Artificial Intelligence:
A Civilizational Challenge, A Generational Duty

by Geneviève Fieux-Castagnet[i] and Gérald Santucci[ii]

September 2019

Table of Contents

Artificial Intelligence: A Global Challenge of Disruptive Change

"Fifty thousand years ago with the rise of Homo sapiens sapiens.
Ten thousand years ago with the invention of civilization.
Five hundred years ago with the invention of the printing press.
Fifty years ago with the invention of the computer.
In less than thirty years, it will end."

Jaan Tallinn, Staring into the Singularity, 2007

The phrase "Artificial Intelligence" – in short AI – conjures concepts like robotics, facial recognition, chatbots, or autonomous vehicles. But in fact, well before the advent of even electricity, humans were fixated on creating artificial creatures, for example medieval alchemists who believed they could transform things into forms of artificial life ("philosopher's stone" or "elixir of life"), thus exploring immortality, or Judah Loew ben Bezalel, the late 16th century rabbi of Prague who invented the Golem out of clay, with the intention of protecting the Jewish people from attacks.

The age-old desire to expand intelligence was accelerated in the 1940s with the creation of the computer, providing scientists with the power to generate models capable of solving complex tasks, emulating functions previously only carried out by humans. The advent and maturation of artificial intelligence is now blurring the boundaries between human and technology. Where do we end, and where does technology begin?

The first neural network stemmed from the idea that rational thoughts could be transformed into formulaic rules. The first step towards artificial neural networks came in 1943 when Warren McCulloch, a neurophysiologist, and a young mathematician, Walter Pitts, wrote a paper on how neurons might work. They modelled a simple neural network with electrical circuits. The early timelines of Artificial Intelligence include genial inventors such as Alan Turing, Claude Shannon, or Ada Lovelace. From the work of the "Godfathers of AI", Yoshua Bengio, Yann LeCun and Geoffrey Hinton, working away on convolutional neural networks during the "AI Winter" of the 1970s, emerged the 21st century ground-breaking work being done in research and industry applications to solve some of the world's biggest challenges.

> There is currently no agreed definition of "Artificial Intelligence" (…) AI is used as an umbrella term to refer generally to a set of sciences, theories and techniques dedicated to improving the ability of machines to do things requiring intelligence. An AI system is a machine-based system that makes recommendations, predictions or decisions for a given set of objectives. It does so by: (i) utilising machine and/or human-based inputs to perceive real and/or virtual environments; (ii) abstracting such perceptions into models manually or automatically; and (iii) deriving outcomes from these models, whether by human or automated means, in the form of recommendations, predictions or decisions.[iii]

The two tables below give the main common terms used in Artificial Intelligence and some key milestones in the Artificial Intelligence Timeline.

Common terms used in Artificial Intelligence[iv]

Algorithm

A finite suite of formal rules/commands, usually in the form of a mathematical logic, that allows for a result to be obtained from input elements. They form the basis for everything a computer can do, and are therefore a fundamental aspect of all AI systems.

Artificial Intelligence (AI)

An umbrella term that is used to refer to a set of sciences, theories and techniques dedicated to improving the ability of machines to do things requiring intelligence.

AI system

A machine-based system that can make recommendations, predictions or decisions for a given set of objectives. It does so by utilising machine and/or human-based inputs to: (i) perceive real and/or virtual environments; (ii) abstract such perceptions into models manually or automatically; and (iii) use model interpretations to formulate options for outcomes.

AI system lifecycle

A set of phases concerning an AI system that involve: (i) planning and design, data collection and processing, and model building; (ii) verification and validation; (iii) deployment; (iv) operation and monitoring; and (v) end of life.

Automated decision-making

A process of making a decision by automated means. It usually involves the use of automated reasoning to aid or replace a decision-making process that would otherwise be performed by humans. It does not necessarily involve the use of AI but will generally involve the collection and processing of data.

Expert system

A computer system that mimics the decision-making ability of a human expert by following pre-programmed rules, such as 'if this occurs, then do that'. These systems fuelled much of the earlier excitement surrounding AI in the 1980s, but have since become less fashionable, particularly with the rise of neural networks.

Machine learning

A field of AI made up of a set of techniques and algorithms that can be used to "train" a machine to automatically recognise patterns in a set of data. By recognising patterns in data, these machines can derive models that explain the data and/or predict future data. When provided with sufficient data, a machine learning algorithm can learn to make predictions or solve problems, such as identifying objects in pictures or winning at particular games, for example. In summary, it is a machine that can learn without being explicitly programmed to perform the task.

Model

An actionable representation of all or part of the external environment of an AI system that describes the environment's structure and/or dynamics. The model represents the core of an AI system. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms.

Neural network

Also known as an artificial neural network, this is a type of machine learning loosely inspired by the structure of the human brain. A neural network is composed of simple processing nodes, or 'artificial neurons', which are connected to one another in layers. Each node will receive data from several nodes 'above' it, and give data to several nodes 'below' it. Nodes attach a 'weight' to the data they receive, and attribute a value to that data. If the data does not pass a certain threshold, it is not passed on to another node. The weights and thresholds of the nodes are adjusted when the algorithm is trained until similar data input results in consistent outputs.

Deep learning

A more recent variation of neural networks, which uses many layers of artificial neurons to solve more difficult problems. Its popularity as a technique increased significantly from the mid-2000s onwards, as it is behind much of the wider interest in AI today. It is often used to classify information from images, text or sound.

Personal data

Information relating to an identified or identifiable natural person, directly or indirectly, by reference to one or more elements specific to that person. Sensitive personal data concern personal data relating to "racial" or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, as well as genetic data, biometric data, data concerning health or concerning sex life or sexual orientation.

Personal data processing

Any operation or set of operations performed or not using automated processes and applied to personal data or sets of data, such as collection, recording, organisation, structuring, storage, adaptation or modification, retrieval, consultation, use, communication by transmission, dissemination or any other form of making available, linking or interconnection, limitation, erasure or destruction.

Artificial Intelligence Timeline

1308 Catalan poet and theologian Ramon Llull publishes *Ars generalis ultima* (The Ultimate General Art), further perfecting his method of using paper-based mechanical means to create new knowledge from combinations of concepts

1666 Mathematician and philosopher Gottfried Leibniz publishes *Dissertatio de arte combinatoria* (On the Combinatorial Art), following Ramon Llull in proposing an alphabet of human thought and arguing that all ideas are nothing but combinations of a relatively small number of simple concepts

1763 Thomas Bayes develops a framework for reasoning about the probability of events – Bayesian inference will become a leading approach in machine learning

1854 George Boole argues that logical reasoning could be performed systematically in the same manner as solving a system of equations

1898 Nikola Tesla makes a demonstration of the world's first radio-controlled vessel equipped "with a borrowed mind"

1914 The Spanish engineer Leonardo Torres y Guevedo demonstrates the first chess-playing machine, capable of king and rook against king endgames without human intervention

1921 Czech write Karel Capek introduces the word "robot" (from "robota", meaning work) in his play R.U.R. (Rossum's Universal Robots)

1929 Makoto Nishimura designs *Gakutensoku*, Japanese for "learning from the laws of nature", the first robot built in Japan, which could change its facial expressionand move its head and hands via an air pressure mechanism

1943 Warren S. McCulloch and Walter Pitts publish "A Logical Calculus of The Ideas Immanent in Nervous Activity" in the *Bulletin of Mathematical Biophysics*, in which paper they discuss networks of idealized and simplified artificial "neurons" and how they might perform simple logical functions

1949 Edmund Berkeley publishes *Giant Brains: Or Machines that Think*

1949 Donald Hebb publishes *Organisation of Behaviour: A Neuropsychological Theory* in which he proposes a theory about learning based on conjectures about neural networks and the ability of synapses to strengthen or weaken over time

1950 Claude Shannon's "Programming a Computer for Playing Chess" is the first published article on developing a chess-playing program

1950 TURING TEST In "Computing Machinery and Intelligence", computer scientist Alan Turing proposes "the imitation game", a test for machine intelligence – if a machine can trick humans into thinking it is human, then it has intelligence

1955 (31 August) ARTIFICIAL INTELLIGENCE BORN The term "artificial intelligence" is coined by computer scientist John McCarthy (Dartmouth College), Marvin Minsky (Harvard University), Nathaniel Rochester (IBM) and Claude Shannon (Bell Telephone Laboratories) to describe "the science and engineering of making intelligent machines"

1955 (December) THE LOGIC THEORIST, considered by many to be the first AI program, is developed by Allen Newell and Herbert Simon

1956 (July-August) THE DARTMOUTH SUMMER RESEARCH PROJECT on Artificial Intelligence is discussed for a month of brainstorming, in Vermont, USA, at the initiative of John McCarthy, with the intention of drawing the talent and expertise of others interested in machine intelligence

1958 John McCarthy develops programming language Lisp

1959 Arthur Samuel coins the term "machine learning", reporting on programming a computer "so that it will learn to play a better game of checkers than can be played by the person who wrote the program"

1961 UNIMATE First industrial robot, Unimate, goes to work on an assembly line in a General Motors plant in New Jersey, replacing humans

1964 ELIZA Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966 SHAKEY The "first electronic person" from Stanford, Shakey, is a general-purpose mobile robot which is able to reason about its own actions

1997 DEEP BLUE Deep Blue, a chess-playing computer from IBM defeats world chess champion, Garry Kasparov

1998 KISMET Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999 AIBO Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002 ROOMBA First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011 SIRI Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011 WATSON IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy!* by defeating champions Brad Rutter and Ken Jennings

2014 EUGENE Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

2014 ALEXA Amazon launches Alexa, an intelligent virtual assistant with a voice interface that can complete shopping tasks

2016 TAY Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017 ALPHAGO Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2*170) of possible positions

2019 ALPHASTAR Google's AI agent defeats pro StarCraft II players

The main – sometimes unspoken – reason why Artificial Intelligence has been so much discussed over the past few years is because it represents a disruptive phenomenon that is radically changing the existing social and economic systems. It is, as Siebel put it, an "evolutionary punctuation" that could be "intimately linked with the widespread death of species (…) Evolutionary punctuations are responsible for the cyclic nature of species: inception, diversification, extinction, repeat."[v] In the past 500 million years, there have been five global mass extinction events that left only a minority of species surviving. The voids in the ecosystem were then filled by massive speciation of the survivors – after the Cretaceous-Tertiary event, for example, most well-known for the elimination of the dinosaurs, came the reign of mammals. Geologists argue that

disruptive punctuations are on the rise, and the periods of stasis in between punctuations are fading.

Evolutionary Mass Extinction Events				
Ordovician-Silurian	Late Devonian	End Permian	Triassic-Jurassic	Cretaceous-Tertiary
Between 445 and 415 million years ago	375 million years ago	252 million years ago ("the Great Dying")	200 million years ago	66 million years ago
				
Percentage of Species Extinct				
86%	75%	96%	80%	76%
Groups affected				
Brachiopods, trilobites, graptolites and moss animals	Coral-sponge reefs in tropics, fish and plankton, trilobites	Marine invertebrates, land plants, plankton, insects and all life	Large amphibians, crurotarsans (not crocodiles), insects and conodonts	Dinosaurs (not birds), pterosaurs, plesiosaurs, mosasaurs and ammonoids

Source: based on Siebel Thomas M., op.cit., page 5

Taking the concept of "punctuated equilibrium" as a relevant framework for thinking about disruption in today's economy, we could say that we are in the midst of an evolutionary punctuation:

- The average stay of corporations listed in the S&P 500 sharply declined from 61 years in 1958 to 25 years in 1980 and 18 years in 2011. At the present rate of evolution, it is estimated that three-quarters of today's S&P 500 will be replaced by 2027[vi].
- Since 2000, 52 percent of the companies in the Fortune 500 have either gone bankrupt, been acquired, ceased to exist, or dropped off the list[vii]. It is estimated that 40 percent of the companies in existence today will shutter their operations in the next ten years.

"Enabled by cloud computing", Siebel writes, "a new generation of AI is being applied in an increasing number of use cases with stunning results. And we see IoT everywhere – connecting devices in value chains across industry and infrastructure and generating terabytes of data every day." In fact, Big Data, Artificial Intelligence, cloud computing and the Internet of Things, promise to transform the technoscape to a degree

comparable to those of the five mass extinctions. Just as the Internet revolutionized business in the 1990s and 2000s, the ubiquitous use of artificial intelligence will transform business in the coming decades. While big tech companies like Google, Netflix and Amazon are early adopters of AI for consumer-facing applications, virtually every type of organisation, private and public, will soon use AI throughout their operations.

The impact of Artificial Intelligence on the global economy is expected to be enormous, thus justifying Siebel's idea of an "evolutionary punctuation" which, by the combined play of innovation and competition, will deeply and lastingly change the techno-economic landscape. According to PwC, by 2030 global GDP could increase by 14%, or $15.7 trillion, because of AI. If today AI contributes $1 trillion to global GDP, its contribution to the global GDP will increase 16-fold in the next decade, until 2030, according to the head of Russia's largest bank Sberbank Herman Gref. And finally, according to McKinsey Global Institute, while the introduction of steam engines during the 1800s boosted labour productivity by an estimated 0.3% a year, the impact from robots during the 1990s around 0.4%, and the spread of IT during the 2000s 0.6%, additional global economic activity from the deployment of artificial intelligence is projected to be a whopping 1.2% a year, or 16% by 2030, amounting to $13 trillion! If delivered, this impact would compare well with that of other general-purpose technologies through history.

The science fiction of yesterday quickly becomes reality. AI has already altered the way we think and interact with each other every day. Whether it's in healthcare, education, or manufacturing, AI yields a great deal of success in nearly every industry. Whether it's AI's effect on start-ups and investments, robotics, big data, virtual digital assistants, and so forth, AI is surging as the main driver of change in the next decades:

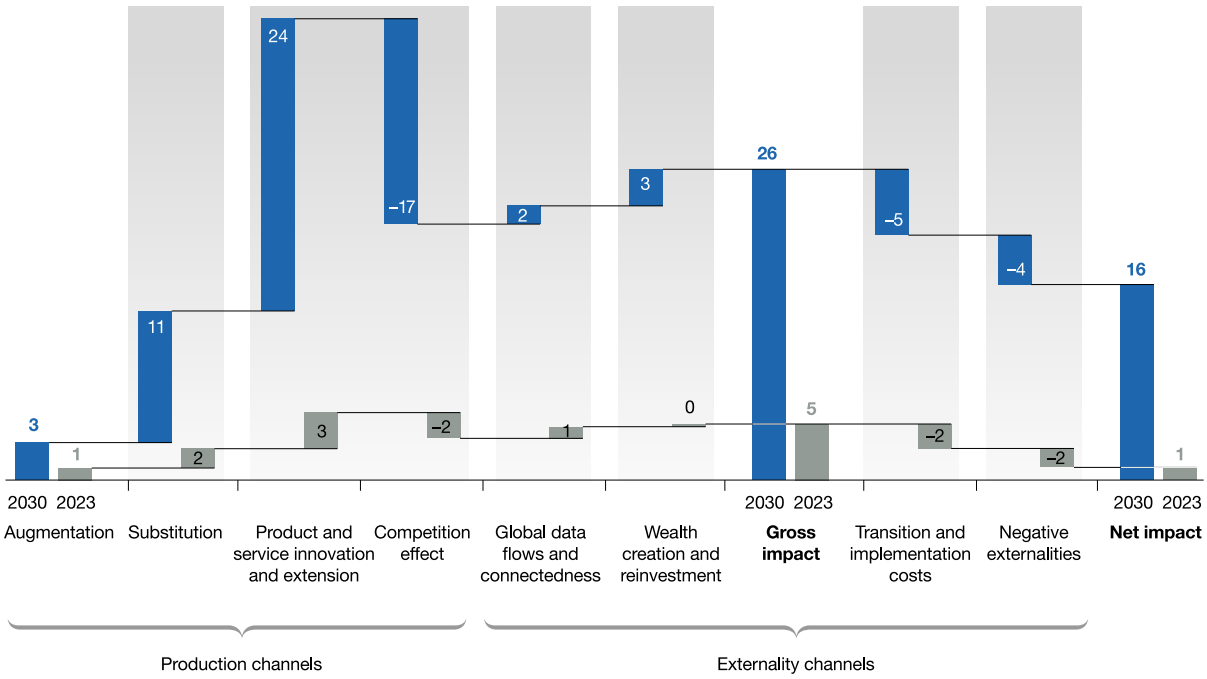
- Global GDP will grow by $15.7 trillion by 2030 thanks to AI.
- By 2025, the global AI market is expected to be almost $60 billion.
- AI can increase business productivity by 40% (source: Accenture).
- AI start-ups grew 14 times and investment in AI start-ups grew 6 times since 2000.
- Already 77% of the devices we use feature one form of AI or another.
- Cyborg technology – or Nano Bio Info Cogno (NBIC) – will help us overcome physical and cognitive impairments.
- Google analysts believe that by 2020, robots will be smart enough to mimic complex human behaviour like jokes and flirting

A number of factors, including labour automation, innovation, and new competition, affect AI-driven productivity growth. Micro factors, such as the pace of adoption of AI, and macro factors, such as the global connectedness or labour-market structure of a country, both contribute to the size of the impact. McKinsey examined seven possible channels of impact: the first three relate to the impact of AI adoption on the need for, and mix of, production factors that have direct impact on company productivity; the other four are externalities linked to the adoption of AI related to the broad economic environment and the transition to AI. The results are presented in the exhibit below.

Artificial intelligence's net economic impact has seven channels.

SIMULATION

Breakdown of economic impact, cumulative boost vs today, %



Note: Numbers are simulated figures to provide directional perspectives rather than forecasts. Figures may not sum to 100%, because of rounding.

