DOI:10.5281/zenodo.2576737
CZU 004.8

THE RISKS OF ARTIFICIAL INTELLIGENCE

Titu-Marius I. Băjenescu
Switzerland
Corresponding Author: T.-M. Băjenescu, *tmbajenesco@gmail.com*

Received: October, 28, 2018
Accepted: December, 5, 2018

Abstract. The artificial intelligence (AI) revolution is the fastest of all revolutions that we have known. The AI is most often presented as a how to improve and refine existing processes in order to spectacular. Much more than a technology, AI represents a new way of interacting with our environment and doing business. While some welcome this innovation favourably, others call for a break to assess the potential risks to consumers. The paper examines the possible risks and benefits of AI for different areas of human activity.

Key words: *Risks of misuse, AI advantages, large-scale cybercrime, safety, reliability.*

Introduction

Artificial intelligence (AI) is found in many devices that enrich our everyday life with modern comfort functions. The spectrum ranges from smartphones to intelligent thermostats. AI is also used to meet important social challenges. It is a branch of computer science in which machines capture, learn, conclude, act and adapt to the real world - and thus strengthen human abilities and automate tedious or dangerous tasks. Some experts believe that it has the potential for a real social revolution [1-14, 16-18].

Perhaps - before we try to define what artificial intelligence (AI) is - we should know what intelligence is. Unfortunately, defining intelligence is a particularly delicate problem, and dictionaries are not very helpful. It is true that most authors agree that intelligence is "the ability to understand" of the human spirit; when it comes to defining it, we are struggling with extremely varied definitions - such as "sensing meaning," "embracing thinking," "making a clear idea," "understanding through knowledge," etc. And when we want to know what knowledge is, we read that it is what we have learned, what I understand, that ideas are representations that form thought, and that the meaning of a word, for example, is the ensemble of the intelligible ideas it conveys. We enter into a loop from where we can no longer go out, for in order to define a word we need other words and so on. There is nothing left to do but to consider intelligence as a primitive word that we will not try to define. Why should we talk about artificial intelligence? This term designates a science that has gradually emerged since the emergence of the modern computers: it is the set of concepts and methods used to make computers behave intelligently, similar to that of people who exercise their intelligence. The domain is vast and is rapidly growing, among other things due to the fact that the power of the handlers is growing fast, so it can manipulate masses of information that are increasingly important with the help of increasingly elaborate algorithms.

What is AI?

Very simply, it's machines doing things that are considered to require intelligence when humans do them: understanding natural language, recognising faces in photos, driving a car, or guessing what other books we might like based on what we have previously enjoyed reading.

It's the difference between a mechanical arm on a factory production line programmed to repeat the same basic task over and over again, and an arm that learns through trial and error how to handle different tasks by itself.

AIs are at work wherever you look, in industries from finance to transportation, monitoring the share market for suspicious trading activity or assisting with ground and air traffic control. They even help to keep spam out of your inbox. And this is just the beginning for artificial intelligence. As the technology advances, so too does the number of applications.

