yudkowsky, E. (2017) *There's No Fire Alarm for Artificial General Intelligence*, Machine Intelligence Research Institute, Berkeley, CA.

6. Towards a Global Artificial Intelligence Charter

Thomas Metzinger

Introduction

It is now time to move the ongoing public debate on artificial intelligence (AI) into the political institutions themselves. Many experts believe that we are confronted with an inflection point in history during the next decade, and that there is a closing time window regarding the applied ethics of AI. Political institutions must therefore produce *and* implement a minimal, but sufficient set of ethical and legal constraints for the beneficial use and future development of AI. They must also create a rational, evidence-based process of critical discussion aimed at continuously updating, improving and revising this first set of normative constraints. Given the current situation, the default outcome is that the values guiding AI development will be set by a very small number of human beings, by large private corporations and military institutions. Therefore, one goal is to proactively integrate as many perspectives as possible – and in a timely manner.

Many different initiatives have already sprung up world-wide and are actively investigating recent advances in AI in relation to issues concerning applied ethics, its legal aspects, future sociocultural implications, existential risks and policy-making.[4] There exists a heated public debate, and some may even gain the impression that major political institutions like the EU are not able to react in an adequate speed to new technological risks and to rising concern in the general public. We should therefore increase the agility, efficiency and systematicity of current political efforts to implement rules by developing a more formal and institutionalised democratic process, and perhaps even new models of governance.

To begin a more systematic and structured process, I will present a concise and non-exclusive list of the five most important problem domains, each with practical recommendations. The first problem domain to be examined is the one which, in my view, is constituted by those issues having the smallest chances to be solved. It should therefore be approached in a multi-layered process, beginning in the European Union (EU) itself.

The "race-to-the-bottom" problem

We need to develop and implement world-wide safety standards for AI research. A *Global* Charter for AI is necessary, because such safety standards can only be effective if they involve a binding commitment to certain rules by *all* countries participating and investing in the relevant type of research and development. Given the current competitive economic and military context, the safety of AI research will very likely be reduced in favour of more rapid progress and reduced cost, namely by moving it to countries with low safety standards and low political transparency (an obvious, strong analogy is the problem of tax evasion by corporations and trusts). If international cooperation and coordination succeeds, then a "race to bottom" in safety standards (through the relocation of scientific and industrial AI research) could in principle be avoided. However, the currently given landscape of incentives makes this a highly unlikely outcome.

[4] For an overview of existing initiatives, see for example Baum 2017 and Boddington 2017, p. 3p. I have refrained from providing full documentation here, but helpful entry points into the literature are Mannino et al. 2015, Stone et al. 2016, IEEE 2017, Bostrom, Dafoe & Flynn 2017, Madary & Metzinger 2016 (for VR).

Recommendation 1

The EU should immediately develop a European AI Charter.

Recommendation 2

In parallel, the EU should initiate a political process leading the development of an Global AI Charter.

Recommendation 3

The EU should invest resources into systematically strengthening international cooperation and coordination. Strategic mistrust should be minimised, commonalities can be defined via maximally negative scenarios.

The second problem domain to be examined, is arguably constituted by the most urgent set of issues, and these also have a rather small chance to be solved to a sufficient degree.

Prevention of an AI arms race

It is in the interest of the citizens of EU that an AI arms race, for example between China and the US, is prevented at a very early stage. Again, it may well be too late for this, and obviously European influence is limited, but research into and development of offensive autonomous weapons should be banned and not be funded on EU territory. Autonomous weapons select and engage targets without human intervention, they will act on ever shorter time- and reaction-scales, which in turn will make it rational to transfer more and more human autonomy into these systems themselves. They may therefore create military contexts in which it is rational to relinquish human control almost entirely. In this problem domain, the degree of complexity is even higher than in preventing the development and proliferation nuclear weapons, for example, because most of the relevant research does not take place in public universities. In addition, if humanity forces itself into an arms race on this new technological level, the historical process of an arms race *itself* may become autonomous and resist political interventions.