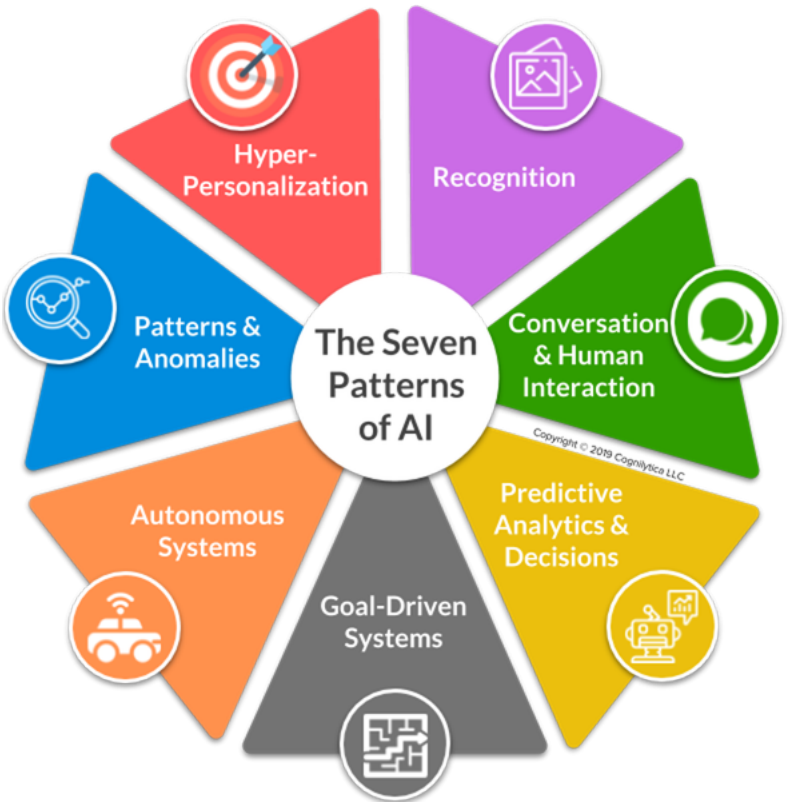
McKinsey&Company | **Source:** McKinsey Global Institute analysis

The process of disruption brought about by AI is something that we've only seen a few times in history. Electricity, the Industrial Revolution, and the Internet are the three that we can all think of, but this time it will be much faster. "Compared to electricity, it took decades for the electrical grid to be built up, and then people had to invent new ways to use the electricity—air conditioners, refrigerators, and so on. It took over a century, and only now we're getting electric cars. But with AI, these engines work on the cloud, on the Internet, and you can program them and connect them with the data that's also on the cloud. And engineers can access them. Open source also allows people to build on each other's work. Compared to electricity, which took decades if not a century to become fully pervasive, AI can be pervasive in years. And this will bring tremendous value, tremendous efficiency, but also tremendous disruptions. Because it will change business practices, it will cause companies to go out of business, it will take away people's jobs, especially if they are routine. It's going to be a very exciting but also a very challenging decade ahead."[viii]

The "datafication" of everything will continue to accelerate, powered by the intersection of separate advances in infrastructure, cloud computing, artificial intelligence, open source and the overall digitalization of our economies and lives. Data science, machine learning and AI allow to add layers of intelligence into many applications, which are now increasingly running in production in all sorts of consumer and B2B products. As those technologies continue to improve and to spread beyond the initial group of early adopters into the broader world economic and societal corners, the discussion is shifting from purely technical matters into a necessary conversation around impact on our economies, societies and lives. In a world where data-driven automation becomes the rule (automated products, automated vehicles,

automated enterprises), what is the new nature of work? How do we handle the social impact? How do we think about privacy, security, freedom?

As highlighted by Kathleen Walch[ix], if the use cases for AI are many, from autonomous vehicles, predictive analytics applications, facial recognition, to chatbots, virtual assistants, cognitive automation, and fraud detection, and so forth, there is also commonality to all these applications. She argues that all AI use cases fall into one or more of seven common patterns (see figure below).



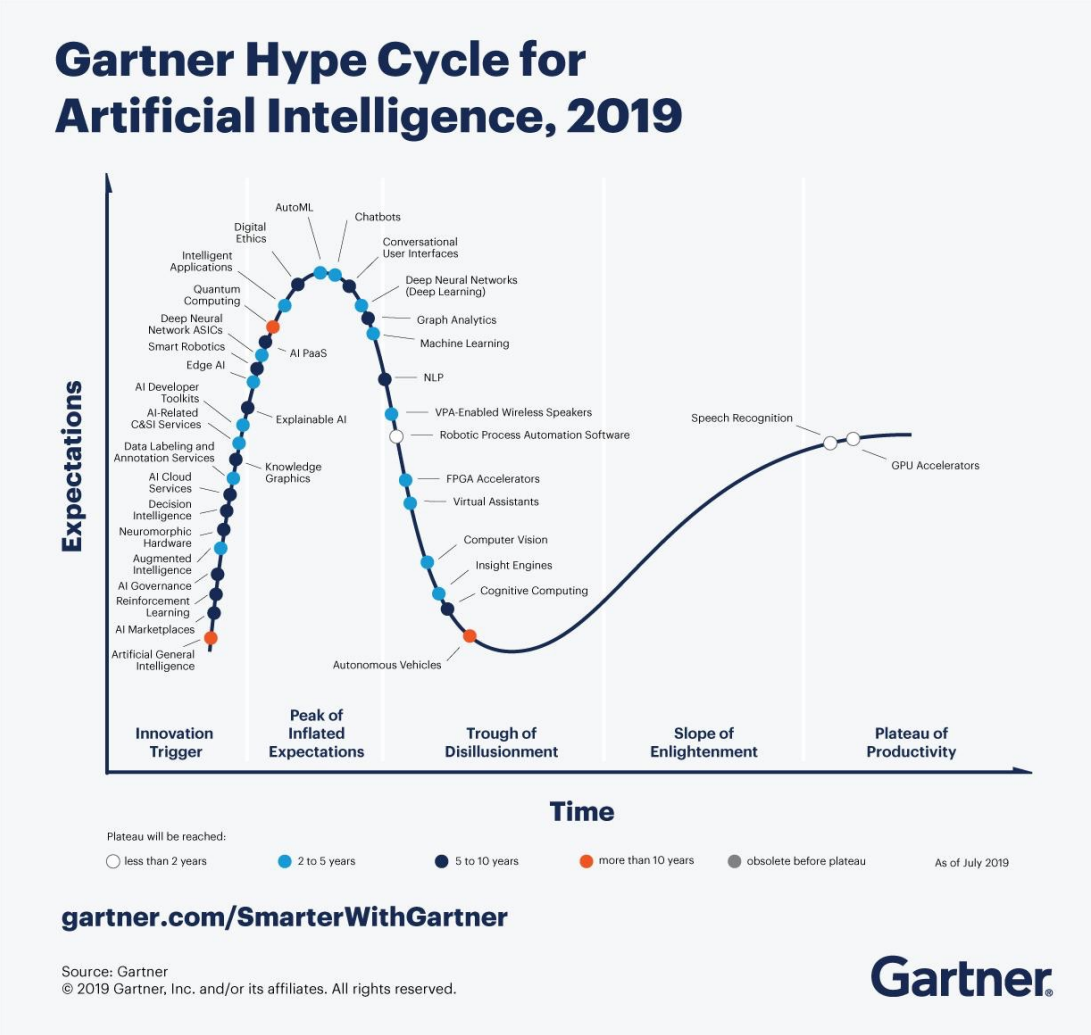
Artificial Intelligence is not just a further discipline or set of technologies and applications. It is a real disruption, an “evolutionary punctuation”, which is going to metamorphize the world economies and societies.

Artificial Intelligence Global Outlook

"The race to become the global leader in artificial intelligence (AI) has officially begun."

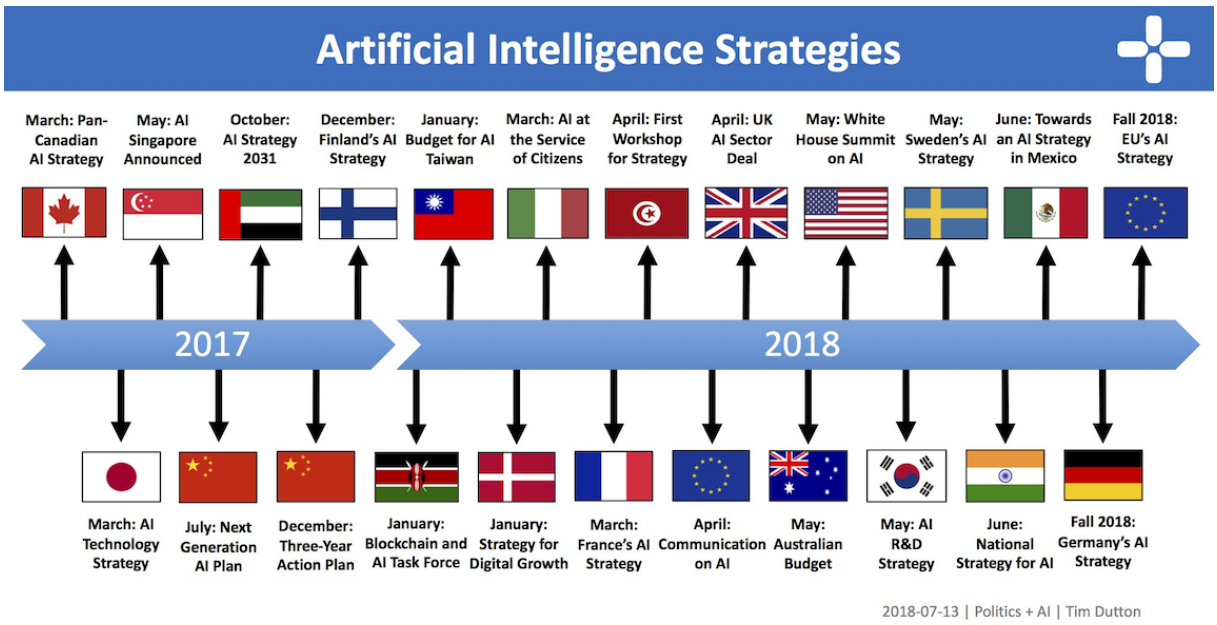
Tim Dutton, *An Overview of National AI Strategies*, 28/05/2018

The Gartner AI Hype Cycle evaluates the business maturity of various dimensions underpinning AI. AI is currently overhyped in the entire world as a socioeconomic phenomenon. Scientists, industrialists, experts, media, governments, and individuals each have an opinion about AI, sometimes based on vague ideas of what it really is. Gartner's Hype Cycle views AI as a pervasive paradigm and an umbrella term for many innovations at the different stages of value creation. The traffic jam at the "Peak of Inflated Expectations" is increasing, as early implementers grow in numbers, but production implementations remain scarce. A long line of high-promise innovation profiles at the "Innovation Trigger" phase are approaching the traffic jam at the Peak of Inflated Expectations, indicating that the AI hype will continue. None of the profiles in this Hype Cycle is obsolete before "Plateau", but not all will survive, and many will morph into something different depending on the choices and decisions that customers are making today.



Aware of the economic importance of Artificial Intelligence, many countries and regions, including Canada, China, Denmark, the EU Commission, Finland, France, India, Italy, Japan, Mexico, the Nordic-Baltic region, Singapore, South Korea, Sweden, Taiwan, the UAE, and the UK have released strategies to promote its use and development. No two strategies are alike, with each focusing on different aspects of AI policy: scientific research, talent development, skills and education, public and private sector adoption, ethics and inclusion, standards and regulations, and data and digital infrastructure.

All countries are seeking to foster fast development of AI technologies, in particular by financing an ever-vibrant ecosystem of start-ups, products and projects[x].



It is a remarkable fact that so many countries of the world are addressing the development of Artificial Intelligence at both public sector and private sector level. Such an engagement has not been observed for the Internet of Things – Europe led initiatives in this field from the mid-2000s (2006-2012), followed by China (2009) and then other countries – or more recently for 5G, though these technologies are also disruptive in technological, socio-economic and ethical terms. We could argue that AI is considerably impacting the further development and deployment of the Internet of Things, and this is true, but this doesn't explain the scope and scale of public programs for AI across the world compared to other transformative technologies.

Another interesting feature concerns the importance given to ethics in addressing the AI challenges. All national and regional programs are primarily designed for research and innovation; however, the ethical dimension is rarely neglected. Everything happens as if humans were committed to reaping the full expected benefits of AI while at the same time carrying out a deep reflection on the normative questions around how AI should confront ethical, societal and legal dilemmas and what applications of AI, specifically with regards to security, privacy and safety, should be considered permissible. Such a balanced approach to an advanced technology is rather new – when work on ethics in the Internet of Things was proposed for discussion by an Expert Group set up by the European Commission in 2010-2012, no follow up was given on

the premise that the issue of ethics had to be addressed, like other IoT-specific challenges, at the level of the global discussions about the Internet.

We could be worried that the proliferation of public programs, including recommendations on AI ethics, leads to a waste of time and public money. Indeed, a lot of synergy could have been created by addressing the AI challenges, and at least the precompetitive aspects, in a more cooperative way by a large number of countries.

Some examples of national strategies for the development and deployment of AI are described below.

Australia

Australia does not yet have an artificial intelligence strategy. However, in the 2018–2019 Australian budget, the government announced a four-year, AU$29.9 million investment to support the development of AI in Australia. The government will create a Technology Roadmap, a Standards Framework, and a national AI Ethics Framework to support the responsible development of AI. In addition, in the its 2017 innovation roadmap, Australia 2030: Prosperity Through Innovation, the government announced that it will prioritise AI in the government's forthcoming Digital Economy Strategy. This report is expected to be released in the second half of 2018.

2018: The federal government's 2018-19 budget earmarks AU$29.9 million over four years to strengthen Australia's capability in artificial intelligence and machine learning (ML). The funding will be split between programs at the Department of Industry, Innovation and Science, which will receive the lion's share of the funding, the CSIRO and the Department of Education and Training. The government said it would fund the development of a "technology roadmap" and "standards framework" for AI as well as a national AI Ethics Framework. Together they will "help identify opportunities in AI and machine learning for Australia and support the responsible development of these technologies." The investment will also support Cooperative Research Centre projects, PhD scholarships, and other initiatives to increase the supply of AI talent in Australia. The budget funding for AI forms part of the government's broader Australian Technology and Science Growth Plan.

January 2019: White Paper on "Artificial Intelligence: Governance and Leadership, Australian Human Rights Commission and World Economic Forum 2019.

05/04/2019 → 31/05/2019: discussion paper, developed by CSIRO's Data61 and designed to encourage conversations about AI ethics and inform the Government's approach to AI ethics in Australia[xi].

Canada

The Canadian government launched the 5-year Pan-Canadian Artificial Intelligence (AI) Strategy in its 2017 Budget with the allocation of CAD$125 million[xii]. Canada was actually the first country to release a national AI strategy. The effort is being led by a group of research and AI institutes: Canadian Institute for Advanced Research (CIFAR), the Alberta Machine Intelligence Institute, the Vector Institute, and the Montreal Institute for Learning Algorithms (MILA).

The AI Strategy has four major goals:

- Increase the number of outstanding artificial intelligence researchers and skilled graduates in Canada;
- Establish interconnected nodes of scientific excellence in Canada's three major centres for artificial intelligence in Edmonton, Montreal and Toronto-Waterloo;
- Develop global thought leadership on the economic, ethical, policy and legal implications of advances in artificial intelligence;
- Support a national research community on artificial intelligence.

The Strategy is expected to help Canada enhance its international profile in research and training, increase productivity and collaboration among AI researchers, and produce socio-economic benefits for all of Canada. Existing programs of the Strategy include the funding of three AI centres throughout the country, supporting the training of graduate students, and enabling working groups to examine the implications of AI to help inform the public and policymakers.

Separately, Canadian Prime Minister Justin Trudeau and French President Emmanuel Macron announced the creation of an international study group for AI on June 7, 2018, ahead of the G7 Summit in Quebec. The independent expert group will bring together policymakers, scientists, and representatives from industry and civil society. It will identify challenges and opportunities presented by AI, and determine best practices to ensure that AI fulfils its potential of creating economic and societal benefits. Trudeau and Macron said they would create a working group to make recommendations about how to form the panel and will invite other nations to join.

Canada's AI strategy is distinct from other strategies because it is primarily a research and talent strategy. It's initiatives – the new AI Institutes, CIFAR Chairs in AI, and the National AI program – are all geared towards enhancing Canada's international profile as a leader in AI research and training. The CIFAR AI & Society Program examines the policy and ethical implications of AI, but the overall strategy does not include policies found in other strategies such as investments in strategic sectors, data and privacy, or skills development. That is not to say that the Canadian government does not have these policies in place, but that they are separate from, rather than part of, the Pan-Canadian Artificial Intelligence Strategy.

What Success Will Look Like

1) Canada will have one of the most skilled, talented, creative and diverse workforces in the world, with more opportunities for all Canadians to get the education, skills and work experience they need to participate fully in the workforce of today, as they—and their children – prepare for the jobs of tomorrow.

2) Canadian businesses will be strong, growing and globally competitive—capable of becoming world leaders in their fields, leading to greater investment and more job creation in Canada.

3) Canada will be on the leading edge of discovery and innovation, with more ground-breaking research being done here at home, and more world class researchers choosing to do their work at Canadian institutions.

4) Canadian academic and research leadership in artificial intelligence will be translated into a more innovative economy, increased economic growth, and improved quality of life for Canadians.

(Department of Finance Canada, Budget 2017. Chapter 1. Canada's Innovation and Skills Plan, 2017)

China

China announced its ambition to lead the world in AI theories, technologies, and applications in its July 2017 plan, *A Next Generation Artificial Intelligence Development Plan*. The plan is the most comprehensive of all national AI strategies, with initiatives and goals for R&D, industrialisation, talent development, education and skills acquisition, standard setting and regulations, ethical norms, and security. It is best understood as a three-step plan:

- first, make China's AI industry "in-line" with competitors by 2020;
- second, reach "world-leading" in some AI fields by 2025; and
- third, become the "primary" centre for AI innovation by 2030.

By 2030, the government aims to cultivate an AI industry worth 1 trillion RMB, with related industries worth 10 trillion RMB. The plan also lays out the government's intention to recruit the world's best AI talent, strengthen the training of the domestic AI labour force, and lead the world in laws, regulations, and ethical norms that promote the development of AI.

Since the release of the Next Generation Plan, the government has published the Three-Year Action Plan to Promote the Development of New-Generation Artificial Intelligence Industry. This plan builds on the first step of the Next Generation plan to bring China's AI industry in-line with competitors by 2020. Specifically, it advances four major tasks:

(1) focus on developing intelligent and networked products such as vehicles, service robots, and identification systems,
(2) emphasize the development AI's support system, including intelligent sensors and neural network chips,
(3) encourage the development of intelligent manufacturing, and
(4) improve the environment for the development of AI by investing in industry training resources, standard testing, and cybersecurity.

In addition, the government has also partnered with national tech companies to develop research and industrial leadership in specific fields of AI and will build a $2.1 billion technology park for AI research in Beijing.

If the United States remain today stronger in deep tech, just like with autonomous vehicles or robots walking like a human, China has more data and more consumers who are more active. Competitiveness in AI depends more on the volume of data than on the quality of scientists. And if data is the new oil of the Digital Economy, in particular of Artificial Intelligence, China is for AI the equivalent of the Organisation of the Petroleum Exporting Countries (OPEC)! Furthermore, China enjoys a strong and dynamic entrepreneurial ecosystem marked by a culture of resilience and the idea that "the winner takes all".