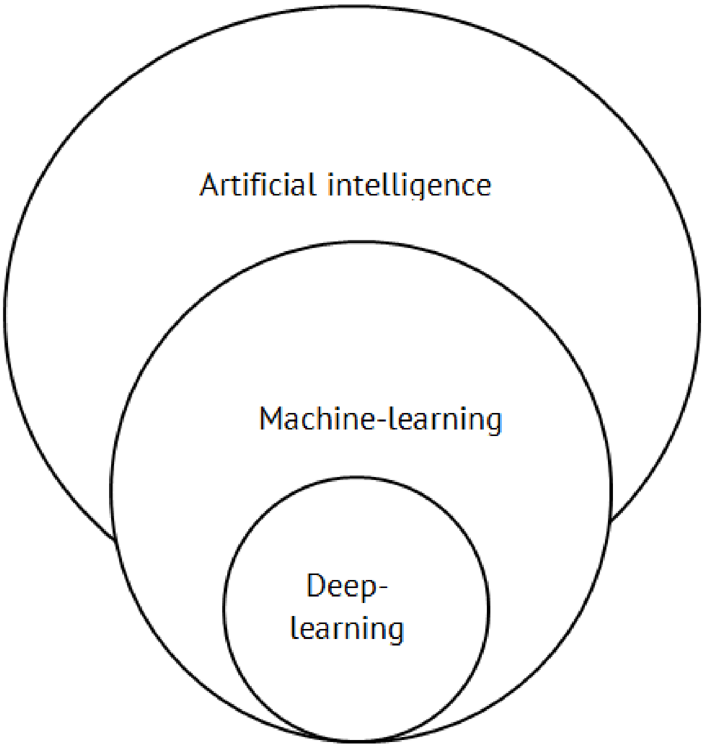

Figure 1. Relationships between AI, Machine-Learning and Deep-Learning.

As AIs are rolled out to assess everything, the risks that they will sometimes get it wrong – without us necessarily knowing – get worse. Since so much of the data that we feed AIs is imperfect, we should not expect perfect answers all the time. Recognising that is the first step in managing the risk. Decision-making processes built on top of AIs need to be made more open to scrutiny. Since we are building artificial intelligence in our own image, it is likely to be both as brilliant and as flawed as we are.

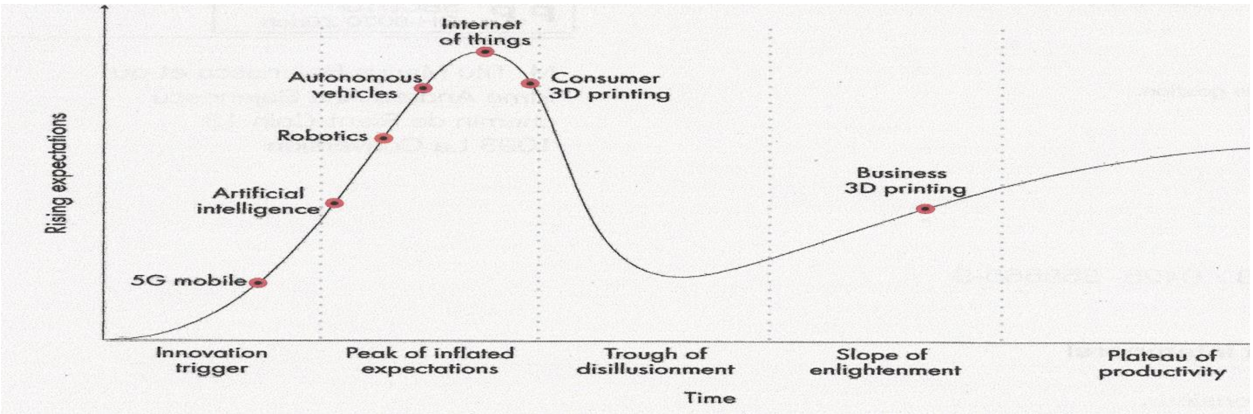

Figure 2. Gartner "hype" cycle applied to selected digital technologies.
Source: WDR 2016 team; adapted from Gartner 2015.