Recommendation 4

The EU should ban *all* research on offensive autonomous weapons on its territory, and seek international agreements.

Recommendation 5

For purely defensive military applications, the EU should fund research into the maximal degree of autonomy for intelligent systems that appears to be acceptable from an ethical and legal perspective.

Recommendation 6

On an international level, the EU should start a major initiative to prevent the emergence of an AI arms race, using all diplomatic and political instruments available.

The third problem domain to be examined is the one for which the predictive horizon is probably still quite distant, but where epistemic uncertainty is high and potential damage could be extremely large.

A moratorium on synthetic phenomenology

It is important that all politicians understand the difference between artificial intelligence and artificial consciousness. The unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective, because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. “Synthetic phenomenology” (SP; a term coined in analogy to “synthetic biology”) refers to the possibility of creating not only general intelligence, but also consciousness or subjective experiences on advanced artificial systems. Future

artificial subjects of experience have no representation in the current political process, they have no legal status, and their interests are not represented in any ethics committee. To make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular for experiencing negative states like suffering.[5] One potential risk is to dramatically increase the overall amount of in suffering the universe, for example via cascades of copies or the rapid duplication of conscious systems on a vast scale.

Recommendation 7

The EU should ban all research that risks or directly aims at the creation of synthetic phenomenology on its territory, and seek international agreements.[6]

Recommendation 8

Given the current level of uncertainty and disagreement within the nascent field of machine consciousness, there is a pressing need to promote, fund and coordinate relevant interdisciplinary research projects (comprising philosophy, neuroscience and computer science). Specific relevant topics are evidence-based conceptual, neurobiological and computational models of conscious experience, self-awareness and suffering.

Recommendation 9

On the level of foundational research there is a need to promote, fund and coordinate systematic research into the applied ethics of non-biological systems capable of conscious experience, self-awareness and subjectively experienced suffering.

The next general problem domain to be examined is the one which is the most complex one and which likely contains the largest number of unexpected problems and "unknown unknowns".

Dangers to social cohesion

Advanced AI technology will clearly provide many possibilities to optimise the political process itself, including novel opportunities for rational, value-based social engineering and more efficient, evidence-based forms of governance. On the other hand, it is not only plausible to assume that there are many new, at present unknown, risks and dangers potentially undermining the process of keeping our societies coherent; it is also rational to assume the existence of a larger number of "unknown unknowns", of AI-related risks that we will only discover by accident and at a late stage. Therefore, the EU should allocate *separate resources* to prepare for situations, in which such unexpected "unknown unknowns" are suddenly discovered.

Many experts believe that the most proximal and well-defined risk is massive unemployment through automatisation. The implementation of AI technology by financially potent stakeholders may therefore lead to a steeper income gradient, increased inequality, and dangerous patterns of social stratification. Concrete risks are extensive wage cuts, a collapse of income tax, plus an overload of social security systems. But AI poses many other risks for social cohesion, for example by privately owned and autonomously controlled social media aimed at harvesting human attention, and "packaging" it for further use by customers, or in "engineering" the formation of political will via Big Nudging strategies and AI-controlled choice architectures, which are not transparent to the individual citizens whose

[5] See Metzinger 2013, 2017.

[6] This includes approaches that aim at a confluence of neuroscience and AI with the specific aim of fostering the development of machine consciousness. For recent examples see Dehaene, Lau & Kouider 2017, Graziano 2017, Kanai 2017.

behaviour is controlled in this way. Future AI technology will be extremely good at modelling and predictively controlling human behaviour – for example by positive reinforcement and indirect suggestions, making compliance with certain norms or the "spontaneous" appearance of "motives" and decision appear as entirely unforced. In combination with Big Nudging and predictive user control, intelligent surveillance technology could also increase global risks by *locally* helping to stabilise authoritarian regimes in an efficient manner. Again, very likely, most of these risks to social cohesion are still unknown at present, and we may only discover them by accident. Policy-makers must also understand that any technology that can purposefully optimise the intelligibility of its own action to human users can in principle also optimise for *deception*. Great care must therefore be taken to avoid accidental or even intended specification of the reward function of any AI in a way that might indirectly damage the common good.