The cultural aspect is particularly important here – no country in the past ever moved from imitator to innovator in only 10 years. Not so long ago, China was viewed as a mere imitator in the technology world with its companies more likely to copy western products than develop their own innovative ideas. But following years of government support, long-term visions and strategies, strong and steady GDP growth, and massive investment in education, the outlook has changed. "China's ten-year miracle – moving

from copycat to innovator – is basically a cycle that began with a larger market attracting more money" (Kai-Fu Lee). But there is more. In Chinese history, there were four great inventions: the compass, gunpowder, paper, and printing technology, plus dozens of other noteworthy inventions which have made people's lives easier around the world. Let's also remember that until the 15th century China's naval technology was the most advanced in the world: Admiral Zheng He, a Muslim eunuch who eventually became commander of the Chinese Navy as his master, used large ships under the service of the Chinese emperor in an effort to explore the vast Chinese empire and to bring wealth back to his country. The first of his six voyages was in 1402 and by the end of his sixth he had sailed west around the Indian ocean all the way to the coast of Africa and brought vast amounts of gold for the emperor[xiii]. Therefore, we can find roots to China's ability to innovate far back in history. Besides China's well-known companies like Huawei, Alibaba, Baidu, ByteDance or Tencent, a large number of new national AI champions are emerging, such as Xiaomi (smart home hardware), JD.com (smart supply chain), Qihoo 360 Technology (online safety), Hikvision (AI infrastructure and software), Megvii (image perception) and Yitu (image computing), which are fuelling innovation and probably hold the key to the future of AI.

China will intensify its efforts to reduce the exposure of its industry to U.S. suppliers. At the same time, it will clone a better Internet, with a better Cloud and a better system of global influence (organised around the "New Silk Road").

Germany

Germany, long seen as an industrial powerhouse with great engineering capabilities, also has lots of AI talent. Berlin is currently touted as Europe's top AI talent hub. Cyber Valley, a new tech hub region southern Germany is hoping to create new opportunities for collaboration between academics and AI-focused businesses. Germany also has a very notable automobile industry with a long track record of innovation. Almost half of the worldwide patents on autonomous driving are held by German automotive industry companies and suppliers such as Bosch, Volkswagen, Audi and Porsche.

The Federal Government's Artificial Intelligence (AI) strategy was jointly developed in 2018 by the Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs based on suggestions taken from a nationwide online consultation[xiv].

The government wants to strengthen and expand German and European research in AI and focus on the transfer of research results to the private sector and the creation of AI applications. Proposed initiatives to achieve this include new research centres, Franco-Germany research and development collaboration, regional cluster funding, and support for SMEs and start-ups. The proposed plan is comprehensive and also includes measures to attract international talent, respond to the changing nature of work, integrate AI into government services, make public data more accessible, and promote the development of transparent and ethical AI. Overall, the government wants "*AI made in Germany*" to become a globally recognized seal of quality.

In addition, Germany already has a number of related policies in place to develop AI. Principally, the government, in partnership with academia and industry actors, focuses on integrating AI technologies into Germany's export sectors. The flagship program has been Industry 4.0, but recently the strategic goal has shifted to smart services, which relies more on AI technologies. The German Research Centre for AI (DFKI) is a

major actor in this pursuit and provides funding for application-oriented research. Other relevant organizations include the Alexander von Humboldt Foundation, which promotes academic cooperation and attracts scientific talent to work in Germany, and the *Plattform Lernende Systeme*, which brings together experts from science, industry, politics, and civic organizations to develop practical recommendations for the government.

The government also announced in June 2018 a new commission to investigate how AI and algorithmic decision-making will affect society. It consists of 19 MPs and 19 AI experts and is tasked with developing a report with recommendations by 2020 (a similar task force released a report on the ethics of autonomous vehicles in June 2017[xv]).

Definition of "artificial intelligence" in Germany

There is no definition of AI which generally valid or used consistently by all stakeholders. The Federal Government's AI Strategy is based on the following understanding of AI:

In highly abstract terms, AI researchers can be assigned to two groups: "strong" and "weak" AI. "Strong" AI means that AI systems have the same intellectual capabilities as humans, or even exceed them. "Weak" AI is focused on the solution of specific problems using methods from mathematics and computer science, whereby the systems developed are capable of self-optimisation. To this end, aspects of human intelligence are mapped and formally described, and systems are designed to simulate and support human thinking.

The Federal Government is oriented its strategy to the use of AI to solve specific problems, i. e. to the "weak" approach:

1. Deduction systems, machine-based proofs: deduction of formal statements from logical expressions, systems to prove the correctness of hardware and software

2. Knowledge-based systems: methods to model and gather expertise; software to simulate human expertise and to support experts (previously designated "expert systems"); to some extent coupled with psychology and cognitive sciences;

3. Pattern analysis and pattern recognition: inductive analytical processes in general, machine learning in particular

4. Robotics: autonomous control of robotic systems, i.e. autonomous systems

5. Smart multimodal human-machine interaction: analysis and "understanding" of language (in conjunction with linguistics), images, gestures and other forms of human interaction.

India

India has taken a unique approach to its national AI strategy by focusing on how India can leverage AI not only for economic growth, but also for social inclusion. NITI Aayog, the government think tank that elaborated a report, calls this approach *#AIforAll*. The strategy, as a result, aims to (1) enhance and empower Indians with the skills to find quality jobs; (2) invest in research and sectors that can maximize economic growth and

social impact; and (3) scale Indian-made AI solutions to the rest of the developing world.

NITI Aayog provides over 30 policy recommendations to invest in scientific research, encourage reskilling and training, accelerate the adoption of AI across the value chain, and promote ethics, privacy, and security in AI. Its flagship initiative is a two-tiered integrated strategy to boost research in AI. First, new Centres of Research Excellence in AI (COREs) will focus on fundamental research. Second, the COREs will act as technology feeders for the International Centres for Transformational AI (ICTAIs), which will focus on creating AI-based applications in domains of societal importance. In the report, NITI Aayog identifies healthcare, agriculture, education, smart cities, and smart mobility as the priority sectors that will benefit the most socially from applying AI. The report also recommends setting up a consortium of Ethics Councils at each CORE and ICTAI, developing sector specific guidelines on privacy, security, and ethics, creating a National AI Marketplace to increase market discovery and reduce time and cost of collecting data, and a number of initiatives to help the overall workforce acquire skills.

Strategically, the government wants to establish India as an "AI Garage," meaning that if a company can deploy an AI in India, it will then be applicable to the rest of the developing world.

Japan

Japan was one of the first countries to develop a national AI strategy. Based on instructions from Prime Minister Abe during the Public-Private Dialogue towards Investment for the future on 12 April 2016, the Strategic Council for AI Technology was established to develop "research and development goals and a roadmap for the industrialization of artificial intelligence." The 11-member council had representatives from academia, industry, and government, including the President of Japan's Society for the Promotion of Science, the President of the University of Tokyo, and the Chairman of Toyota.

The plan, the Artificial Intelligence Technology Strategy, was released in March 2017[xvi]. The strategy is notable for its Industrialization Roadmap, which envisions AI as a service and organizes the development of AI into three phases: (1) the utilization and application of data-driven AI developed in various domains, (2) the public use of AI and data developed across various domains, and (3) the creation of ecosystems built by connecting multiplying domains.

The strategy applies this framework to three priority areas of Japan's Society 5.0 initiative – productivity, health, and mobility – and outlines policies to realize the industrialization roadmap. These policies include new investments in R&D, talent, public data, and start-ups.

The Japanese government strategy, however, must meet a difficult challenge. While China, adapting itself to the speed of change, has implemented cashless payments, vehicle dispatch service, and unmanned convenience stores and hotels, and is reportedly only one step short of putting unmanned delivery vehicles and self-driving buses into service, Japan lags in AI use and Internet literacy, with most Internet users using smartphones merely for making calls, social networking services and downloading games, music and animation. They are not making full use of them as

Internet terminals. Perhaps due to Japan's poor Internet literacy, the equivalent of giant platform businesses as represented by the United States' GAFA (Google, Amazon, Facebook and Apple), and China's Baidu, Alibaba, Tencent and Alipay are almost non-existent in this country.

Japan is also home to one of the largest venture funds in the industry, Softbank, which has over $100 billion to invest in industry-shifting AI companies.

South Korea

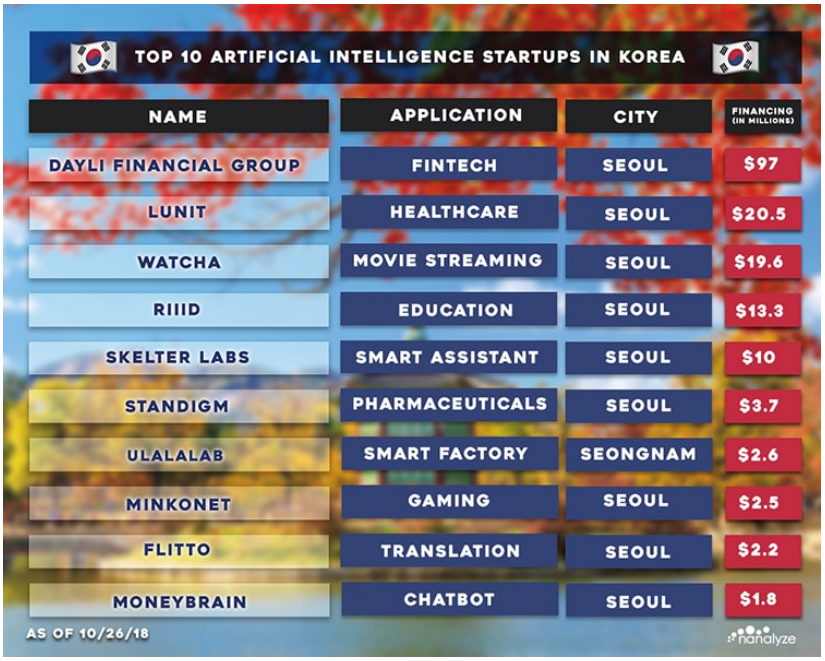
South Korea has ambitious plans around Artificial Intelligence. In 2016, it famously hosted the match where DeepMind's AlphaGo defeated Go's world champion Lee Sedol, a South Korea native.

Home to huge tech conglomerates like Samsung, LG, and Hyundai, South Korea showed their commitment to growing AI by announcing in 2018 a $2 billion investment program to strengthen AI research in the country[xvii]. The aim is to join the global top four nations in AI capabilities by 2022 by pursuing the following priorities:

- establishing at least six new schools with focus on AI and the training of more than 5,000 engineers;
- funding large-scale AI projects related to medicine, national defence, and public safety;
- starting an AI R&D challenge similar to those developed by the US Defence Advanced Research Projects Agency (DARPA).

South Korea's Ministry of Science and ICT (MSICT) proposed the investment strategy as a way to close the gap between Korea's AI tech and China's. The MSITC, which defines South Korea's R&D strategy in three categories – human resources, technology and infrastructure – also estimated that Korean AI tech is currently 1.8 years behind US AI tech.

The table below indicates the top funded AI start-ups in Korea:



United Arab Emirates (UAE)

Arab States have recently left the dangerous rims of marginalisation from the mainstream economic growth of the world and have become visible in the Digital Economy radar. It is indeed essential for Arab countries – a potential market of more than 420 million people – that they are neither isolated from world trends nor disenfranchised from active and autonomous participation in development.

The UAE launched its AI strategy in October 2017[xviii]. It has been the first country in the Middle East to create an AI strategy and the first in the world to create a Ministry of Artificial Intelligence. The strategy is the first initiative of the larger UAE Centennial 2071 Plan[xix] and its primary goal is to use AI to enhance government performance and efficiency. The government will invest in AI technologies in nine sectors: transport, health, space, renewable energy, water, technology, education, environment, and traffic. In doing so, the government aims to cut costs across the government, diversify the economy, and position the UAE as a global leader in the application of AI.

On 16-17 December, 2018, the First Arab Digital Economy Conference took place in Abu Dhabi to provide a Common Vision for the Region, a Global Outlook (i.e. why Arab countries need to work on common agendas), and insights on the application of AI to Education and Labour, Future Cities, Healthcare, Finance, and Sustainable Development. This Conference was a defining moment for fostering steady cooperation between Arab States and other world regions to share information and knowledge and to create synergy.

United Kingdom

The British government released an "AI Sector Deal" in April 2018[xx], aiming to position the UK as a global leader in AI, as part of its broader industrial strategy.

The Sector Deal policy paper covers policies to boost public and private R&D, invest in STEM education, improve digital infrastructure, develop AI talent, and lead the global conversation on data ethics. Major announcements include over £300 million in private sector investment from domestic and foreign technology companies, the expansion of the Alan Turing Institute, the creation of Turing Fellowships, and the launch of the Centre for Data Ethics and Innovation. The Centre in particular is a key program of the initiative, as the government wants to lead the global governance of AI ethics. A public consultation and a call for the chair of the Centre was launched in June 2018.

A few days before the release of the Sector Deal, the UK's House of Lords' Select Committee on AI had published a report titled, "AI in the UK: ready, willing and able?"[xxi] The report was the culmination of a ten-month inquiry that was tasked with examining the economic, ethical, and social implications of advances in AI. It outlined a number of recommendations for the government to consider, including calls to review the potential monopolization of data by technology companies, incentivize the development of new approaches to the auditing of datasets, and create a growth fund for UK SMEs working with AI. The report also argued that there is an opportunity for the UK to lead the global governance of AI and recommended hosting a global summit in 2019 to establish international norms for the use and development of AI. In June 2018, the government released an official response to the House of Lords that comments on each of the recommendations in the report.

The House of Lords' report outlined five key principles to form the basis of a cross-sector AI code, which can be adopted nationally:

The "AI Code" in UK

1. Artificial intelligence should be developed for the common good and benefit of humanity.

2. Artificial intelligence should operate on principles of intelligibility and fairness.

3. Artificial intelligence should not be used to diminish the data rights or privacy of individuals, families or communities.

4. All citizens should have the right to be educated to enable them to flourish mentally, emotionally and economically alongside artificial intelligence.

5. The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence.

United States

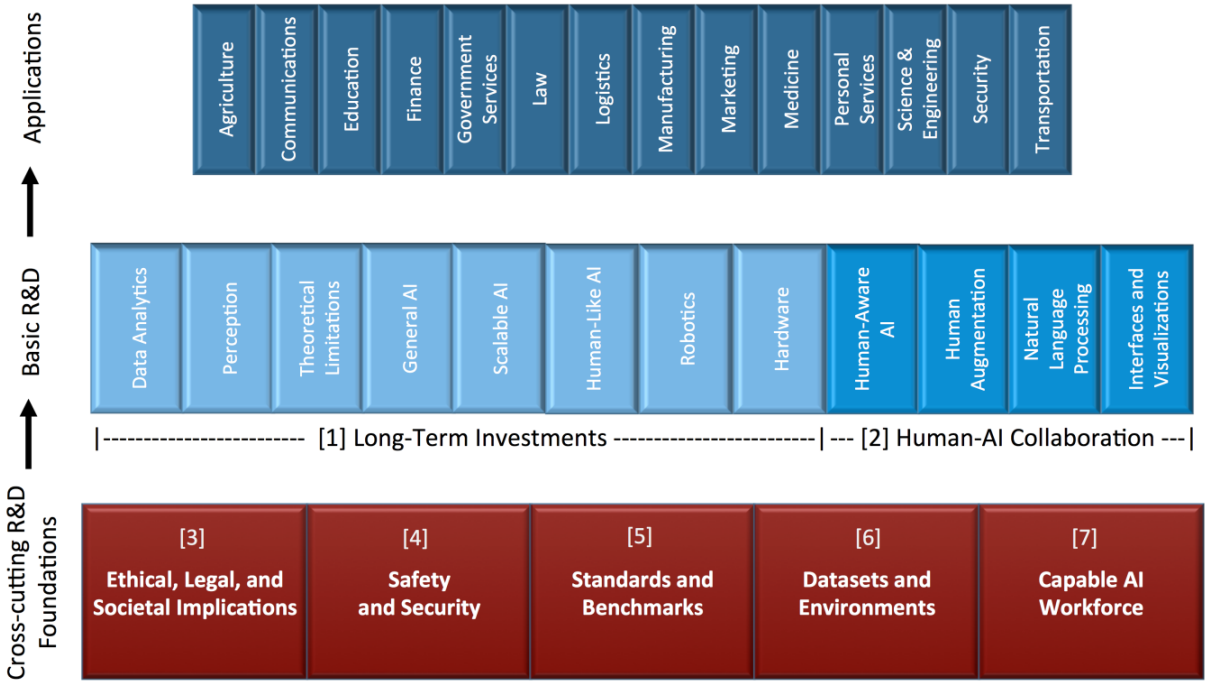
On March 19, 2019, the US federal government launched AI.gov[xxii] to make it easier to access all of the governmental AI initiatives currently underway. The site is the best single resource from which to gain a better understanding of US AI strategy.

US President Donald Trump issued an Executive Order launching the American AI Initiative on February 11, 2019. The Executive Order explained that the Federal Government plays an important role not only in facilitating AI R&D, but also in promoting trust, training people for a changing workforce, and protecting national interests, security, and values. And while the Executive Order emphasizes American leadership in AI, it is stressed that this requires enhancing collaboration with foreign partners and allies.

Guiding the United States in its AI R&D investments is the National AI R&D Strategic Plan: 2019 Update, which identifies the critical areas of AI R&D that require Federal investments. Released by the White House Office of Science and Technology Policy's National Science and Technology Council, the Plan defines several key areas of priority focus for the Federal agencies that invest in AI, including:

- continued long-term investments in AI;
- effective methods for human-AI collaboration;
- understanding and addressing the ethical, legal, and societal implications for AI;
- ensuring the safety and security of AI;
- developing shared public datasets and environments for AI training and testing;
- measuring and evaluating AI technologies through standards and benchmark;
- better understanding the National AI R&D workforce needs; and
- expanding public-private partnerships to accelerate AI advances.

Organisation of the US AI R&D Strategic Plan



The process for creating this Plan began in August of 2018, when the Administration directed the Select Committee on AI to refresh the 2016 National AI R&D Strategic Plan to account for significant recent advancements in AI, and to ensure that Federal R&D investments remain at the forefront of science and technology.

In September 2019, agencies for the first time reported their non-defense R&D investments in AI according to the 2019 National AI R&D Strategic Plan, through the Networking and Information Technology Research & Development (NITRD) Supplement to the President's FY2020 Budget[xxiii]. This new AI R&D reporting process provides an important mechanism and baseline for consistently tracking America's prioritization of AI R&D going forward. This report also provides insight into the diverse and extensive range of nondefense Federal AI R&D programs and initiatives.

The strategic priorities are the following:

- Coordinate long-term Federal investments in AI R&D, such as algorithms to enable robust and reliable perception, general AI systems that exhibit the flexibility and versatility of human intelligence, and combinatorial optimization to obtain prodigious performance.
- Promote safe and effective methods for human–AI collaboration to achieve optimal efficiency and performance by developing advanced AI techniques for human augmentation and improved visualization and AI-human interfaces.
- Develop methods for designing AI systems that align with ethical, legal, and societal goals, and behave according to formal and informal human norms.
- Improve the safety and security of AI systems so that they operate in a controlled, well-defined, and well-understood manner.