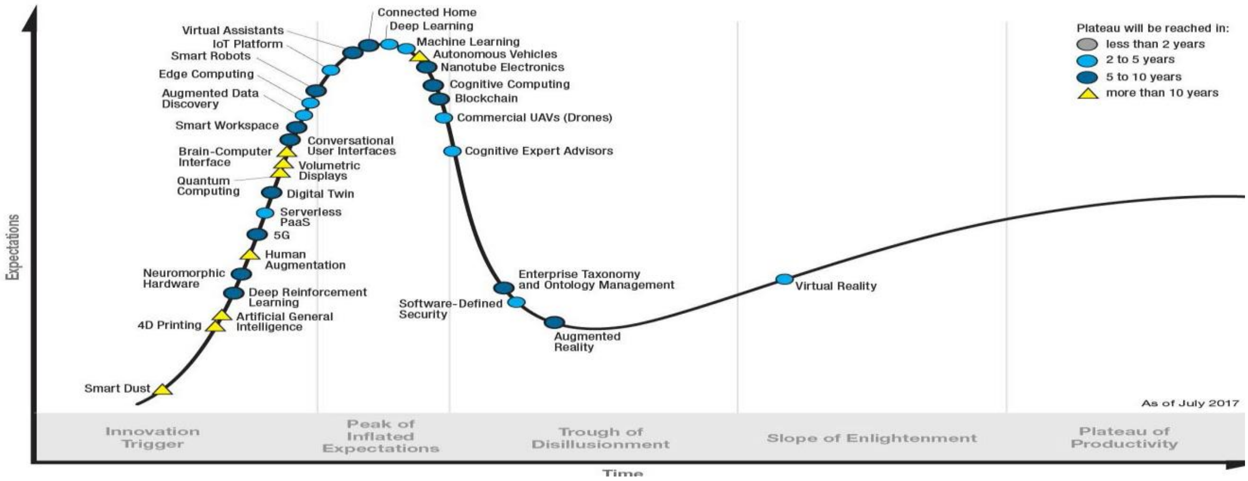

Figure 3. Overview of AI technologies (Source: Capgemini).

Machine-learning [2][12] is a very general and useful framework, but it is not “magic” and will not always work. In order to better understand when it will and when it will not work, it is useful to formalize the learning problem more. This will also help us develop debugging strategies for learning algorithms.

"Artificial intelligence will advance humanity," said Indian Prime Minister Narendra Modi in a speech in New Delhi with a view to a new *eGovernment initiativ*e using appropriate technologies to create paperless offices. "There will be arguments about whether jobs are left over or not. However, experts believe that there is a high probability that AI will create jobs," Modi said. And he added that AI is increasingly gaining influence and has the potential to transform economies (figure 4).

The question is which jobs are most at risk in which sectors. According to MIT economist David Autor, automation will substitute for more routinized occupations and complement high-skill, non-routine jobs. Whereas the effects on low-skill jobs will remain relatively unaffected, medium-skill jobs will gradually disappear, while demand for high-skill jobs will rise.

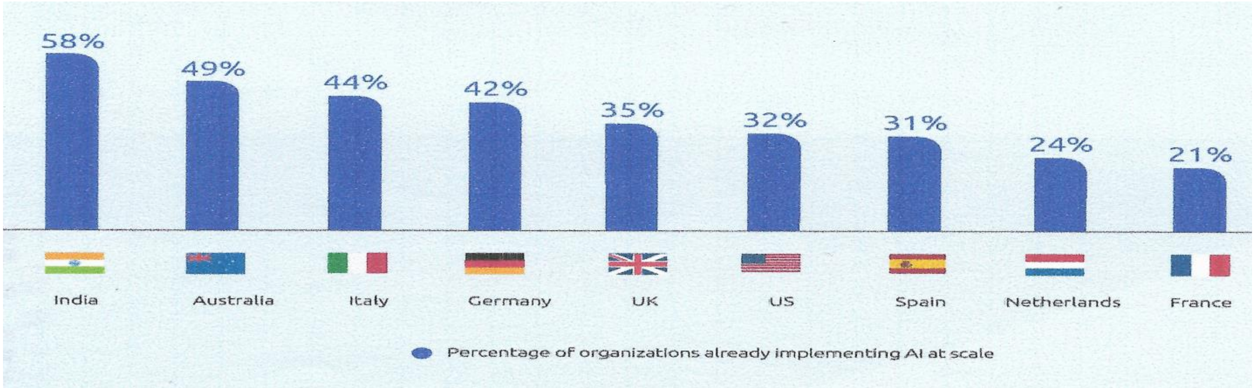

Figure 4. India leads in AI implementation at scale.
(Source: Capgemini Digital Transformation Institute, State of AI survey, June 2017).

Global risks

Many experts believe that alongside global opportunities, AI poses global risks, which will be greater than, say, the risks of nuclear technology - which in any case have historically been underestimated. Furthermore, scientific risk analysis suggests that high potential damages should be taken very seriously even if the probability of their occurrence were low.

While the risks from domain-specific AIs appear limited in the near future, there are long-term developments to take into consideration: in the not-so-distant future, artificial intelligence could in principle pose an existential threat, similar in scope to the pandemic risks associated with biotechnology [4, 6, 17].