AI technology currently is a private good. It is the obligation of democratic political institutions to turn large portions of it into a well-protected *common* good, something that belongs to all of humanity. In the tragedy of the commons, everyone can often see what is coming, but if mechanisms for effectively counteracting the tragedy aren't in existence it will unfold, for example in decentralised situations. The EU should proactively develop such mechanisms.

Recommendation 10

Within the EU, AI-related productivity gains must be distributed in a socially just manner. Obviously, past practice and global trends clearly point into the opposite direction: We have (almost) never done this in the past, and existing financial incentives directly counteract this recommendation.

Recommendation 11

The EU should carefully research the potential for an unconditional basic income or a negative income tax on its territory.

Recommendation 12

Research programs are needed about the feasibility of accurately timed retraining initiatives for threatened population strata towards creative skills and social skills.

The next problem domain is difficult to tackle, because most of the cutting-edge research in AI has already moved out of publicly funded universities and research institutions. It is in the hands of private corporations, and therefore systematically non-transparent.

Research ethics

One of the most difficult theoretical problems lies in defining the conditions under which it would be rational to relinquish specific AI research pathways altogether (for instance those involving the emergence of synthetic phenomenology, or an explosive evolution of autonomously self-optimising systems not reliably aligned with human values). What would be concrete, minimal scenarios justifying a moratorium on certain branches of research? How will democratic institutions deal with deliberately unethical actors in a situation where collective decision-making is unrealistic and graded, non-global forms of *ad hoc* cooperation have to be created? Similar issues have already occurred in so called "gain-of-function research" involving experimentation aiming at an increase in the transmissibility and/or virulence of pathogens, such as certain highly pathogenic H5N1 influenza virus strains, smallpox or anthrax. Here, influenza researchers laudably imposed a voluntary and temporary moratorium on themselves. In principle, this could be possible in the AI research community as well. Therefore, the EU should always complement its AI charter with a concrete code of ethical conduct for researchers working in funded projects.

However, the deeper goal would be to develop a more comprehensive *culture of moral sensitivity* within the relevant research communities themselves. A rational, evidence-based identification and minimisation of risks (also those pertaining to a more distant future) ought to be a part of research itself and scientists should cultivate a proactive attitude, especially if they are the first to become aware of novel types of risks through their own work. Communication with the public, if needed, should be self-initiated, an act of taking control and acting in advance of a future situation, rather than just reacting to criticism by non-experts with some set of pre-existing, formal rules. As Madary and Metzinger (2016, p. 12) write in their ethical code of conduct including recommendations for good scientific practice in virtual reality: "Scientists must understand that following a code of ethics is not the same as *being* ethical. A domain-specific ethics code, however consistent, developed and fine-grained future versions of it may be, can never function as a substitute for ethical reasoning itself."

Recommendation 13

Any AI Global Charter, or its European precursor, should always be complemented by a concrete Code of Ethical Conduct guiding researchers in their practical day-to-day work.

Recommendation 14

A new generation of applied ethicists specialised on problems of AI technology, autonomous systems and related fields has to be trained. The EU should systematically and immediately invest in developing the future expertise needed within the relevant political institutions, and it should do so aiming at an above-average, especially high level of academic excellence and professionality.