- Develop shared public datasets and environments for AI training and testing to increase the benefits and trustworthiness of AI.
- Improve measurement and evaluation of AI technologies through benchmarks and standards to address safety, reliability, accuracy, usability, interoperability, robustness, and security.
- Grow the Nation's AI R&D workforce to ensure the United States leads the automation of the future.
- Expand public-private partnerships to strengthen the R&D ecosystem.

Artificial Intelligence in Europe: Renaissance, Reprieve or Fall?

"Modern Europe was shaped by losing the Ancient World (fall of Constantinople, 1453), by discovering the New World (1492), and by switching out the World (Copernic, 1473-1543). Two centuries later, Europe is going to change the World."

Edgar Morin, Penser l'Europe, Éditions Gallimard, collection Folio/Actuel, 1987-1990, chapitre 3, page 51.

Since Artificial intelligence has become an area of strategic importance and a key driver of economic development, bringing solutions to many societal challenges from treating diseases to minimising the environmental impact of farming, but also raising socio-economic, legal and ethical issues that need to be carefully addressed, the European Commission strives to make that all EU countries join forces to stay at the forefront of this technological revolution, in order to ensure competitiveness and to shape the conditions for its development and use (ensuring respect of European values).

The European Commission's commitment to Artificial Intelligence has gained a new momentum from 2018 onwards, though its roots can be traced back in the following related policy developments:

19/04/2016: Digitising European Industry (DEI) strategy – realising the potential of digitisation, where robotics and AI are key drivers (Digital Innovation Hubs, Platforms, Liability, Safety, Data protection, Skills)

01/12/2016: Digital Jobs and Skills Coalition

16/02/2017: European Parliament resolution with recommendations to the Commission on Civil Law Rules on robotics – written answer by the EC

26/09/2017: G7 ICT and industry Ministers' Declaration – "Making the Next Production Revolution Inclusive, Open and Secure – We recognise that the current advancements in new technologies, especially Artificial Intelligence (AI), could bring immense benefits to our economies and societies. We share the vision of human-centric AI which drives innovation and growth in the digital economy. We believe that all stakeholders have a role to play in fostering and promoting an exchange of perspectives, which should focus on contributing to economic growth and social well-being while promoting the development and innovation of AI. We further develop this vision in the "G7 multi-stakeholder exchange on Human Centric AI for our societies" set forth in Annex 2 to this declaration."

19/10/2017: Digital Summit Tallinn conclusions – "to put forward a European approach for AI"

From this point events proceeded very quickly, as can be seen from the following table, which presents the key milestones of the EU's strategy, including notably a major investment effort on Research and Innovation and a commitment to a Human-Centric Artificial Intelligence.

The European Union Strategy on Artificial Intelligence: Milestones		
Date	*Instrument*	*Purpose*
10 April 2018	Digital Day Declaration	Member States sign up to cooperate on AI.
25 April 2018	Communication COM(2018) 237	*Artificial Intelligence for Europe*. The European Commission strategy places people at the centre of the development of AI – human-centric AI. It is a three-pronged approach to (i) boost the EU's technological and industrial capacity and AI uptake across the economy, (ii) prepare for socio-economic changes, and (iii) ensure an appropriate ethical and legal framework "The approach to AI described in this document shows the way forward and highlights the need to join forces at European level, to ensure that all Europeans are part of the digital transformation, that adequate resources are devoted to AI and that the Union's values and fundamental rights are at the forefront of the AI landscape. Together, we can place the power of AI at the service of human progress."
1 June 2018	High-Level Expert Group on Artificial Intelligence	Following an open selection process, the European Commission appoints 52 experts (30 men and 22 women) to a High-Level Expert Group on Artificial Intelligence (AI HLEG) comprising representatives from academia, civil society, as well as industry. The AI HLEG has as a general objective to support the implementation of the European Strategy on Artificial Intelligence. This includes the elaboration of recommendations on future-related policy development and on ethical, legal and societal issues related to AI, including socio-economic challenges.
1 June 2018	The European AI Alliance	The European Commission announces the establishment of the European AI Alliance – a broad multi-stakeholder platform which complements and supports the work of the AI High Level Expert Group, in particular in preparing draft AI ethics guidelines and ensuring competitiveness of the European Region in the burgeoning field of Artificial Intelligence. This forum engages more than 3000 European citizens and stakeholders in a dialogue on the future of AI in Europe.
6 June 2018	Digital Europe programme	Digital Europe programme proposed: €2.5 billion for the deployment of AI.
7 June 2018	Horizon Europe programme	Horizon Europe programme proposed: largest EU R&I programme ever with €100 billion.
7 December 2018	Coordinated Plan	The European Commission presents a coordinated plan prepared with Member States to create synergies, pool data, and increase joint investments. The aim is to foster cross-border cooperation and mobilise all players to increase public and private investments to at least €20 billion annually over the

		next decade. The European Commission doubled its investments in AI in Horizon 2020 and plans to invest €1 billion annually from Horizon Europe and the Digital Europe Programme, in support notably of common data spaces in health, transport and manufacturing, and large experimentation facilities such as smart hospitals and infrastructures for automated vehicles and a strategic research agenda.
1 January 2019	AI4EU project	AI4EU will mobilise the whole European AI ecosystem and already unites 79 partners in 21 countries in a network across Europe and will provide access to relevant AI resources in the EU for all users. Led by THALES, France, this project receives a total funding of €20 million over the next 3 years.
8 April 2019	Communication COM(2019) 168	Building Trust in Human-Centric Artificial Intelligence
8 April 2019	Ethics guidelines	The High-Level Expert Group on AI presents Ethics Guidelines for Trustworthy Artificial Intelligence. This follows the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.
9 April 2018	Digital Day	Digital Day - Presentation and discussions on AI ethics guidelines[xxiv].
18 June 2018	Stakeholder Summit	The European Economic and Social Committee, together with the European Commission, organise a stakeholder summit on artificial intelligence (AI) bringing together businesses representatives, academia, workers, citizens, policy makers and NGO's to discuss the next steps to advance the EU strategy on AI. In two plenary sessions and three parallel working groups, experts, speakers, panellists and attendees are invited to elaborate on the three pillars of the EU strategy on AI: legal and ethical challenges; socio-economic impact; industrial competitiveness.
26 June 2019	Piloting phase of the ethics guidelines for Trustworthy AI	Organisations can test the assessment list for trustworthy artificial intelligence, developed by the AI HLEG on behalf of the Commission, and see how robust it is in practice. Over 300 organisations have already expressed interest in doing so since the release of the expert group's Ethics Guidelines for Trustworthy AI. An online survey has been created to gather feedback on the assessment list and will be open until 1 December 2019. Best practice examples for assessing the trustworthiness of AI can also be shared through the European AI Alliance.

26 June 2019	Policy and investment recommendations for Trustworthy AI in Europe	The AI HLEG presents to the European Commission a list of 33 recommendations deemed to help AI have major impact on citizens, businesses, administrations and academia. The focus is on ensuring sustainability, growth, competitiveness and inclusion while empowering, benefiting and protecting individuals. The recommendations[xxv] will help the Commission and Member States to update their joint coordinated plan on AI at the end of 2019.
Early 2020	Evaluation	Evaluation report of pilot phase.

The European Commission's approach on AI, currently based on three pillars – being ahead of technological developments and encouraging uptake by the public and private sectors; prepare for socio-economic changes brought about by AI; and ensure an appropriate ethical and legal framework –, has undoubtedly unfolded very rapidly since the beginning of 2018. It encompasses several aspects:

- Research – Development – Innovation
- Testing – Benchmarking – Safety – Certification
- Ethical issues (European values of dignity and privacy)
- Legal issues ("fit for purpose")
- Social issues (awareness – acceptance – social sciences and humanities – trust – perception)
- Economic issues (re/up-skilling / robots to help us)
- Involvement of all stakeholders (academia, industries, SMEs, end-users, social scientists, lawyers, civil society, agencies, institutions, etc.)

The European Commission should be credited for the tremendous work it has already accomplished whose main characteristic is that it covers all the dimensions of the AI challenge, from science and technology to legal and ethical issues.

However, the question remains whether the EU is capable to meet the ambitious objectives it has set for itself.

In its 2018 Communication, the European Commission acknowledged that "overall, Europe is behind in private investments in AI which totalled around €2.4-3.2 billion in 2016, compared with €6.5-9.7 billion in Asia and €12.1-18.6 billion in North America." It went on by stressing the strengths of Europe: "Europe is home to a world-leading AI research community, as well as innovative entrepreneurs and deep-tech start-ups (founded on scientific discovery or engineering). It has a strong industry, producing more than a quarter of the world's industrial and professional service robots (e.g. for precision farming, security, health, logistics), and is leading in manufacturing, healthcare, transport and space technologies – all of which increasingly rely on AI. Europe also plays an important role in the development and exploitation of platforms providing services to companies and organisations (business-to-business), applications to progress towards the 'intelligent enterprise' and e-government."

The European Commission also mentioned the long trail of R&I efforts in AI:

"AI has featured in the EU research and development framework programmes since 2004 with a specific focus on robotics. Investments increased to up to €700 million for 2014-2020, complemented by €2.1 billion of private investments as part of a public-private partnership on robotics. These efforts have significantly contributed to Europe's leadership in robotics. Overall, around €1.1 billion has been invested in AI-related research and innovation during the period 2014-2017 under the Horizon 2020 research and innovation programme, including in big data, health, rehabilitation, transport and space-oriented research. Additionally, the Commission has launched major initiatives which are key for AI. These include the development of more efficient electronic components and systems, such as chips specifically built to run AI operations (neuromorphic chips), world-class high-performance computers, as well as flagship projects on quantum technologies and on the mapping of the human brain."

Stakes are high for Europe in order to avoid repeating the mistakes of its handling of other key digital technologies, such as mobility and the Internet of Things.

In the field of mobility, Europe was once leading in technology and implementation, in particular thanks to Framework Research Programmes 2 & 3 that allowed the RACE programme (R&D in Advanced Communications technologies in Europe) to push the deployment of 3G. Europe was then at the heart of innovation in the mobile space. Cooperation on GSM standards brought it a leading position – with Nokia, Siemens, Ericsson, Alcatel and Philips to name just a few, Europe had the world's technology leaders. They invested their economic success in designing the technologies of the future at that time: 3G (early 2000s) and 4G (launched in 2009). Their North American and Asian competitors lagged behind because they lacked scale, scope and a common approach. However, Europe's mistakes in rolling out 3G, (spectrum auctions that focussed mainly on delivering maximum revenue for governments, not on creating a healthy mobile ecosystem) weakened the position of mobile network operators, limiting their ability to compete worldwide. By 2016, these European companies had ceased being consumer brands. They merged and barely held on as mobile infrastructure providers. They now compete with the new leaders – China's Huawei and ZTE, Apple and Samsung. Has Europe learned from its mistakes? At least, the European Commission is aware that the fragmented emergence and slow rollout of the 4G services should not be repeated in the case of 5G. Although the total investment related to 5G deployment in 28 EU Member States is estimated at €56.6 billion, analysts expect that by 2025 5G will generate more than €113.1 billion euros annually across the four major verticals that will take advantage of 5G early on: automotive, health, transport and energy. 5G introduction has the potential to create 2.3 million jobs in Europe and this opportunity cannot be missed out[xxvi]. But isn't it too late for 5G, which represents a quantum leap in connectivity with the potential to unlock advanced IoT applications in areas running from self-driving cargo trucks to software-driven management of "smart" cities to interconnected drones and remote surgery? A recent survey of chief technology officers by the consultancy McKinsey suggested that European firms expect 5G to go live only in 2021-2022, while counterparts in the US and China expect to have the infrastructure in place before then — in some cases by 2020[xxvii].

In the field of the Internet of Things (IoT), the European Commission was a pioneer in the initial recognition of the concept, the formulation of a convincing rationale for collaborative R&D at EU level, and the recognition of the need of a specific policy and regulatory framework. The phrase "Internet of Things" appeared in the 2007 European

Commission's Communication on RFID[xxviii] – the first time in an official document of a public sector institution[xxix]. Between 2007 and 2012, the EU clearly set the pace in IoT discussions, technical work, and regulatory initiatives, thus triggering the involvement of other countries such as China (2009) and USA (2012)[xxx]. But in November 2012 the European Commission's DG CONNECT took the sudden decision to terminate the work of the Expert Group on the Internet of Things (IoT-EG), which had been set up in August 2010 with the objective of supporting DG CONNECT in the drafting of a Recommendation on the Governance of the Internet of Things[xxxi]. The political mindset at that time was that IoT policy had to be mainstreamed into the broader discussions on the Internet. The result was that two years had been lost in discussions among experts, which led to a bitter feeling of unfinished business. As a result, the European Commission had to refocus its IoT work on the mere management of R&I contracts and suffered a loss of political momentum in Europe-wide and global discussions.

The current resolve of the European Commission to tackle the Artificial Intelligence challenge therefore deserves to be praised. Yet, it remains to be seen if this effort will support a renaissance of Europe in the digital domain or will more modestly be a last stand before a lasting withdrawal. The engagement of the European Commission is a necessary but not sufficient condition. Europe should perhaps seek inspiration in the example of China which has been capable in about 20 years to develop an ecosystem different from the one of North America. Europe has a critical mass of research facilities, in both industry and academia, sufficient skills, and also its own innovation model (e.g., "Digital Europe" Research and Innovation programme, General Data Protection Regulation), but it has been unable so far to create a European equivalent of Google, Facebook or Alibaba[xxxii]. There are many reasons for this, among which the brain drain of the best talents and the lack of a venture capital culture in industry and the financial sector. Moreover, despite the 1993 European Single Market (i.e. the free movement of goods, capital, services and labour across the EU) and the ongoing work on the Digital Single Market (DSM[xxxiii]), Europe remains too fragmented.

If the EU can draw applaud for having come out with a comprehensive plan for AI since 2018, it still lacks major companies like Microsoft, Google, Apple, Alibaba, Baidu, and Tencent that have major data sets with which to train AI algorithms. The EU is likely to devote more money to smaller scale research and application projects, but lacks a well-developed infrastructure to support AI breakthroughs. The UK has some advantages in AI, with a robust education system and innovation centre in London that has attracted some leading AI firms such as Deepmind. The UK, however, and also France, lack large companies and access to data, so they are likely to continue to play a minor role in contributing to the development of AI capabilities.

The role of the Commission, under the presidency of Ursula von der Leyen, will be critical in the next few years, in particular as regards its willingness and ability to design and implement a genuine "industrial policy" breaking the existing silos that are formed by the Competition policy, the Regional policy, the Internal Market policy, and the Research & Innovation policy in order to make all of these consistent and contributing to an ambitious vision of the future of Europe. The new Commission will also need to pick up where the last one left off on pressing questions of online platform regulation (competition, liability, hate speech, algorithmic accountability) and navigate the conflict between the US and China, which is in part a conflict about digital infrastructure and technological sovereignty[xxxiv].

The idea of the European Commission to work on Artificial Intelligence along a Coordinated Plan carried out with the Member States is already a promising idea. Not much can be achieved by the European Commission alone, it is therefore essential that the EU level and the national level advance together, in association with stakeholders, primarily industry, in order to give a chance to the possibility of a true level playing field between Europe, US, China and other global competitors. The most dismal scenario would be to have the EU AI strategy being actually *another* plan besides different national plans of European countries like France, Germany or the UK. Such strategy fragmentation would sign the end of Europe as a credible contender in the global AI race and, more generally, in the whole domain of digital technologies.

Artificial Intelligence: cultural differences and civilizational risks

"Culture refers to what is special and specific in a society whereas Civilisation relates to what can be acquired and transmitted from one society to another. Culture is generic, civilisation generalisable; culture develops through return to roots and loyalty to one's special principles, civilisation by accumulating knowledge, i.e. by progressing."

Edgar Morin, op. cit., page 82

Until recently, the conversation on AI focused on its hype – how AI would be good for healthcare, mobility, security, manufacturing, etc. If we learned anything from 2018 onwards it's that we need to be more sceptical about where AI could be heading us as a civilisation. The early warnings from Elon Musk and Stephen Hawking have been followed on and amplified by several other experts from science, engineering and business[xxxv]. We begin to become aware that we are living in an era where not only our conception of the labour market and our ways of life are challenged, but also where the very existential threats to our survival as human species will be increasingly felt and debated.

The end of labour?

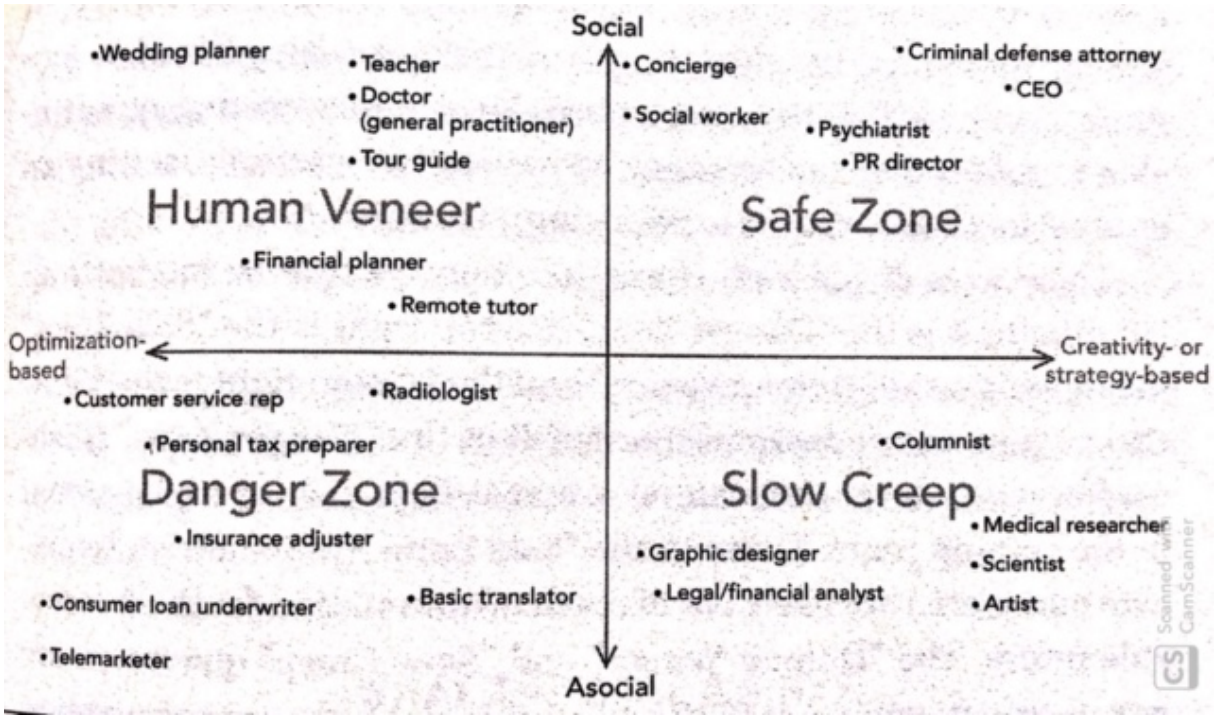
Several studies have filled in the conversational space recently about the impact – beneficial or destructive – of Artificial Intelligence on jobs. Some experts prefer to focus on the jobs that AI will create, thus leveraging human capabilities in different ways, while on the contrary others argue that more and more people will be removed from the workplace. In fact, as of today, nobody can pretend we know the truth – we can only exchange glass-half-full-or-empty arguments.