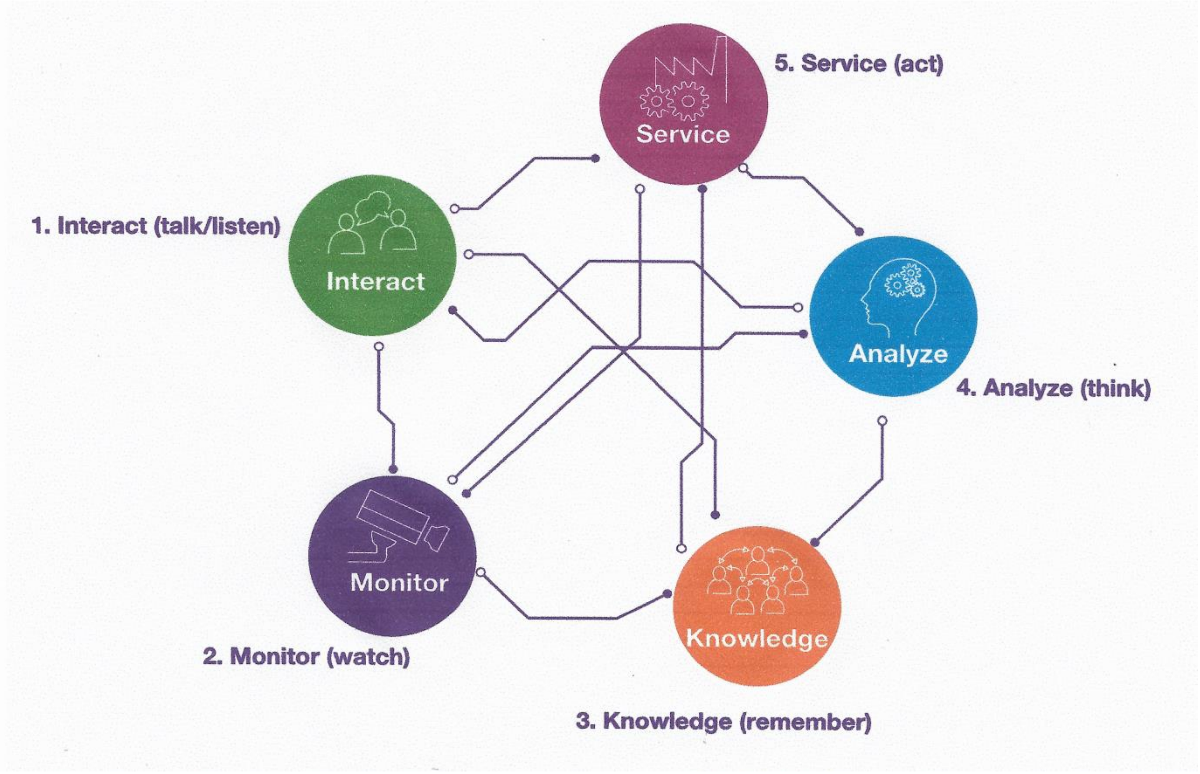

Figure 5. Five senses of artificial intelligence.
(Source: Capgemini, "The five senses of artificial intelligence: Christopher Stancombe", May 2017).

While the entertainment industry does offer significant opportunities for better education through personalized AI teaching and the gamification of learning material [66], it also increases the risk that a growing proportion of young people will have trouble completing their education due to a pathological addiction to video games and/or the internet [3].

On terrorism, crime, and other sources of risk:

The dangerous applications for AI would be criminals or large terrorist organizations using it to disrupt large processes or simply do pure harm. Terrorists could cause harm via digital warfare, or it could be a combination of robotics, drones, with AI and other things as well that could be really dangerous.

And, of course, other risks come from things like job losses. If we have massive numbers of people losing jobs and don't find a solution, it will be extremely dangerous. Things like lethal autonomous weapons systems should be properly governed - otherwise there's massive potential of misuse.