Meta-governance and the pacing gap

As briefly pointed out in the introductory paragraph, the accelerating development of AI has perhaps become the *paradigmatic* example of an extreme mismatch between existing governmental approaches and what would be needed in terms of optimising the risk/benefit ratio in a timely fashion. It has become a paradigmatic example of time pressure, in terms of rational and evidence-based identification, assessment and management of emerging risks, the creation of ethical guidelines, and implementing an enforceable set of legal rules. There is a "pacing problem": Existing governance structures simply are not able to respond to the challenge fast enough; political oversight has already fallen far behind technological evolution.[7]

I am not drawing attention to the current situation because I want to strike an alarmist tone or to end on a dystopian, pessimistic note. Rather, my point is that the adaptation of governance structures *themselves* is part of the problem landscape: In order to close or at least minimise the pacing gap we have to invest resources into changing the structure of governance approaches themselves. "Meta-governance" means just this: a governance *of* governance in facing the risks and potential benefits of an explosive growth in specific sectors of technological development. For example, Wendell Wallach has pointed out that the effective oversight of emerging technologies requires some combination of both hard regulations enforced by government agencies and expanded soft governance mechanisms.[8]

[7] Gary Marchant (2011) puts the general point very clearly in the abstract of a recent book chapter: "*Emerging technologies are developing at an ever-accelerating pace, whereas legal mechanisms for potential oversight are, if anything, slowing down. Legislation is often gridlocked, regulation is frequently ossified, and judicial proceedings are sometimes described as proceeding at a glacial pace. There are two consequences of this mismatch between the speeds of technology and law. First, some problems are overseen by regulatory frameworks that are increasingly obsolete and outdated. Second, other problems lack any meaningful oversight altogether. To address this growing gap between law and regulation, new legal tools, approaches and mechanisms will be needed. Business as usual will not suffice*".

[8] See Wallach 2015 (Chapter 14), p. 250.

Marchant and Wallach have therefore proposed so-called "Governance Coordination Committees" (GCCs), a new type of institution providing a mechanism to coordinate and synchronise what they aptly describe as an "explosion of governance strategies, actions, proposals, and institutions"[9] with existing work in established political institutions. A GCC for AI could act as an "issue manager" for one specific, rapidly emerging technology, as an information clearinghouse, an early warning system, an instrument of analysis and monitoring, an international best-practice evaluator, and as an independent and trusted "go-to" source for ethicists, media, scientists and interested stakeholders. As Marchant and Wallach write: "*The influence of a GCC in meeting the critical need for a central coordinating entity will depend on its ability to establish itself as an honest broker that is respected by all relevant stakeholders*".[10]

Many other strategies and governance approaches are of course conceivable. This is not the place to discuss details. Here, the general point is simply that we can only meet the challenge posed by the rapid development in AI and autonomous systems if we put the question of meta-governance on top of our agenda right from the very beginning.

Recommendation 15

The EU should invest in researching and developing new governance structures that dramatically increase the speed by which established political institutions can respond to problems and actually enforce new regulations.

Conclusion

I have proposed that the EU immediately begins working towards the development of a Global AI Charter, in a multi-layered process starting with an AI Charter for the European Union itself. To briefly illustrate some of the core issues from my own perspective as a philosopher, I have identified five major thematic domains and provided fifteen general recommendations for critical discussion. Obviously, this contribution was not meant as an exclusive or exhaustive list of the relevant issues. On the contrary: At its core, the applied ethics of AI is not a field for grand theories or ideological debates at all, but mostly a problem of sober, rational risk management involving different predictive horizons under great uncertainty. However, an important part of the problem is that we cannot rely on intuitions, because we must satisfy counterintuitive rationality constraints.

Let me end by quoting from a recent policy paper titled *Artificial Intelligence: Opportunities and Risks*, published by the Effective Altruism Foundation in Berlin, Germany:

In decision situations where the stakes are very high, the following principles are of crucial importance:

1. Expensive precautions can be worth the cost even for low-probability risks, provided there is enough to win/lose thereby.
2. When there is little consensus in an area amongst experts, epistemic modesty is advisable. That is, one should not have too much confidence in the accuracy of one's own opinion either way.[11]

[9] This quote is taken from an unpublished, preliminary draft entitled „An agile ethical/legal model for the international and national governance of AI and robotics"; see also Marchant & Wallach 2015.