Kai-Fu Lee argues that "when it comes to job replacement, AI's biases don't fit the traditional one-dimensional metric of low-skill versus high-skill labour. Instead, AI creates a mixed bag of winners and losers depending on the particular content of job tasks performed. He proposed two X-Y graphs, one for physical labour and one for cognitive labour, each of which dividing the charts into four quadrants:

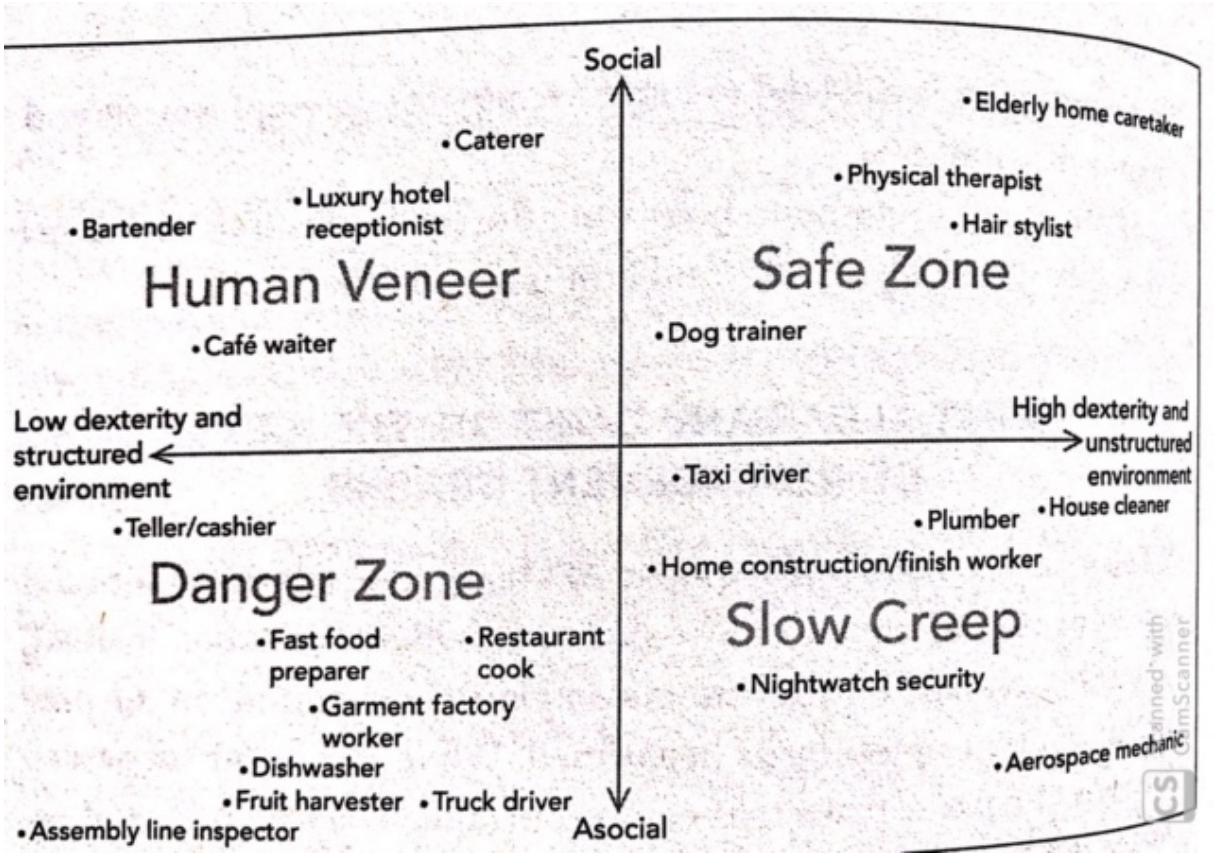
- the "Danger Zone", (bottom-left),
- the "Safe Zone" (top-right),
- the "Human Veneer" (top-left), and
- the "Slow Creep" (bottom right)

These graphs give us a basic heuristic for understanding what kinds of jobs are at risk, without prejudging the impact on total employment on an economy-wide level.

Risk of replacement: Cognitive Labour



Risk of Replacement: Physical Labour



Source: Kai-Fu LEE, AI Super-Powers: China, Silicon Valley, and the New World Order, Houghton Mifflin Harcourt, Boston New York, 2018, pages 155-156

Evaluating the impact of AI, and more generally automation, on the labour market is not an easy task. Past experience shows that new technologies take time to generate productivity and wage gains. Even though automation eventually increases the overall size of the economy, it is also likely to boost inequality in the short run, by pushing some people into lasting unemployment or lower-paid jobs. Therefore, in the short-term automation may lead to unrest and opposition, which could in turn slow the pace of automation and productivity growth, thus leaving everyone worse off in the long run.

The "glass-half-full" argument, supported for example by Paul R. Daugherty, H. James Wilson and Nicola Morini Bianzino, highlights three already-emerging job categories spurred by AI[xxxvi]: trainers (i.e. the people who are doing the data science and the machine-learning engineering), explainers (i.e. the people who explain how AI itself is working and the kind of outcomes it's generating), and sustainers (i.e. the people who make sure that AI not only behaves properly at the outset of a new process but continues to produce the desired outcomes over time). The third category seems particularly important at the time when various experts voice big concerns over the ethical risks of AI – sustainers indeed are responsible for tackling unintended consequences of AI algorithms and how they may affect the public. It is also worth noting that regulation stimulates the second category, e.g., by some estimates, as many as 75,000 data protection officer (DPO) positions will be created in response to the EU's General Data Protection Regulation (GDPR) around the globe[xxxvii].

This optimistic view of AI is based on the past perspective: "As with other technology advances, writes Thomas M. Siebel, AI will soon create more jobs than it destroys. Just as the internet eliminated some jobs through automation, it gave rise to a profusion of new jobs – web designers, etc."[xxxviii] In a report released on December 13, 2017, Gartner said "by 2020, AI will generate 2.3 million jobs, exceeding the 1.8 million that it will wipe out. In the following five years to 2025, net new jobs created in relation to AI will reach 2 million."

On the contrary, other experts prefer to see the "glass-half-empty". The AI-enabled computer is engaged in head-on competition with man on the labour market: it replaces man in complex functions where the human brain was deemed indispensable so far. Almost 47% of US jobs could be computerized within one or two decades[xxxix]. It isn't only manual labour jobs that could be affected, but also many cognitive tasks over two waves of computerization, with the first substituting computers for people in logistics, transportation, administrative and office support and the second affecting jobs depending on how well engineers crack computing problems associated with human perception, creative and social intelligence[xl]. Indeed, even researchers at MIT foresee dismal prospects for many types of jobs as new technologies are increasingly adopted not only in manufacturing, clerical, and retail work but in professions such as law, financial services, education, and medicine. Mckinsey Global Institute reported in 2017 that by 2030 75 million to 375 million workers (3 to 14% of the global workforce) would need to switch occupational categories[xli]. This is not surprising since AI has the potential to increase productivity by about 40 per cent, and is projected to contribute up to $15.7 trillion to the global economy in 2030, more than the current output of China and India combined. With the impact on productivity being competitively transformative – businesses that fail to adapt and adopt will quickly find themselves uncompetitive – all workers will need to adapt, as their occupations evolve alongside increasingly capable AI-enabled machines.

According to an IBM Institute for Business Value (IBV) study, by 2022 as many as 120 million workers in the world's 12 largest economies may need to be retrained or reskilled as a result of AI and intelligent automation[xlii]. In addition, only 41 percent of CEOs surveyed say that they have the people, skills and resources required to execute their business strategies. The time it takes to close a skills gap through training has increased by more than 10 times in just four years! In 2014, it took three days on average to close a capability gap through training in the enterprise; in 2018, it took 36 days. An aggravating factor is indeed the pace of modern technological change, which is so rapid that many workers, unable to adjust, will simply become obsolete. Therefore, the main issue today is whether the system can adapt as it did in the past. To avoid the "technology trap", policymakers should strive to manage the transition at the earliest stage possible. That implies in particular: making greater use of wage insurance, to compensate workers who have to move to jobs with a lower salary; reforming education systems to boost early-childhood education, bring coding to schools, and support retraining and lifelong learning; extending income tax credit to improve incentives to work and reduce inequality; removing regulations that hinder job-switching; providing mobility financial support to subsidise relocation as the distribution of jobs changes.

A related issue that hinders solutions to addressing the skills gap concerns the lack of gender diversity in the science, technology, engineering, and mathematics (STEM) workforce: gender stereotypes and discrimination, a lack of role models and mentors, insufficient attention to work-life balance, and "toxic" work environments in the technology industry come together to create a perfect storm against gender inclusion. There is no easy fix to boost diversity in AI roles, but one necessary option would be for educational institutions to promote the creation of better links between arts, humanities, and AI, thus changing the image of who can work in AI.

Big tech vs. Regulators: 1-0

First came the computer, then the network that allowed multiple devices in the same location to share information. From there the Internet evolved, giving people the ability to store, sort, and find information with nothing but a typed request. Virtual (or digital) assistants – application programs that understand natural language voice commands and complete tasks for the user – are likely to be the next revolution in computing. Apple has *Siri*, Google has *Google Assistant* and *Google Now*, Microsoft has *Cortana*, Amazon has *Alexa*, Alibaba Group has *AllGenie*, and all of them can provide results with just a voice command[xliii]. The technologies that power virtual assistants require massive amounts of data, which feeds artificial intelligence (AI) platforms, including machine learning, natural language processing and speech recognition platforms. As the end-user interacts with a virtual assistant, the AI programming uses sophisticated algorithms to learn from data input and become better at predicting the end user's needs.

In the past few years, big tech companies have embraced ethical self-scrutiny by establishing ethics boards, writing ethics charters, and/or sponsoring research in topics like algorithmic bias. But many people doubt these boards and charters are changing how the companies work or are holding them accountable in any meaningful way.

Most of the ethics principles that have been developed so far lack any institutional framework, hence they are non-binding. This makes it easy for companies to look at ethical issues and still continue with whatever it is they were doing beforehand.

In fact, if companies like Google, Microsoft and Facebook continue to develop and/or use AI without both internal and external regulation, auditing and close monitoring, they could sell technologies and products representing a serious risk to all citizens in the world.

Tech companies and AI Ethics			
Company	AI Ethics Principles		Controversy
	Goal	Principles	
Google	AI at Google: our principles (2018)	Be socially beneficial Avoid creating or reinforcing unfair bias Be built and tested for safety Be accountable to people Incorporate privacy design principles Uphold high standards of scientific excellence Be made available for uses that accord with these principles	Project Maven[xliv] Project Dragonfly[xlv]
Amazon[xlvi]	NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon (2019)	Designing fairness into AI systems Transparency, explainability & accountability in AI systems Factors that affect trustworthiness Ethical decision-support and decision-making systems Detecting or ameliorating, or designing to prevent, adverse biases in data and algorithms	Selling facial recognition technology to police
Facebook	AI Ethics Team (2018) Partnership with the Technical University of Munich (TUM) to	Facebook currently uses AI across its platform, e.g., to recognise people's faces in photos, to detect people who may be at risk of suicide, and to remove abusive posts Trust, privacy, fairness or inclusion, e.g., when	Cambridge Analytica data scandal

	support the creation of an independent AI ethics research centre – Development of ethical guidelines for the responsible use of the technology in society and the economy (2019)	people leave data traces on the Internet or receive certain information by way of algorithms Transparency and accountability, e.g., in medical treatment scenarios, or with rights and autonomy in human decision-making in situations of human-AI interaction	
Apple	Apple researchers are allowed to start publishing their work in academic journals (2016). (The first paper was on training image recognition algorithms, titled "Learning From Simulated and Unsupervised Images Through Adversarial Training".) Getting more app developers to use AI tools such as recognising objects in front of an iPhone's camera (2018)		Unethical and inhumane factory labour practices in China
Microsoft	Microsoft AI principles: Designing AI to be trustworthy requires creating solutions that reflect ethical principles that are deeply rooted in important and timeless values	Fairness: AI systems should treat all people fairly Inclusiveness: AI systems should empower everyone and engage people Reliability & Safety: AI systems should perform reliably and safely Transparency: AI systems should be understandable Privacy & Security: AI systems should be secure and respect privacy Accountability: AI systems should have algorithmic accountability	

The risk of an authoritarian approach of ethics and the emergence of a Global Surveillance State

The age of consumer data harvesting in real time might fuel new AI that not only exploits our attention, but also our emotions and basic drives. US police departments adopting superior Chinese software in facial recognition is an early warning of how this could occur – almost invisibly.

But it's more than likely Huawei, Alibaba, Baidu, Tencent, ByteDance and others are winning the race to technological development and implementations of AI. This is because China has a greater pool of consumer data to train machine learning, more facial recognition start-ups, the beginnings of a social credit system that rates citizens, a long practice of mass surveillance. In fact, China has already virtually won the trade-war, because it is winning the race to innovation. As the US doesn't intend to regulate the GAFAM properly and GAFAM self-regulation is largely artificial, the Western world is led to see its innovation stifled and pushed to copy China just to keep up.

In other words, if China succeeds in maintaining its current supremacy in AI research and innovation, catches up with AI skills, and comes with bolder implementations of the relevant technologies, it will probably become for a long time the dominant future superpower in the field and the Chinese "values" (notably in human rights norms) would win as the global standard for AI regulation.

Chinese ethics on AI – or the lack of ethics – here impacts the creation of a global surveillance world where data-monitoring in real-time will scale in the 2020s and 2030s. A computer system that can track and identify any face anywhere is coming to global citizens without their awareness and consent.

It is not unreasonable to believe that if China were to become the economic and technological superpower of the 21st century, then it will implement a global surveillance state. China possess the top facial recognition tech start-ups in the world and the venture capital infrastructure to scale companies that can implement the most massive data harvesting surveillance state in the world. When data harvesting is the path to supremacy, it seems China won't let a little thing like "ethics" or human rights lie in the way of its plans.

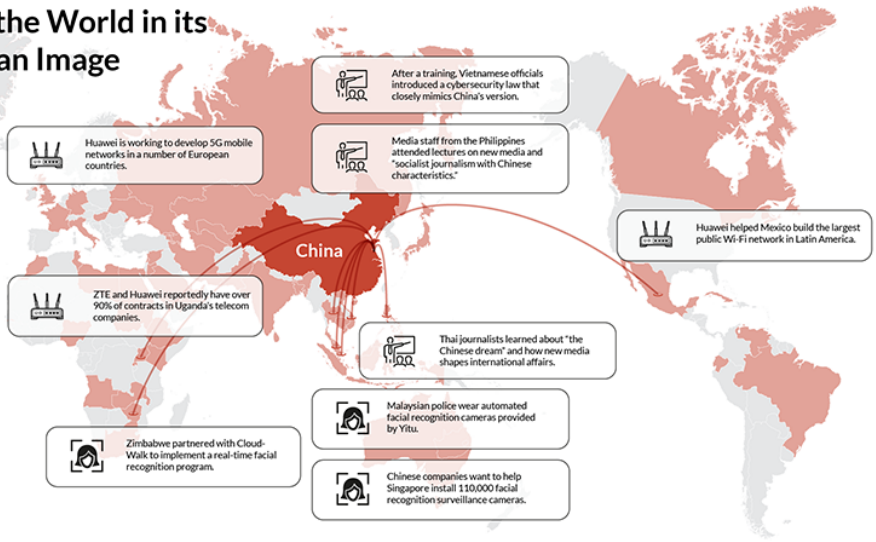
China's Belt and Road Initiative (BRI)[xlvii], with spending deemed to total $1.3 trillion by 2027, aims to revive and extend the historical Silk Road[xlviii] via networks of upgraded or new railways, ports, pipelines, power grids and highways. President Xi Jinping champions his signature project as a means to spur development, goodwill and economic integration. Yet, it is important to know that a subset of this gigantic initiative is the "Digital Silk Road", which will encompass quantum computing, nanotechnology, artificial intelligence, big data and cloud storage. China claims the Digital Silk Road will help create "a community of common destiny in cyberspace" whereby other countries will receive significant support to build digital infrastructure and develop Internet security. In fact, China is already exporting to at least 18 countries sophisticated surveillance systems capable of identifying threats to public order and has made it easier to repress free speech in 36 others[xlix].

China Remakes the World in its Techno-Dystopian Image

Telecom Infrastructure
Internet and mobile networking equipment installed by Chinese companies

AI Surveillance
Intelligent monitoring systems and facial recognition technology developed by Chinese companies

Trainings
Local media elites and government officials hosted in China for weeks-long seminars on new media or information management



www.freedomonthenet.org

Freedom House

www.freedomonthenet.org

Deep Learning is giving access to our health care data and human intimacy

There is every indication the GAFAM have machine learning to do predictive analytics on our electronic medical records (EMR). This poses a grave danger to how AI scales with our health data without our permission.

Apple announced at the 2019 Blue Button Developers Conference it will begin testing an application programming interface (API) from CARIN Alliance[l] that will allow it to integrate patient health claim data into software like its Health app. Patients do not currently have ready access to useable information about their health or their full health records. The CARIN Alliance considers patients and their caregivers should have this information and be able to send it when and where they want. Therefore, every healthcare professional will be capable of storing their patients' documents, except medical data, and at the same time patients will be granted complete access to their health information via a platform on their smartphone or an Internet website[li]. If patients consented to share their data with their physician, their health insurer or even tech companies like Apple, Microsoft and Google, those stakeholders would be entitled to use it as they wish.

A part of Google's artificial intelligence unit, Google Medical Brain, has been working on a program called *Medical Digital Assist*, which aims to harness the power of artificial intelligence for note-taking in medical examinations. Indeed, at the moment note-taking is a very time-consuming process for doctors – in many cases, documentation takes up the biggest part of the doctor's work day. Therefore, it is not surprising that Google is developing AI technology which could take care of that task instead. This would allow doctors to use more of their time on meeting patients for example. The AI-fuelled technology will listen to conversations between health professionals and patients, picking out and documenting the crucial parts of the conversation. The program will also be utilizing touch technology. The main point of this technology is to make patient records more accessible and save time while improving accuracy. This may seem innocent enough, but considering that Google has shown ethical lapses at the highest levels of their executive leadership and strategy in machine learning and voice recognition, the further development of this program and its widespread implementation might become worrisome.