But this is the duality of this technology. Certainly, my conviction is that AI is not a weapon; AI is a tool. It is a powerful tool, and this powerful tool could be used for good or bad things. Our mission is to make sure that this is used for the good things, the most benefits are extracted from it, and most risks are understood and mitigated.

Some of the main risks associated with AI include:

Algorithmic bias: Machine-learning algorithms identify patterns in data and codify them in predictions, rules and decisions. If those patterns reflect some existing bias, the algorithms are likely to amplify that bias and may produce outcomes that reinforce existing patterns of discrimination.

Overestimating the capabilities of AI: Since AI systems do not understand the tasks they perform, and rely on their training data, they are far from infallible.

The reliability of their outcomes can be jeopardized if the input data is biased, incomplete or of poor quality.

Programmatic errors: Where errors exist, algorithms may not perform as expected and may deliver misleading results that have serious consequences.

Risk of cyber attacks: Hackers who want to steal personal data or confidential information about a company are increasingly likely to target AI systems.

Legal risks and liabilities: At present, there is little legislation governing AI, but that is set to change. Systems that analyze large volumes of consumer data may not comply with existing and imminent data privacy regulations, especially the EU's General Data Protection Regulation.

Reputational risks: AI systems handle large amounts of sensitive data and make critical decisions about individuals in a range of areas including credit, education, employment and health care. So any system that is biased, error-prone, hacked or used for unethical purposes poses significant reputational risks to the organization that owns it.

In the example of autonomous cars, the blind use of deep neural nets (DNNs) coupled directly to vehicle action control systems would be very risky: it would be equivalent to asking a taxi driver who has lost more than 80% of his brain as a result of an accident (and only keeping this occipito-temporary lane) to drive a car. It is simply not possible to ask these systems to do more than what they were originally designed for, at the risk of producing dramatic accidents.

Artificial intelligence brings with it not only enormous potential, but also a number of challenges, fears and risks.

There are five types of major risks of AI: bugs, cybersecurity, the "Sorcerer's Apprentice" phenomenon, shared autonomy, and possible negative socio-economic effects. Bugs refer to programming errors in the AI software. Just as there are dangers from cyberattacks, this risk is not different from current IT systems. "Sorcerer's Apprentice" risks refer to AI systems that perform incorrect actions when unclear instructions are present from the user.

Not a few experts are of the opinion that, in addition to global opportunities, the AI also poses global risks, which have been historically underestimated for a long time, such as those of nuclear technology. In addition, a knowledge-based risk analysis suggests that high potential damage levels should be taken very seriously even if the probabilities of occurrence were low.

The risks of AI are real and important:

Misuse risks. The discussion of risk is *not* dependent on the view that AI is now on a successful path toward superintelligence - though it gains urgency if such “success” is a no negligible possibility in the coming decades. It also gains urgency if the stakes are set high, even up to human extinction. If the stakes are so high as to include extinction of humankind, even a fairly small possibility of a disastrous outcome (say, 3%) is entirely sufficient to motivate the research. Consider that if there were a 3% possibility that a plane you are about to board will crash: that would be sufficient motivation for getting off.

The utility at stake in scientific or philosophical research is usually quite a bit lower. It appears that the outcome of superintelligence is more likely to be extreme: either extremely bad or extremely good for humanity. Therefore, there is the risk that "the machines take over" and this loss of control is a significant risk, perhaps an existential risk for humanity [14].

When Bostrom [5, 6] discuss about the transition to machine intelligence and existential risk, he is not referring to artificial intelligence systems as they exist in today's world.

He is thinking about the future and the advent of what might be called "artificial, global intelligence", the kind of global intelligence of reasoning and problem solving that allows us to dominate our environment as human species have done. We are not at that point at all at the moment and experts differ as to how long it will take us to achieve this. However, it is very likely to take more than a decade and even, according to some, several decades. However, if we do, the world will be so upset that it is difficult for us to imagine what it will look like.

By 2050 the probability of high-level machine intelligence (that surpasses human ability in nearly all respects) goes beyond the 50% mark, that is, it becomes more probable than not [10]. 2050 is also the year that "RoboCup" set itself for fielding a robot team that can beat the human football world champions (actually an aim that does not make much sense).