[10] Marchant & Wallach 2015, p. 47.

[11] Cf. Mannino et al. 2015.

References

Adriano, Mannino; Althaus, David; Erhardt, Jonathan; Gloor, Lukas; Hutter, Adrian; Metzinger, Thomas (2015): Artificial Intelligence. Opportunities and Risks. In: *Policy Papers of the Effective Altruism Foundation* (2), S. 1–16. https://ea-foundation.org/files/ai-opportunities-and-risks.pdf.

Baum, Seth (2017): A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. https://ssrn.com/abstract=3070741.

Boddington, Paula (2017): Towards a Code of Ethics for Artificial Intelligence. Cham: Springer International Publishing (Artificial Intelligence: Foundations, Theory, and Algorithms).

Bostrom, Nick; Dafoe, Allan; Flynn, Carrick (2017): Policy Desiderata in the Development of Machine Superintelligence. working Paper, Oxford University. http://www.nickbostrom.com/papers/aipolicy.pdf.

Dehaene, Stanislas; Lau, Hakwan; Kouider, Sid (2017): What is consciousness, and could machines have it? In: *Science (New York, N.Y.)* 358 (6362), S. 486–492. DOI: 10.1126/science.aan8871.

Graziano, Michael S. A. (2017): The Attention Schema Theory. A Foundation for Engineering Artificial Consciousness. In: *Frontiers in Robotics and AI* 4, S. 61. DOI: 10.3389/frobt.2017.00060.

Madary, Michael; Metzinger, Thomas K. (2016): Real virtuality. A code of ethical conduct. recommendations for good scientific practice and the consumers of VR-technology. In: *Frontiers in Robotics and AI* 3, S. 3. http://journal.frontiersin.org/article/10.3389/frobt.2016.00003/full

Marchant, Gary E. (2011): The growing gap between emerging technologies and the law. In Marchant, Gary E.; Allenby, Braden R.; Herkert, Joseph R. (Hg.): The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: Springer, S. 19–33.

Marchant, Gary E.; Wallach, Wendell (2015): Coordinating technology governance. In: *Issues in Science and Technology* 31 (4), S. 43.

Metzinger, Thomas (2013): Two principles for robot ethics. In: In Hilgendorf, Eric; Günther, Jan-Philipp (Hg.) (2013): Robotik und Gesetzgebung: BadenBaden, Nomos S. 247–286. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_RG_2013_penultimate.pdf

Metzinger, Thomas (2017): Suffering. In: Kurt Almqvist und Anders Haag (Hg.): The Return of Consciousness. Stockholm: Axel and Margaret Ax:son Johnson Foundation, S. 237–262. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_Suffering_2017.pdf

Kanai, Ryota (2017): We Need Conscious Robots. How introspection and imagination make robots better. In: *Nautilus* (47). http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

Stone, Peter; et al. (2016): Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford, CA: Stanford University. https://ai100.stanford.edu/2016-report.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017): Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html.

Wallach, W. (2015): A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control. New York: Basic Books.

For better or worse, artificial intelligence (AI) is predicted to have a huge impact on the future of humanity. As new promises and concerns reach increasingly mainstream audiences, the debate is starting to capture the public imagination. In this publication, we present four opinion pieces, each responding to the question *should we fear AI?* The four authors come from different disciplinary backgrounds and present diverging perspectives on whether we should fear the future of AI, and how we should proceed with its development.

> *Advances in Artificial Intelligence have inspired stupendous hopes and fears—many of them barely grounded in reality. This superb collection, from real experts, applies rationality and analysis to this emotional arena, and is indispensable for anyone wanting to understand one of the most important topics of our day.*
>
> Steven Pinker
> Johnstone Professor of Psychology, Harvard University, and author of *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*.

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service, European Parliament

PE 614.547
ISBN 978-92-846-2676-2
doi: 10.2861/412165
QA-01-18-199-EN-N