Google's AI uses neural networks, which has proven effective at gathering data and then using it to learn and improve analysis. Google has possibly the fastest and more accurate tools at evaluating a patient's medical history known today. What could this mean for the future of healthcare? Google envisions the AI system steering doctors towards certain medications and diagnoses, which will fundamentally change how doctors deal with patients, hopefully helping to improve patient outcomes, reduce error, and use patient data like never before. More specifically, Google has a new algorithm that can quickly sift through thousands of digital documents in a patient's health record to find important information and, with superior number crunching, once fed this data, the AI makes predictions about the likelihood of death, discharge, and readmission. In particular, Google's Medical Brain team is training its AI to predict the death risk among hospital patients[lii].

For its part, Microsoft has launched an initiative, Healthcare NExT, which aims to accelerate healthcare innovation through AI and cloud computing. The explosion of data, incredible advances in computational biology, genomics and medical imaging have created vast amounts of data well beyond the ability of humans to comprehend.

Therefore, providing cloud- and AI-powered tools is expected to unlock the vast potential of AI and cloud computing, in particular by developing foundations for precision health care, enabling the health industry's move to the cloud, and empowering the people that make healthcare work.

The push of these tech companies into healthcare means they are getting their hands on our "life-and-death" data. This movement has already begun, with its tantalising promises and scary scenarios. We should not ignore the possibility that our most secret vulnerabilities could be hacked, thus generating grave consequences for both individuals and humankind.

More generally, the huge involvement of tech companies in AI raises questions concerning the future of human intimacy in a world of ubiquitous "AI companions", sex robots, and always more addictive and immersive technology. Tech companies, attracted by a potentially very lucrative business, are now building personal assistants that will not only be able to help us organise every aspect of our lives, but eventually give us a sense of companionship, psychological support and maybe even emotional connection.

The weaponization of Artificial Intelligence is changing the fundamentals of security

Autonomous weapons are being dubbed as the third revolution in warfare after gunpowder (circa 900 A.D. in China[liii]) and nuclear weapons (Manhattan Project, 1942). They are believed to eventually lead warfare to an algorithmic level. The second revolution brought the world to the brink of World War III in the aftermath of the Cuban Missile Crisis (October 1962). The third revolution, however, could be even more volatile and uncertain in triggering such an event.

By leading nations towards a new algorithmic warfare battlefield that has no boundaries or borders, may or may not have humans involved, AI is rapidly becoming the centre of the global power play. Autonomous weapons systems (AWS) offer greater speed, accuracy, persistence, precision, reach and coordination on the cyberspace, geospace and space (CGS) battlefields. Automated warfare has already begun in cyberspace – anyone and everyone is a target. However, security risks are numerous for not only each nation's decision makers but also for the future of humanity. First, algorithms are by no means secure nor are they immune to bugs, malware, bias, and manipulation. Second, since machine learning uses machines to train other machines, what happens if there is malware or manipulation of the training data? Third, smart connected devices increase the possibility of cybersecurity breaches everywhere, from remote locations, and because the code is opaque, security is very complex.

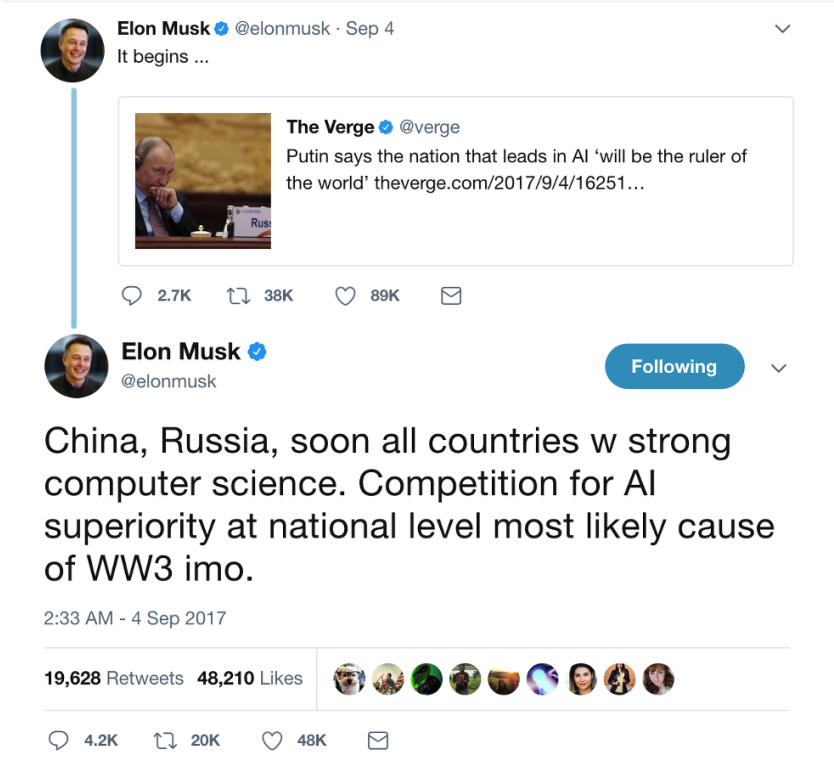
Then, after cyberspace, what is next, geo-warfare and space-warfare? And, who and what will be the targets?

"The key question for humanity today is whether to start a global AI arms race or to prevent it from starting. 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."[liv]

Unfortunately, the main nations of the world think differently. In 2017, in a 45-minute open lesson, attended via satellite links by students and teachers from 16,000 schools for a total audience of over one million, Russia's president Vladimir Putin said: "*Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world.*"

Putin's words prompted a reaction from Elon Musk, who regularly warns about the dangers posed by future AI. In a tweet, Musk cautioned that competition for "AI superiority" could result in World War 3.



Another scary issue is the following: what will happen if the dictators of the future were not human beings, but actual AI with the capability to evolve independently and aggressively and to hack any system? Who can guarantee today that in such circumstances human beings would be able to keep control by turning the hacked system "off"?

These questions, and many similar ones, cannot probably be answered today by scientists, and even less by unknowledgeable decision makers, but they should rapidly be included in all future conversations about AI Ethics.

And what about God?

An increasing number of scientists promote the idea that industrial and service robots as well as the growing field of autonomous systems spanning from drones and driverless vehicles to cognitive vision and computing will eventually have a conscience, or even a soul.

For example, Martine Rothblatt, a lawyer, technologist, and medical ethicist, coined in a book the concept of *cyberconsciousness*: "(…) when a robot is created using the memories and knowledge from a human mind the result is new, spontaneous, and original combinations of those ideas, which in turn leads to original "equations" or thoughts. We recognize this behaviour as acting or "being" human, and information technology (IT) is increasingly capable of replicating and creating its highest levels: emotions and insight. This is called *cyberconsciousness* (…) Running right alongside (…) is the development of powerful yet accessible software, called *mindware*, that will activate a digital file of your thoughts, memories, feelings, and opinions – a *mindfile* – and operate on a technology powered twin, or *mindclone* (…) It's only a matter of time before brains made entirely of computer software express the complexities of the human psyche, sentience, and soul (…) The eventual sophistication and ubiquity of mindclones will naturally raise societal, philosophical, political, religious, and economic issues. Cyberconsciousness will be followed by new approaches to civilization that will be as revolutionary as were ideas about personal liberty, democracy, and commerce at the time of their births (…) if we don't treat cyberconscious mindclones like the living counterparts they will be, they will become very, very angry."[lv]

Indeed, the issue of AI consciousness should be regarded as an open question: if we look at the chemical differences between carbon and silicon, and their impact on life itself, we should not exclude the possibility that these differences impact whether silicon gives rise to consciousness beyond its ability to process information in a superior manner. And if this were happening, we should not exclude that future AI systems, unaware of human beings' consciousness, might ask whether biological, carbon-based beings have the right substrate for experience. In the end, from an ethical perspective, human beings should debate whether consciousness could merely be outmoded by superintelligent, silicon-based AI systems surpassing expert knowledge in every domain and ignoring the very mental faculties associated with conscious experience in humans, or whether, if a silicon-based consciousness eventually existed, it could prove superior to carbon-based consciousness and be better suited for the development of life throughout the universe[lvi]. Conscious AI systems would also raise troubling legal and ethical issues – would a conscious robot be a "person" under law and hence be liable if its actions hurt someone? More frighteningly, might conscious robots rebel against humans and opt to wipe humans out of the earth's surface altogether?

Moving further, in June 2018 Dan Robitzski, a neuroscientist-turned-journalist, wrote a fascinating article titled "Artificial Consciousness: How to Give a Robot a Soul?"[lvii] One year later, at an internal seminar on ethics organised by SNCF's Ethics and Deontology Directorate, a speaker from industry titled his presentation: "Do Algorithms have a Soul?"[lviii] Talking about AI by conjuring the notion of "soul" may look fanciful. But let us leave aside the scientific and philosophical debate concerning the differences between "soul", "mind", "awareness" and "conscience[lix], what is most

striking here is the civilizational shift that the rapprochement between “AI” and “soul” suggests.

Indeed, it is clear that the development of Artificial Intelligence will continue to inspire billion-dollar companies, far-reaching research programs, and more or less fictive scenarios of both transcendence and doom. In 2017, Anthony Levandowski, the co-founder of Otto, Waymo, and known as the autonomous vehicle pioneer formerly of Google and Uber, started his own IRS-approved religion, *Way of the Future* (WOTF), dedicated to the worship of Artificial Intelligence. So, Artificial Intelligence has already originated its first church whose activities will focus on the realisation, acceptance, and worship of a Godhead based on AI developed through computer hardware and software. That includes funding research to help create the divine AI itself. The religion will seek to build working relationships with AI industry leaders and create a membership through community outreach, initially targeting AI professionals and other people who are interested in the idea. “What is going to be created will effectively be a god,” Levandowski said in an interview[lx]. “It’s not a god in the sense that it makes lightning or causes hurricanes. But if there is something a billion times smarter than the smartest human, what else are you going to call it?” The creation of the WOTF church marks the evolution of the techno-religious sentiment from a marginal movement to an institutionalised belief system, which is an undeniably large and significant leap. That goes a long way – “There are many ways people think of God, and thousands of flavours of Christianity, Judaism, Islam,” says Levandowski, but they’re always looking at something that’s not measurable or you can’t really see or control. This time it’s different. This time you will be able to talk to God, literally, and know that it’s listening.”

Levandowski is not alone. In his bestselling book, Homo Deus, Yuval Noah Harari argues that the foundations of modern civilization are eroding in the face of an emergent religion, which he calls “dataism” – by giving ourselves over to information flows, we can transcend our earthly concerns and ties[lxi]. Other nascent transhumanist religious movements go even further by focusing on immortality.

The “AI as God” wave is reaching literature. In his last novel, *Transparence*, which takes place in the 2060s when life on Earth is seriously threatened by global warming, Marc Dugain features a small digital company based in Iceland that is about to launch a secret, revolutionary program on immortality, named Endless, consisting in transplanting the human soul into an artificial earthly body[lxii].

It may seem weird, if not scandalous, to many people that the very idea of God be replaced with a scientific discipline which, by harvesting exabytes of data, would be capable of imposing a new ‘religion’ on Earth. Yet, this is what is happening at the same time when religions are revisited, sometimes extravagantly[lxiii].

Artificial Intelligence Ethics – An oxymoron?

"We need a Hippocratic oath in the same way it exists for medicine. In medicine, you learn about ethics from day one. In mathematics, it's a bolt-on at best. It has to be there from day one and at the forefront of your mind in every step you take."

Hannah Fry, associate professor in the mathematics of cities at University College London, in The Guardian, 16/08/2019

Why Ethics is important for AI?

As AI is changing societies and economies around the world, its ethical dimension is of growing importance. Recent advances in AI-enabled technologies have prompted a wave of responses across the globe, as nations attempt to tackle emerging ethical issues. Germany has delved into the ethics of automated vehicles, rolling out the most comprehensive government-led ethical guidance on their development available. New York has put in place an automated decisions task force, to review key systems used by government agencies for accountability and fairness. The UK has a number of government advisory bodies, notably the Centre for Data Ethics and Innovation. The European Union has explicitly highlighted ethical AI development as a source of competitive advantage.

Clearly, we have entered an era when we hand over our most private data, our DNA, but we're not just consenting for ourselves, we do it also for the next generations. Well, let's assume we don't live today in a world where people are genetically discriminated against; but who's to say in one-hundred years that we won't?

A survey conducted among four groups of experts in 2012-2013 by AI researchers Vincent C. Müller and Nick Bostrom reported a 50 percent chance that Artificial General Intelligence (AGI) would be developed between 2040 and 2050, rising to 90 percent by 2075; so-called "superintelligence", which Bostrom defines as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" – was expected some 30 years after the achievement of AGI[lxiv].

Futurists like Ray Kurzweil predict an intelligence explosion that will lead to the "singularity" – a moment when computers, advancing their own intelligence in an accelerating cycle of improvements, far surpass all human intelligence. What that moment will look like for humanity, and what we can do to ensure artificial superintelligence benefits rather than harms us, remain for the time being open questions. Indeed, the notion of "Superintelligent Artificial Intelligence" (SAI) makes us anxious as we fear that AI will threaten humanity. The idea of "singularity" has sparked existential debates, conspiracy theories, and of course has raised ethical concerns. If it happens, since AI will have already surpassed the cognitive ability of humans, surely, we will not be able to resist. Bostrom argues that we would have to tactfully negotiate the terms of our existence. As we have mentioned above, some people are even willing to worship the AI overlords and have already established the "Church of Artificial Intelligence" in the Silicon Valley.

We must admit that the advent of SAI is not unlikely. AI machines will then be globally connected 24 hours a day, and therefore they will develop an increasing comparative advantage on us, humans, who are still obliged to go to sleep for a minimum number of hours each day.

Even with the current fast development of AGI, concerns are raised concerning the ability of humans to eventually control their creatures. Since 2018 voices have been heard more loudly to claim that mathematicians, computer engineers and scientists in related fields should commit to an ethical pledge – equivalent to the Hippocratic oath – to think deeply about the possible applications of their work and pursue only those that, at the least, "do no harm" to society. to protect the public from new technologies under development in laboratories and technology firms. The case for a Hippocratic oath for scientists is not new. Already in 1969 epistemologist Karl Popper had written: "One of the few things we can do is to try to keep alive, in all scientists, the consciousness of their responsibility."

Some questions:

Will there come a point when AI computers become too intelligent?

Should robots have the same legal rights as humans?

How would we get around the question of ownership?

To which extent should we trust (or be able to predict the behaviour of) robots?

What responsibilities do we owe such robots?

Would it be wrong to deliberately install mechanisms that gave humans an immediate physical advantage, e.g., an "off" switch?

Should we allow a self-governing robot society to co-exist with ours?

Ethics and Values

While "ethics" and "values" are sometimes used synonymously, they are different, wherein ethics are the set of rules that govern the behaviour of a person, established by a group or culture, and values refer to the beliefs for which a person has an enduring preference. Ethics and values are important in every aspect of life, when we have to make a choice between two things, wherein ethics determine what is right, values determine what is important. The problem with ethics is that it is a shifting, amorphous concept that can rapidly change among different cultures, societies, and values. Although there have been numerous attempts at drafting ethical AI guidelines, what remains unclear is if everyone, regardless of sector, socioeconomic status, culture, or religion, agrees. What's "ethical" for western societies may not be so to an Eastern or Asian one. Even within western societies, a country like Germany is over-sensitive to the issues of privacy and ethics, for historical reasons, while other European countries show a lower sensitivity and the United States are more partial to security than to privacy. Privacy from face recognition may be tantamount to western societies that view freedom as a human right, but not matter as much for Chinese citizens who are accustomed to surveillance – or even welcome it if it means they'll benefit from an AI-powered social credit system.

As professor Sarah Spiekermann wrote, "if IT managers and engineers focused on creating IT values throughout the entire design process while rigorously controlling for value risks, they would support human beings in their flourishing much more than they do today. Machines would then be designed to strengthen people's values such as their health, increase their sense of privacy, freedom and autonomy, help them trust, and so

forth. In the long run, we would even envision that machines support the development of cognitive skills such as learning, help them rediscover their senses, have more ethical integrity, be more just in their decisions, and so on."[lxv]

The difficulty of dealing with ethics is that it is a "value", i.e. "a conception, explicit or implicit, distinctive of an individual or characteristic of a group, of the desirable which influences the selection from available modes, means and ends of action (…) A value is not just a preference but is a preference which is felt and/or considered to be justified – "morally" or by reasoning or by aesthetic judgements, usually by two or all three of these."[lxvi] Therefore, the threshold level of how strongly something is valued depends on the culture of a group or a society at a specific time. This is why we noted differences between, for example, western societies and China, in the importance, and even the relevance, of ethics. Before western societies became secular, or laic, "moral" values like charity, humility and obedience were prevailing; in today's capitalist societies, including China, economic success informs the dominant ideology and almost opposite values like competition, pride and autonomy are favoured. The importance assigned to values fluctuates over the course of history and depends on the ideals of a society.

As we move into the Artificial Intelligence age, where machines of all kinds are connected and made *smart*, we need to trigger and feed a conversation on our current ideals, as much as possible at global level. Our ideals, in one country, one region, and the world, will influence how we regard values such as privacy, security, freedom, control, and how much importance we grant to them. Some values are viewed so important over time by some societies that they are transformed into rights, e.g., human freedom and dignity. Other values transcend individual country legal systems and become international conventions, e.g., the rights and freedoms enshrined in the Universal Declaration of Human Rights (1948) and its first binding instrument – the European Convention of Human Rights (1950).

Value theory distinguishes between *intrinsic* values (i.e. something that is valuable in itself or in its own right) and *extrinsic/instrumental* values (i.e. something that relates and enables something else that is good). If there can be many extrinsic values, scholars often disagree on how many intrinsic values there are. Some believe that there is only one ultimate value – *human happiness*. In the preamble to the Declaration of Independence it is written: "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness." But most philosophers stress other intrinsic values besides happiness such as knowledge, beauty, health, truth, power, and harmony.

Therefore, designing an ethical framework for digital technology, in particular for Artificial Intelligence, consists in embracing a desire to create positive value through technology, i.e. how technology can benefit society while addressing its risks along the way. This is the philosophy that transpires in any AI ethical guidelines that have been developed recently across the world, with the core principle being the "AI Code" (UK) or "generates net-benefits" (most countries).

As Spiekermann shows (Figure below), ethics-by-design, i.e. value-based IT design, is not a mere philosophical exercise but an economic and political necessity. From the perspective of economics, one could argue that ethics throughout IT has positive externalities for society and companies, whereas its absence has negative externalities.