Like many people working in AI, Bishop [4] remains unimpressed by the discussion about risks of superintelligence because he thinks that there are principled reasons why machines will not reach these abilities: they will lack phenomenal consciousness, understanding, and insight.

International experts are sounding the alarm about the risks of the misuse of artificial intelligence (AI) by "rogue states, criminals, terrorists", in a report published recently. According to them, in the next ten years, the increasing effectiveness of AI is likely to increase cybercrime, but also to lead to the use of drones or robots for terrorist purposes. It is also likely to facilitate the manipulation of elections via social networks through automated accounts (bots).

This 100-page report entitled *The Malicious Use of Artificial Intelligence* was written by 26 experts in artificial intelligence, cybersecurity and robotics.

They work for universities (Cambridge, Oxford, Yale, Stanford) and non-governmental organizations (OpenAI, Center for a New American Security, Electronic Frontier Foundation). These experts call on governments and the various stakeholders to put in place measures to limit potential threats related to artificial intelligence.

One of the main ways in which AI could be transformative is by *enabling/accelerating the development of one or more enormously powerful technologies*. In the wrong hands, this could make for an enormously powerful tool of authoritarians, terrorists, or other power-seeking individuals or institutions.

The potential damage in such a scenario is nearly limitless (if transformative AI causes enough acceleration of a powerful enough technology), and could include long-lasting or even permanent effects on the world as a whole. It is worth asking whether there is anything we can do today to lay the groundwork for avoiding misuse risks in the future.

Accident risks. Tthere is a substantial class of potential “accident risks” that could rise (like misuse risks) to the level of *global catastrophic risks*.

We’ve seen substantial (though far from universal) concern that such risks could arise and no clear arguments for being confident that they will be easy to address. These risks are difficult to summarize.

The above risks could be amplified if AI capabilities improved relatively rapidly and unexpectedly, making it harder for society to anticipate, prepare for, and adapt to risks.

Some signs of interrogation arise:

What kinds of technical research are most important for reducing the risk of unexpected/undesirable outcomes from progress in artificial intelligence? Who are the best people to do this research?

What could be done - especially in terms of policy research or advocacy - to reduce risks from the weaponization/misuse of artificial intelligence?

What is the comparative size of the risk from intentional misuse of artificial intelligence (e.g. through weaponization) vs. loss of control of an advanced artificial intelligence agent with misaligned values?

Why the advantages of artificial intelligence outweigh the risks

The arguments against artificial intelligence (AI) are clearly anxiety-driven: fear of the unknown and fear of (information) intelligence. If Stephen

Hawking is believed, we have every reason to guard against the consequences of the further development of artificial intelligence, including the possibility of sealing the end of humanity.

However, the rise of the machines does not pose an immediate threat today, as artificial intelligence is still in a primitive stage. The most realistic scenario is probably the fear that artificial intelligence will destroy jobs.

AI and large-scale cybercrime

"We believe that the attacks that will be made possible by the increasing use of AI will be particularly effective, finely targeted and difficult to attribute," the report says.

To illustrate their fears, these specialists refer to several "hypothetical scenarios" of misguided use of AI. They point out that terrorists could modify commercially available AI systems (drones, autonomous vehicles) to cause crashes, collisions or explosions.

The authors imagine the case of a manipulated cleaning robot that surreptitiously slips among other robots in charge of cleaning in a Berlin ministry.

One day, the intruder would attack after visually recognizing the Minister of Finance. He would approach him and explode autonomously, killing his target.

In addition, "cybercrime, already on the rise, is likely to increase with the tools provided by AI," *Seán Ó hÉigeartaigh*, (or Sean O'Hegarty) director of the Centre for the Study of Existential Risk at Cambridge University, one of the report authors, told AFP. Spear phishing attacks could this become much easier to conduct on a large scale.

The most serious risk, although less likely, is political risk. We have already seen how people use technology to try to interfere in elections and democracy. If AI allows these threats to become stronger, more difficult to identify and attribute, this could pose major problems of political stability and perhaps contribute to outbreaks of war," says *Seán* Ó *hÉigeartaigh.*

With AI, it should be possible to make very realistic fake videos and this could be used to discredit politicians, the report warns. Authoritarian states will also be able to rely on AI to strengthen surveillance of their citizens, he added.

This is not the first time that concerns have been expressed about AI. As early as 2014, astrophysicist Stephen Hawking warned of the risks it could pose to humanity, surpassing human intelligence. Entrepreneur Elon Musk and others have also sounded the alarm.

Specific reports on the use of killer drones or how AI could affect US security have also been published. This new report provides "an overview of how AI creates new threats or changes the nature of existing threats in the areas of digital, physical and political security," explains Sean O'Hegarty.

Safety and reliability

The safety and reliability of software systems are a top priority in the areas of embedded systems, communication and application software.

Software technology must help ensure that data and communication channels are secure against unauthorized access and unintentional changes. In addition, the reliable and correct operation of components and services under normal operating conditions must be ensured.

The openness and flexibility of service-oriented architectures are a particular security challenge. Here, cross-instrument activities on ICT security for service-oriented architectures can form a decisive basis for the success of lead innovations.

Quality problems in large hardware and software systems limit the possibilities of functional innovation. Reliability must be increased here.

Google's banned AI applications

Technologies that could cause major harm: "When such a risk is identified, we will ensure that the benefits far outweigh the risks and that the integrated safety rules are sufficient".

Weaponing-related technologies: Google is committed to not developing technologies that can cause injuries to people (main objective or possible use of AI-based technology).

Surveillance technologies: mechanisms that collect information on individuals or use this information to monitor them, "in violation of globally accepted standards" will be banned. Google obviously does not want to be associated with the slogan "Big brother is watching you".

Technologies contrary to human rights: more generally, the Mountain View firm undertakes not to develop technologies that would contravene the principles of international law and human rights.

The European Economic and Social Committee *EESC* takes stock of its benefits and risks

Artificial intelligence will be a major upheaval for workers, explains the European Economic and Social Committee *EESC*. In a report, he unveils his recommendations on the subject, with an emphasis on the social side of course. Employment, ethics, influence on our decisions and education are all covered.

For several years, artificial intelligence - in the broadest sense of the term - has been growing rapidly in many areas. The latest example to date is Alpha Go's 3-0 clear victory (created by *DeepMind,* a Google subsidiary) over *Ke Jie,* which is considered the world's number one in the game of Go. However, artificial intelligence also raises many ethical, technological and societal questions.

As is often the case in such situations, the report begins by looking at the benefits that artificial intelligence could bring to everyday life: "making agriculture more sustainable and production processes less polluting, improving transport and workplace safety, making the financial system more stable, providing better quality medical care" and so on. In its synthesis, the European Union's advisory body goes so far as to assume that artificial intelligence "could even contribute to the eradication of disease and poverty". A whole program....

However, before it comes to that, the road is long and full of obstacles. The ***EESC*** identifies eleven areas where a framework needs to be defined: ethics, security, privacy, transparency and accountability, work, education and skills, (in) equality and inclusion, legislation and regulation, governance and democracy, war, and finally *superintelligence*.

The delicate issue of autonomous weapons and cyberwarfare like many other organizations, the EESC supports the call to ban autonomous weapon s systems, but believes that issue should be raised more broadly in *cyberwarfare*.

The idea is attractive on paper, but unfortunately there is little chance that all governments and private arms companies will leave out AI for the development of their weapons, digital or not. It is already more or less too late with drones and the automatic interception of messages on the Internet.

According to this report, it is "necessary to prevent AI from reaching the hands of people or regimes likely to use it for terrorist purposes"... certainly easier said that did.

The report then raises several questions about machine safety. He would like specific requirements to be put in place to answer and anticipate the following questions: is the algorithm reliable and effective, even in an unknown or unpredictable situation? Can it stop working or be hacked? The answers should not be taken lightly.

As we regularly see in IT security, there is no such thing as zero risk. However, the consequences of a failure could be quite different between an AI in charge of filtering sites for parental control and that of an autonomous car.

There is also the question of the legal personality of robots. The EESC has a clear-cut opinion on the subject: recognising it would represent an "unacceptable moral risk".