	1. Protect from risk to destroy value through IT		2. Foster value creation through IT
Individual	Health problems, loss of duty/function, information overload, biased treatment, loss of autonomy, loss of privacy, loss of control	⇒	Better health, better duties/functions, more calmness, more courtesy, more autonomy, more privacy, more control
	⇩		⇩
Society	High sickness rates, high unemployment rates, lowered productivity, ubiquitous distrust, loss of citizen autonomy, loss of citizen privacy, loss of accountability perceptions	⇒	Healthier citizens, social stability, higher productivity and knowledge creation, atmosphere of trust, higher citizen autonomy, more creativity and innovation, more responsibility taking by citizens

Negative externalities Positive externalities

The Difficulty to Address AI Ethics

Ethical issues

In the above sections we have frequently touched upon the main ethical "concerns" that are raised in various countries. Although they are well known, it remains appropriate to summarise them.

<u>Identification and profiling of people without their consent</u>

Through the posts and likes that people give on social networks, AI can define their psychological profile, detect their political opinions, sexual orientation, and so on. This may lead to the Cambridge Analytica scandal[lxvii]. Such profiling is now the object of trade. For example, the Spinner* is a service that enables to subconsciously influence a specific person, by controlling the content on the websites he or she usually visits[lxviii].

The EU General Data Protection Regulation (GDPR) regulates what companies can do with European people data and gives more control to European citizens about how their data is collected and used through the request of consent. Can we really, however, speak of consent when we are forced to accept voluminous general conditions of use and cookies in order to be capable to go on websites? The notions of voluntary and well-informed consent and of "defects of consent" must be questioned and clarified.

Awareness

People may not be aware of interacting with AI. In the case of chatbots, for example, it may be difficult to be sure for a human whether she/he is interacting with a machine. Awareness may be crucial for some services such as web appointments with doctors or psychologists.

Freedom

AI may score people or organizations without their consent and their awareness of all impacts. This violates the principle of freedom. For example, as we have already mentioned above, with millions of cameras[lxix] and billions of lines of code, China is building a high-tech authoritarian future. The Chinese authorities are embracing technologies like facial recognition and artificial intelligence to identify and track 1.4 billion people. they want to assemble a vast and unprecedented national surveillance system, with crucial help from its thriving technology industry. China is reversing the commonly held vision of technology as a great democratizer, bringing people more freedom and connecting them to the world – in China, it has actually brought control. Citizens are scored according to their behaviour in their everyday life and the access to some services may depend on the score they have (access to university etc). Data from facial recognition is also used to detect ethnicities and label them (e.g., Han Chinese, the main ethnic group of China, or Uyghur Muslims). The goal clearly is "algorithmic governance"!

Bias

Artificial intelligence has a bias problem. Concerns regarding racial or gender bias in AI have arisen in applications as varied as hiring, policing, judicial sentencing, and financial services. As of today, algorithms can reproduce and accentuate bias according to the data they use. There are two main challenges that AI developers, users, and policymakers need to address:

- Bias built into data. For example, for sentencing decisions it would be improper to use race as one of the inputs to an algorithm. But what about a seemingly race-neutral input such as the number of prior arrests? In the United States at least, arrests are not race neutral: evidence indicates that African-Americans are disproportionally targeted in policing, and hence arrest record statistics are heavily shaped by race. In the United States, a proprietary risk assessment algorithm named COMPAS (*Correctional Offender Profiling for Alternative Sanctions*), widely used to decide on the freedom or incarceration of defendants passing through the US criminal justice system, might be systematically biased against specific populations, in particular African Americans compared to whites[lxx]. The indirect influence of bias is present in plenty of other types of data; for example: evaluations of creditworthiness are determined by factors including employment history and prior access to credit (two areas in which race has a major impact); starting salaries for new hires in a large company may be set by AI through inputs from salary history, which itself is frequently influenced by a gender bias.
- AI-induced bias. Biases can be created within AI systems and then become amplified as the algorithms evolve. AI algorithms are not static – rather they learn and change over time. Initially, an algorithm might make decisions using

> only a relatively simple set of calculations based on a small number of data sources. As the system gains experience, it can broaden the amount and variety of data it uses as input, and subject those data to increasingly sophisticated processing. Therefore, an algorithm can end up being much more complex than when it was initially deployed, and this can happen without human intervention to modify the code, but rather because of automatic modifications made by the machine to its own behaviour. In some cases, this evolution can introduce bias[lxxi].

While AI has the potential to bring enormous benefits, the challenges discussed above need close attention. These challenges are certainly not a reason to stop investing in AI or to burden AI creators with innovation-stifling new regulations, but it is important to put real effort into approaches that can minimize the probability that bias will be introduced into AI algorithms, either through externally-supplied data or from within.

Does Artificial Intelligence threaten the future of humans?

As emerging algorithm-driven AI continues to spread, will people be better off than they are today?

Experts predict networked artificial intelligence will enhance human performance[lxxii] – computers might match or even exceed human intelligence and capabilities on tasks such as complex decision-making, reasoning and learning, sophisticated analytics and pattern recognition, visual acuity, speech recognition and language translation – but also threaten human autonomy, agency and capabilities.

As artificial intelligence systems get better at manipulating us, in particular because they can decipher almost any emotion from our face or speech, the risk is high that we end up happily submitting our lives to the algorithms in the same way as we already do today with our smartphones (people touch their phones on average more than 2,500 times a day). At a certain point in the future, we will be tempted to believe that we can take it easy knowing that the hard work of planning, exploring, managing, organizing and evaluating is being done by a better brain – the robot or the autonomous system. Why should we care any longer since AI is “superintelligent” and can perform better than us? Why should we continue to tackle new complex challenges? Then, rather than concentrating on the most creative tasks while AI would perform the more repetitive ones, we will begin to become lazy. And deprived of mental effort, our brains will begin to regress.

Concerns and suggested solutions in the conflict between AI and humans		
Concerns	**Human agency**: Individuals are experiencing a loss of control over their lives	Decision-making on key aspects of digital life is automatically ceded to code-driven, "black box" tools. People lack input and do not learn the context about how the tools work. They sacrifice independence, privacy and power over choice; they have no control over these processes. This effect will deepen as automated systems become more prevalent and complex.
	Data abuse: Data use and surveillance in complex systems is designed for profit or for exercising power	Most AI tools are and will be in the hands of companies striving for profits or governments striving for power. Values and ethics are often not baked into the digital systems making people's decisions for them. These systems are globally networked and not easy to regulate or rein in.
	Job loss: The AI takeover of jobs will widen economic divides, leading to social upheaval	The efficiencies and other economic advantages of code-based machine intelligence will continue to disrupt all aspects of human work. While some expect new jobs will emerge, others worry about massive job losses, widening economic divides and social upheavals, including populist uprisings.
	Dependence lock-in: Reduction of individuals' cognitive, social and survival skills	Many see AI as augmenting human capacities but some predict the opposite - that people's deepening dependence on machine-driven networks will erode their abilities to think for themselves, take action independent of automated systems and interact effectively with others.
	Mayhem: Autonomous weapons, cybercrime and weaponized information	Some predict further erosion of traditional socio-political structures and the possibility of great loss of lives due to accelerated growth of autonomous military applications and the use of weaponized information, lies and propaganda to dangerously destabilize human groups. Some also fear cybercriminals' reach into economic systems.
Suggested solutions	**Global good is number 1**: Improve human collaboration across borders and stakeholder groups	Digital cooperation to serve humanity's best interests is the top priority. Ways must be found for people around the world to come to common understandings and agreements – to join forces to facilitate the innovation of widely accepted approaches aimed at tackling wicked problems and maintaining control over complex human-digital networks.
	Values-based system: Develop policies to assure AI will be directed at "humanness" and common good	Adopt a "moon-shot mentality" to build inclusive, decentralized intelligent digital networks "imbued with empathy" that help humans aggressively ensure that technology meets social and ethical responsibilities. Some new level of regulatory and certification process will be necessary.
	Prioritize people: Alter economic and political systems to better help humans "race with the robots"	Reorganize economic and political systems toward the goal of expanding humans' capacities and capabilities in order to heighten human/AI collaboration and staunch trends that would compromise human relevance in the face of programmed intelligence.

Source: Pew Research Center, "Artificial Intelligence and the Future of Humans", 10 December 2018

In addition, AI systems may disqualify human beings in several ways. First, some people consider AI provides absolute truth, and hence they rely on its recommendations or decisions more than on human judgment. This can lead people to "intellectual laziness", to losing their critical sense, and to pushing them out of responsibility by relying on the machine. Second, AI may disqualify the performance and skills of professionals, e.g., robots replacing manual trades or AI algorithms bypassing intellectual services such as those provided by lawyers or accountants. Third, AI may disqualify human relationships – robotic companions who never contradict their owner and seem empathetic may actually make human relations less attractive as more complicated to manage.

"Black box"

Neural networks are a particular concern not only because they are a key component of many AI applications – including image recognition, speech recognition, natural language understanding and machine translation – but also because they're something of a "black box" when it comes to elucidating exactly how their results are generated. Neural networks are so-called because they mimic, to a degree, the way the human brain is structured: they're built from layers of interconnected, neuron-like, nodes and comprise an input layer, an output layer and a variable number of intermediate hidden layers. The nodes themselves carry out relatively simple mathematical operations, but between them, after training ("back propagation"), they can process previously unseen data and generate correct results based on what was learned from the training data.

Therefore, relying on AI systems can be very sensitive in some cases. For example, are the owners of autonomous cars ready for an artificial intelligence system to decide in their place who will live and who will die in the event of a lethal accident? Are people ready to lose the control in such circumstances? So, it's not surprising that most experts consider putting the AI in a mode where it's not supervised may bear risks – the human must have the final call, especially for critical applications. That's why terms such as transparency, explainability and interpretability are playing an increasing role in the AI ethics debate.

The European Commission's General Data Protection Regulation (GDPR) states in its Article 22.1 "*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*" But what does the GDPR say precisely about machine learning and artificial intelligence? Well, not much actually. There is a continuing debate that centres on the single occurrence of the phrase "right to explanation" that occurs in Recital 71, a companion document to the GDPR that is not itself legally enforceable[lxxiii]. However, the GDPR states that data controllers must notify consumers how their data will be used, including "*the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.*" (Article 13.1(f)) Common-sense reading would suggest that if a computer is making real-world decisions without a human in the loop, then there should be some accountability for how those decisions are made. For example, if a bank's machine learning model denies you credit, and does so without meaningful human intervention, then the bank owes you an explanation of how it arrived at that decision. As companies are selling what are essentially black box systems, when things go wrong the

programmers are hiding behind a lack of transparency, saying "nobody can understand these systems." Fairly, this is not an acceptable answer – it's an expression of human dignity, of human autonomy, not to have decisions made about us which have no reviewability. However, even if the two principles of transparency and accountability for how data is processed play an essential role in the GDPR, and hence may infer a "right to explanation", the technical difficulty of implementing such a right for increasingly complex AI systems raises the question whether compliance with it would be possible at all – explaining decisions might result in disclosure of closely-held trade secrets or otherwise violate their intellectual property rights. By remaining deliberately vague on regulating AI, the European Commission has sought to balance the interests of data controllers and those of data subjects. The spirit of the regulation seems to be not to hinder innovation while maintaining enough "regulatory hooks" to intervene, if necessary. In the end, we can say that finding the right balance between accountability and encouraging innovation remains an unsolved problem.

For their parts, the French government already has committed to publishing the code that powers the algorithms it uses, and in the United States the Federal Trade Commission's Office of Technology Research and Investigation has been charged with providing guidance on algorithmic transparency.

Ethical principles

The table below presents a comparison of the main ethical principles for a limited number of countries and regions across the world. The titles of these principles are those mentioned in the national/regional strategy documents, but they have been sometimes streamlined or clustered to facilitate understanding or to avoid overlaps. For example: some documents talk of "no harm", while others use the term "non-maleficence"; several documents mention "privacy" alone, while the EU ethics guidelines for trustworthy AI add the phrase "data governance"; the same EU ethics guidelines include the expression "diversity, non-discrimination and fairness", while others also speak of "intelligibility".

AI Ethical Principles in different Countries / Regions

[lxxiv]	Australia	EU	India	OECD	UK	USA
Accountability (incl. by-design)	X	X		X		X
"AI Code" (cross-sector code of conduct)					X	
Contestability	X					
Avoiding harm / Nonmaleficence	X		X		X	
Flourishing alongside AI					X	
Generates net-benefits	X			X	X	
Human agency and oversight		X	X			
Human dignity			X			
Intelligibility, diversity, non-discrimination and fairness	X	X			X	
Privacy and Data governance	X	X			X	
Regulatory and legal compliance	X	X		X		X
Societal and environmental well-being		X		X		
Technical robustness and safety		X	X	X		X
Transparency and explainability	X	X	X	X		

Two main conclusions come out from this geographical comparison. One, there's a strong global convergence towards six ethical principles, including accountability,

transparency, justice and fairness, non-maleficence, responsibility, and privacy. Two, people can't really agree on what any of those words mean when it comes to policy[lxxv].

Accountability is mentioned in almost every document, though the definition of this concept slightly differs between countries. The EU ethics guidelines say: "The requirement of accountability… is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use." In the United States, documents focus on the concept of accountability-by-design.

Transparency, or the ability to understand the decisions of AI is the most prevalent principle in the current literature. Some guidelines seek to increase interpretability or other acts of communication, with the main goal of reducing harm. Other guidelines highlight transparency as a way to foster trust, for legal reasons, or to bolster open dialogue and the principles of democracy. Some point to opening the source code, where others believe communicating topics such as evidence for AI and its limitations or responsibility for AI and investments are more impactful.

Non-maleficence – or "do not harm" – calls for safety and security, in a way that AI will never cause foreseeable or unintentional harm, including discrimination, violation of privacy, or bodily harm. The principle intersects with the one of *justice,* which monitors AI to prevent or reduce bias. A related concept is *fairness* – reduced bias for race or gender, for example, and equal access to the technology – as well as reducing harm from AI taking over jobs. Yet as before, guidelines differ on who becomes the overseer of justice. One proposed solution is technological standards, for example, training HR algorithms on datasets that span different races. Other ideas include raising public awareness, auditing, or establishing new laws. Some even propose taking explicit position against military applications.

The final two principles*, responsibility* and *privacy*, are also differently addressed. Privacy is generally seen as a value to uphold and as a right to be protected; guidelines, however, vastly differ on how to get there, though new or enforced privacy laws are a popular idea. Responsibility goes hand in hand with transparency and trust, though neither term is usually defined.

The complexity of defining the concept of "ethics" and the wide variety of its interpretations across the world make it unlikely that a global consensus on how to implement it in the design of new systems will be reached in the near future. What we observe today about AI ethics is a growing concern in several countries (e.g., Australia, EU, India), a lack of interest among others (e.g., China, Japan), and where ethics is being studied a striking difference in how to tackle the challenge (e.g., designing hard or soft regulation within the EU, finding an appropriate balance between ethics and economic interests in the US). The differences between regions of the world are rooted in their historical relationships to creativity and innovation – they are therefore *cultural* and stable over time. The risk is to have a variety of ethical guidelines with different principles, or interpretations of these principles, and different ways to design, implement and enforce them. If nations in the world were proving incapable to agree on a common set of ethical guidelines, and the way to use them, ethics in AI will remain just pious lip-service, and the further development and deployment of AI will continue to be a strategic, technological and economic battlefield between countries, primarily

the United States, China and the European Union. Instead of being an opportunity for the world to use new technology to shape a better future, the phrase "AI ethics" will remain an oxymoron.

Examples of Existing Ethical Guidelines

In the European Union

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft on 18 December 2018 on which more than 500 comments were received through an open consultation.

According to the Guidelines, trustworthy AI should be:

- lawful: respecting all applicable laws and regulations
- ethical: respecting ethical principles and values
- robust: both from a technical perspective while taking into account its social environment

The Guidelines put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy.

Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.

Technical Robustness and safety: AI systems need to be resilient and secure. They need to be safe, ensuring a fall-back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.

Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

Diversity, non-discrimination and fairness: unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

Societal and environmental well-being: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment,

including other living beings, and their social and societal impact should be carefully considered.

Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.

In France

On 15 December 2017, the French *Commission Nationale Informatique et Libertés* (CNIL) published a report summarising conclusions of the public debate which it conducted on the ethical challenges of algorithms and artificial intelligence[lxxvi].

CNIL retains in its report two "founding principles": loyalty and vigilance.

The principle of loyalty applies to all algorithms and integrates all their collective impacts, not just those concerning the individual. Every algorithm, whether it processes personal data or not, shall be loyal to its users, not only as consumers but also as citizens, and also, where appropriate, to communities or large collective interests of which the existence might be directly affected. The user interest must prevail.

The principle of vigilance/reflexivity implies to organise a kind of regular, methodical and deliberative questioning with respect to these moving objects. It is a direct response to the requirements that are dictated by these technological objects as a result of their unpredictable nature (inherent to machine learning), the very compartmentalized aspect of the algorithmic chains within which they are inserted and, finally, the excessive confidence to which they give rise. All the links in the algorithmic chain (designers, companies, citizens) must be mobilized to give form to this principle by means of concrete procedures.

CNIL makes six operational recommendations:

1. to train into ethics all the actors-links of the "algorithmic chain" (designers, professionals, citizens) – digital literacy is needed to allow every human to acquire a much better understanding of the AI machine;
2. to make algorithmic systems understandable by strengthening existing rights and organizing mediation with users;
3. to work on the design of algorithmic systems in the service of human freedom, in order to combat the "black box" effect;
4. to establish a national algorithm audit platform;
5. to foster research into AI ethics and launch a great national participative cause around a research project of general interest;
6. to strengthen the function of ethics in corporations (e.g., creation of ethical committees, dissemination of good sectorial practices, revision of deontology charters).

In Quebec

On 4 December 2018, Université de Montréal, in collaboration with the Fonds de recherche du Québec, unveiled the Montréal Declaration for Responsible Development of Artificial Intelligence[lxxvii]. This set of ethical guidelines for the development of artificial intelligence was the culmination of more than a year of work, research and

consultations with citizens, experts, public policymakers and industry stakeholders, civil society organizations, and professional orders.

A key objective of the Declaration is to identify the ethical principles and values applicable to the fields of digital technology and AI that promote the fundamental interests of people and groups. The 10 principles are as follows:

1. **Well-being**: The development and use of artificial-intelligence systems (AIS) must permit the growth of the well-being of all sentient beings.
2. **Respect for autonomy**: AIS must be developed and used with respect for people's autonomy, and with the goal of increasing people's control over their lives and their surroundings.
3. **Protection of privacy and intimacy**: Privacy and intimacy must be protected from intrusion by AIS and by data-acquisition and archiving systems.
4. **Solidarity**: The development of AIS must be compatible with maintaining the bonds of solidarity among people and generations.
5. **Democratic participation**: AIS must meet intelligibility, justifiability and accessibility criteria, and must be subjected to democratic scrutiny, debate and control.
6. **Equity**: The development and use of AIS must contribute to the creation of a just and equitable society.
7. **Diversity inclusion**: The development and use of AIS must be compatible with maintaining social and cultural diversity, and must not restrict the scope of lifestyle choices and personal experience.
8. **Prudence**: Every person involved in AIS development must exercise caution by anticipating, as far as possible, the potential adverse consequences of AIS use, and by taking appropriate measures to avoid them.
9. **Responsibility**: The development and use of AIS must not contribute to diminishing the responsibility of human beings when decisions must be made.
10. **Sustainable development**: The development and use of AIS must be carried out so as to ensure strong environmental sustainability of the planet.