Conclusion

To sum up, the EESC's assessment focuses mainly on the impact that artificial intelligence will have on people's lives, both at work and in their private lives. They will be more or less important, but everyone will a priori be concerned, so we must be prepared for them. In any case, the Committee advocates that humans maintain control of AI in all circumstances, and prepare for the arrival of strong artificial intelligence.

It also recommends that Europe place itself at the heart of the debates, but this will not be an easy fight given that the vast majority of societies at the forefront of this field are outside our borders. There are certainly many start-ups and research centres in France (with a good reputation in general), but those with massive access to and use of data are located outside Europe.

References

1. Arnold, K. (2010): Signals: Applying academic analytics: EDUCAUSE Quarterly (33).

2. Arnold, K.; Pistilli, M. D. (Ed.) (2012): Course signals at Purdue: using learning analytics to increase student success. 2nd International Conference on Learning Analytics and Knowledge. Vancouver, BC, Canada, 29th-May 2nd. New York.
3. Bavelier, D., Green, S., Hyun Han, D., Renshaw, P., Merzenich, M., & Gentile, D. (2011). Viewpoint: Brains on Video Games. *Nature Reviews Neuroscience*, 12, 763–768.
4. Bishop, M. (2009). Why computers can't feel pain. *Minds and Machines*, *19*(4), 507–516.
5. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press
6. Bostrom, N. (2002). Existential Risks. *Journal of Evolution and Technology*, 9(1).
7. Daugherty, P.; Wilson, J. (2018): Human + Machine. Reimagining work in the age of AI, Boston, Mass.: Harvard Business Review Press
8. Eberl, U. (2018, 5 February): Was ist Künstliche Intelligenz – Was kann sie leisten? Aus Politik und Zeitgeschichte (APuZ), p. 8–14.
9. Lossau, Norbert (2017): Diese Super-Software bringt sich übermenschliche Leistungen bei. Available online at https://www.welt.de/wissenschaft/article169782047/Diese-Super-Software-bringt-sich-uebermenschliche-Leistungen-bei.html, last checked 22.01.2019.
10. Müller, V. C., and Bostrom, N. (2015). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence*. Berlin, Germany: Springer.
11. Neander, Joachim (1996): Computer schlägt Kasparow. Available online at https://www.welt.de/print-welt/article652666/Computer-schlaegt-Kasparow.html, last checked 22.01.2019.
12. Kapp, K. M. (2012). The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education. Pfeiffer.
13. Ramge, T. (2018): Mensch fragt, Maschine antwortet. Wie Künstliche Intelligenz Wirtschaft, Arbeit und unser Leben verändert. Aus Politik und Zeitgeschichte (APuZ). p. 15–21.
14. Sotala, K. and Yampolski R. V., Responses to catastrophic AGI risk: a survey, Physica Scripta, 2014. Technical Report. 2013. http://intelligence.org/files/ResponsesAGIRisk.pdf.
15. Spiegel Online o. V. (2015): Künstliche Intelligenz bewältigt 49 Atari-Spiele. Available online at http://www.spiegel.de/netzwelt/games/google-ki-computer-lernt-atari-spiele-wiespace-invaders-a-1020669.html, last checked am 14.02.2019.
16. Suich Bass, A. (2018, 31 March): GrAIt expectations. The Economist, Vol. 426, N° 9085, S. 3–12 (Special report 'AI in Business'). The World Bank Group (2016): Digital Dividends. World Development Report 2016. Available online https://openknowledge.worldbank.org/bitstream/handle/10986/23347/9781464806711.pdf, last checked 22.02.2019.
17. Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Global Catastrophic Risks*, 1, 303.
18. Weber, Christian (2017): Computer spielt Go gegen sich selbst – und wird unschlagbar. Available online at https://www.welt.de/wissenschaft/article169782047/Diese-Super-Software-bringt-sich-uebermenschliche-Leistungen-bei.html, last checked 21.02.2019.
19. http://www3.weforum.org/docs/WEF_FOW_Reskilling_Revolution.pdf, last checked am 22.10.2018.