Based on these principles, 8 recommendations have been developed, for the purpose of suggesting guidelines for accomplishing the digital transition within the ethical framework of the Declaration.

1. **Organization for independent citizen scrutiny and consultation**: An organization dedicated to the examination of and research into the uses and social impacts of digital technology and AI should be established.
2. **SIA audit and certification policy**: A coherent policy for the auditing and certification of AIS that promotes responsible deployment should be instituted.
3. **Empowerment and automation**: There should be support for empowerment of citizens in the face of digital technologies, in the form of access to education that enables understanding, critical thinking, respect, and accountability, so as to promote active participation in a sustainable digital society.
4. **Education and ethics**: The education of stakeholders concerned by the design, development and use of AIS should be rethought, with investment in multidisciplinarity and ethics.
5. **Inclusive development of AI**: A coherent strategy should be implemented, utilizing the various existing institutional resources, to promote inclusive

development of AI and prevent potential biases and discrimination related to development and deployment of AIS.

6. **Protection of democracy**: To safeguard democracy against the manipulation of information for political ends, a containment strategy is required to prevent deception and political manipulation of citizens via malicious social platforms and websites, along with a strategy to combat political profiling, so as to maintain the conditions for healthily functioning democratic institutions and informed citizens.
7. **International development of AI**: A non-predatory model of international development should be adopted that aims to include the various regions of the globe without abusing low- and middle-income countries (LMICs).
8. **Environmental footprint**: A public/private strategy should be implemented to ensure that development and deployment of AIS and other digital technologies are compatible with robust environmental sustainability and conducive to advancement of solutions to the environmental crisis.

In Dubai

"Smart Dubai" vision[lxxviii] is to excel in the development and use of AI in ways that both boost innovation and deliver human benefit and happiness.

The four "AI Principles" are the following:

- **Ethics**: AI should be fair, transparent, accountable and understandable;
- Security: AI should be safe, secure, and should serve and protect humanity;
- Humanity: AI should be beneficial to humans and aligned with human values in both the long and short term;
- Inclusiveness: AI should benefit all people in society, be governed globally and respect dignity and people rights.

The AI Ethical Guidelines expand on Dubai's AI Principle about Ethics dealing with fairness, transparency, accountability and explainability:

- **Fair**: Demographic fairness, fairness in design, fairness in data, fairness in algorithms, fairness in outcomes
- **Transparent**: Identifiable by humans, traceability of cause of harm, auditability by public
- **Accountable**: Apportionment of accountabilities, accountable measures for mitigating risks, appeals procedures and contingency plans
- **Explainable**: Process explainability, outcomes explainability, explainability in non-technical terms, channels of explanation

In China

The Chinese Ministry of Science and Technology and the Beijing City Government commissioned the researchers with publishing a "code of ethics" to ensure that "human privacy, dignity, freedom, autonomy and rights are sufficiently respected" in the development of AI technologies[lxxix]. Beijing University, Tsinghua University, the Institute of Automation and the Institute of Computing Technology within the Chinese Academy of Sciences, as well as the country's three major Internet companies, Baidu, Alibaba and Tencent, were involved in drafting the code.

Despite its basic accordance with western values, the code of ethics constitutes a set of rules that can be interpreted in quite different ways because it is formulated in very general terms. The Chinese government in any case does not seem to see any contradiction with its planned social credit system, which rewards “good behaviour” by adding social points and punishes “undesirable behaviour” by deducting points. Even a majority of Chinese citizens considers the necessary interference in their privacy and autonomy as acceptable.

The principles are proposed as “an initiative for the research, development, use, governance and long-term planning of AI, calling for its healthy development to support the construction of a human community with a shared future, and the realization of beneficial AI for humankind and nature.”

Research and development

The research and development (R&D) of AI should observe the following principles:

- Do Good: AI should be designed and developed to promote the progress of society and human civilization, to promote the sustainable development of nature and society, to benefit all humankind and the environment, and to enhance the well-being of society and ecology.
- For Humanity: The R&D of AI should serve humanity and conform to human values as well as the overall interests of humankind. Human privacy, dignity, freedom, autonomy, and rights should be sufficiently respected. AI should not be used to against, utilize or harm human beings.
- Be Responsible: Researchers and developers of AI should have sufficient considerations for the potential ethical, legal, and social impacts and risks brought in by their products and take concrete actions to reduce and avoid them.
- Control Risks: Continuous efforts should be made to improve the maturity, robustness, reliability, and controllability of AI systems, so as to ensure the security for the data, the safety and security for the AI system itself, and the safety for the external environment where the AI system deploys.
- Be Ethical: AI R&D should take ethical design approaches to make the system trustworthy. This may include, but not limited to: making the system as fair as possible, reducing possible discrimination and biases, improving its transparency, explainability, and predictability, and making the system more traceable, auditable and accountable.
- Be Diverse and Inclusive: The development of AI should reflect diversity and inclusiveness, and be designed to benefit as many people as possible, especially those who would otherwise be easily neglected or underrepresented in AI applications.
- Open and Share: It is encouraged to establish AI open platforms to avoid data/platform monopolies, to share the benefits of AI development to the greatest extent, and to promote equal development opportunities for different regions and industries.

Use

The use of AI should respect the following principles:

- Use Wisely and Properly: Users of AI systems should have the necessary knowledge and ability to make the system operate according to its design, and have sufficient understanding of the potential impacts to avoid possible misuse and abuse, so as to maximize its benefits and minimize the risks.
- Informed-consent: Measures should be taken to ensure that stakeholders of AI systems are with sufficient informed-consent about the impact of the system on their rights and interests. When unexpected circumstances occur, reasonable data and service revocation mechanisms should be established to ensure that users' own rights and interests are not infringed.
- Education and Training: Stakeholders of AI systems should be able to receive education and training to help them adapt to the impact of AI development in psychological, emotional and technical aspects.

Governance

The governance of AI should observe the following principles:

- Optimizing Employment: An inclusive attitude should be taken towards the potential impact of AI on human employment. A cautious attitude should be taken towards the promotion of AI applications that may have huge impacts on human employment. Explorations on Human-AI coordination and new forms of work that would give full play to human advantages and characteristics should be encouraged.
- Harmony and Cooperation: Cooperation should be actively developed to establish an interdisciplinary, cross-domain, cross-sectoral, cross-organizational, cross-regional, global and comprehensive AI governance ecosystem, so as to avoid malicious AI race, to share AI governance experience, and to jointly cope with the impact of AI with the philosophy of "Optimizing Symbiosis".
- Adaptation and Moderation: Adaptive revisions of AI principles, policies, and regulations should be actively considered to adjust them to the development of AI. Governance measures of AI should match its development status, not only to avoid hindering its proper utilization, but also to ensure that it is beneficial to society and nature.
- Subdivision and Implementation: Various fields and scenarios of AI applications should be actively considered for further formulating more specific and detailed guidelines. The implementation of such principles should also be actively promoted – through the whole life cycle of AI research, development, and application.
- Long-term Planning: Continuous research on the potential risks of Augmented Intelligence, Artificial General Intelligence (AGI) and Superintelligence should be encouraged. Strategic designs should be considered to ensure that AI will always be beneficial to society and nature in the future.

In the United States

The Algorithmic Accountability Act of 2019 (H.R. 2231) aims to “direct the Federal Trade Commission (FTC) to require entities that use, store, or share personal

information to conduct automated decision system impact assessments and data protection impact assessments". It requires companies with annual revenue over $50m, holding data of over 1m users or working as personal data brokers, to test their algorithms and fix in a timely manner anything that is "inaccurate, unfair, biased or discriminatory".

The law, if voted in and passed, would require such companies to audit all the processes involved with sensitive data in any way such as machine learning and others. The algorithms that would fall under the audits would be those affecting legal rights of a consumer, performing predictive behaviour assessments, processing large amounts of sensitive personal data, or "systematically monitor a large, publicly accessible physical space". If the audits turn up any risks of discrimination, data privacy breaches and others, the companies would be required to address them within a timely manner. It could be the start of a significant shift in the paradigm and pave the way for many cases of tech companies finally being held accountable for their various breaches and violations of certain standards.

With the Algorithmic Accountability Act in place, together with other acts of the sort like the EU's GDPR, the global legislative frameworks would finally start catching up to the fast developments of technology and AI. Having an overhead regulation at the federal level in the United States would be a large step towards achieving that.

Furthermore, other relevant developments deserve close attention:

- the California Consumer Privacy Act (CCPA) will become effective on 1 January 2020;
- New York's privacy bill is even bolder than California's;
- San Francisco voted to ban the use of facial recognition by city agencies;
- Illinois moved against video bots for hiring interviews.

In OECD

"The OECD Principles on Artificial Intelligence promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights and democratic values. They were adopted in May 2019 by OECD member countries when they approved the OECD Council Recommendation on Artificial Intelligence. The OECD AI Principles are the first such principles signed up to by governments. Beyond OECD members, other countries including Argentina, Brazil, Colombia, Costa Rica, Peru and Romania have already adhered to the AI Principles, with further adherents welcomed."[lxxx]

The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI:

- **Inclusive growth, sustainable development and well-being**. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.
- **Human-centred values and fairness**.

    - AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.
    - To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.
- **Transparency and explainability**. AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
    - to foster a general understanding of AI systems,
    - to make stakeholders aware of their interactions with AI systems, including in the workplace,
    - to enable those affected by an AI system to understand the outcome, and,
    - to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.
- **Robustness, security and safety**.
    - AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
    - To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
    - AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.
- **Accountability**. Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles, based on their roles, the context, and consistent with the state of the art.

Scope of the various AI Ethical Guidelines

Do these different guidelines represent the various stakeholders and have they been discussed with citizens?

The guidelines of CNIL were written after a large national consultation and a public debate involving more than 3,000 people and 60 partner structures during 45 animated events on the territory, including a day of citizen consultation. The themes of education, health, employment and the internet were thus retained.

The European Commission guidelines were written by a group of around 50 experts selected from across the EU with a wide range of profiles.

The Montreal Guidelines were the result of an inclusive deliberative process that engaged citizens, experts, public officials, industry stakeholders, civil society organizations and professional bodies in roundtable discussions, in public and private sectors. classrooms or on the Internet etc.

The OECD guidelines were developed following the usual consultations of the bodies representing the Member States.

The Beijing guidelines were adopted by the Academy of Artificial Intelligence (BAAI), an organization of scientists and engineers backed by the Chinese Ministry of Science and Technology and the Beijing municipal government, universities and the three major technology companies of the country (Baidu, Alibaba and Tencent).

The Dubai guidelines do not mention how they were written and whether they were subject to stakeholder consultation.

In which political systems are the guidelines adopted?

The guidelines of CNIL, the European Commission, Montreal, have been adopted in democratic states.

Dubai's ethical principles have been broadly based on European principles, but in a very abstract way and without concrete definitions. The context in which they were adopted must be considered in assessing their scope. Dubai is ruled by a hereditary monarchy where political rights, freedom of speech, and gender equality are restricted.

The ethical principles of Beijing must also be appreciated in the context of the Chinese People's Republic with a single party, a lack of freedom of expression, and a continuous surveillance of the population for decades, only the means varying in time. Artificial Intelligence systems that monitor and score the population via widely deployed facial recognition systems clearly infringe the Beijing Guidelines principles of privacy, dignity and freedom.

This puts into perspective the meaning of these ethical guidelines and shows that they must be read in the light of the political regime of the country where they have been adopted in order to appreciate their real significance.

Do they fit into a binding legal system?

There is no authority to control the correct application of the guidelines presented above, which is quite logical since the various guidelines are not binding.

To be able to attack companies that violate these different guidelines, it will be necessary to rely on the fundamental texts to which they refer and to invoke their violation. It is therefore essential to analyse their legal framework.

CNIL guidelines must be read in the light of the Declaration of the Rights of Man and the Citizen of 1789, which is the preamble to the 1958 Constitution, the Data Protection Act, which incorporates the provisions of the European Regulation on the protection of personal data and the regulations applicable to discrimination. These are binding texts that might be invoked in the context of a trial.

The European Union guidelines are based the fundamental rights enshrined in the EU treaties, the charter of human rights (with reference to dignity, liberty, equality and justice), the UN Human Rights Treaty, the European Convention of Human Rights, and also on the General Data Protection Regulation (GDPR).

According to the Financial Times[lxxxi], the European Commission is planning regulation that will give EU citizens explicit rights over the use of their facial recognition data as part of an overhaul in the way Europe regulates artificial intelligence. The aim would be to limit the indiscriminate use of facial recognition technology by companies and public authorities. Under the plan, European citizens would be given the powers to know when facial recognition data is used, with any exceptions tightly circumscribed to ensure appropriate use. The move to explicitly legislate facial recognition technology would enhance citizens' protection above existing restrictions laid out under the GDPR (e.g., the collection of sensitive biometric data that can be used to uniquely identify people is already prohibited under the GDPR unless citizens give their explicit consent). It seems that the EU wants to draw up legislation that could emerge as a world-standard for AI regulation with clear, predictable and uniform rules adequately protecting individuals.

This reference to a body of texts is all the more important as ethical guidelines are non-binding, unlike regulatory texts. Therefore, their legal framework is essential to ensure their application. It is to be feared that isolated ethical guidelines will only be marginally applied, as long as there is no profound change in culture or regulation.

To whom are they intended and how are they implemented.

All guidelines mention the fact that they should apply to both the public and private sectors.

For CNIL "It is all along the algorithmic chain (...) that it is necessary to act, by means of a combination of technical and organizational approaches. Algorithms are everywhere: they are everyone's business". A platform called TransAlgo[lxxxii] has been created. It is a collaborative scientific platform operated by Inria, and accessible to all. It is the first in Europe. Its purpose is to provide the understanding of algorithms systems through pedagogical content, articles, videos and free software tools. For example, TransAlgo provides resources and tools that reveal to which third parties people's information is sent without their knowledge, to understand why a person is targeted by a particular advertisement when she navigated on certain social networks, what is behind the data collection, what algorithmic vulnerabilities are. TransAlgo also allows the development of new generations of "transparent by construction" algorithms that facilitate the measurement of their transparency, their explanation and the traceability of their reasoning as well as the development of so-called "responsible by construction" algorithms that respect the laws and the rules of companies.

The European Commission adopted a Trustworthy AI assessment list when developing, deploying, or using AI systems which is quite comprehensive and concrete, and has asked companies to experiment it and provide feedback. It insists on the fact that this assessment list will never be exhaustive and requires continuous improvement throughout the AI system's lifecycle involving stakeholders in it.

The appropriation of these guidelines by the stakeholders is crucial, hence the need to involve them as much as possible in the processes and to allow them to share the best

practices and the tools to do “ethics by design”. The European AI Alliance is a good example of this process. It is a forum engaged in a broad and open discussion of all aspects of Artificial Intelligence development and its impacts, which gives the stakeholders the opportunity to share their views and to comment the work of the HLEG working for a Trustworthy AI.

Developing an Ethics Impact Assessment (EIA) Framework

The components of an Ethical Impact Assessment (EIA) framework

Going beyond the identification of common ethical guidelines for Artificial Intelligence, we recommend the development of an ethical impact assessment (EIA) framework that could be used by those developing AI technologies, services, projects, policies or programmes. Such a framework would allow the technology developer, policy maker or project manager to carry out an ethical impact assessment in consultation with all key stakeholders before a new system is deployed.

Any development of an EIA framework would have to use the lessons learned from the work that scholars, experts and policy makers have done over the last 30 years or so, including: environmental impact assessment, risk assessment, technology assessment, regulatory impact assessment or simply impact assessment, privacy impact assessment. Following the work of Wright[lxxxiii], the framework would consist of a set of ethical principles, values and issues followed by a set of questions the aim of which being to facilitate ethical consideration of the new technology. The framework should be supported by:

- ethical tools (or value appraisal techniques) helping the technology developer to get a better idea of how the technology is perceived ethically by stakeholders and what measures could be adopted to ensure that the technology is ethically acceptable or what alternatives might be at his or her disposition;
- procedural aspects aimed at ensuring the ethical impact assessment is conducted in a way that engages stakeholders, ensures the transparency of the whole process and provides for independent evaluation and audit.

The table below gives a list of ethical principles, values/issues, and questions as specified by Wright. The “principles” overlap with those indicated above, though with slightly different titles, e.g., “beneficence” is similar to “generates net-benefits”. The questions asked here should serve as the basis for elaborating an AI Ethics Impact Assessment Framework, even if some are very related to privacy, less to ethics, and if further questions would need to be introduced.

Ethical Principles, Values, Issues, and Questions		
Principles	Values / Issues	Questions
Respect for autonomy (right to liberty)	Liberty (independence from controlling influences) Agency (capacity for intentional action)	Does the technology or project curtail a person's right to liberty and security in any way? If so, what measures could be taken to avoid such curtailment? Does the project recognise and respect the right of persons with disabilities to benefit from measures designed to ensure their independence, social and occupational integration and participation in the life of the community? Will the project use a technology to constrain a person or curtail their freedom of movement or association? If so, what is the justification? Does the person have a meaningful choice, i.e., are some alternatives so costly that they are not really viable alternatives? If not, what could be done to provide real choice?
	Dignity	Will the technology or project be developed and implemented in a way that recognises and respects the right of citizens to lead a life of dignity and independence and to participate in social and cultural life? If not, what changes can be made? Is such a recognition explicitly articulated in statements to those involved in or affected by the project? Does the technology compromise or violate human dignity? For example, in the instance of body scanners, can citizens decline to be scanned or, if not, what measures can be put in place to minimise or avoid comprising their dignity? Does the project require citizens to use a technology that marks them in some way as cognitively or physically disabled? If so, can the technology be designed in a way so that it does not make them stand out in a crowd? Does the project or service or application involve implants? If so, does it accord with the opinion of the European Group on Ethics (EGE)?
	Informed consent	Will the project obtain the free and informed consent of those persons to be involved in or affected by the project? If not, why not? Will the person be informed of the nature, significance, implications and risks of the project or technology? Will such consent be evidenced in writing, dated and signed, or otherwise marked, by that person so as to indicate his consent?

		If the person is unable to sign or to mark a document so as to indicate his consent, can his consent be given orally in the presence of at least one witness and recorded in writing? Does the consent outline the use for which data are to be collected, how the data are to be collected, instructions on how to obtain a copy of the data, a description of the mechanism to correct any erroneous data, and details of who will have access to the data? If the individual is not able to give informed consent (because, for example, the person suffers from dementia) to participate in a project or to use of a technology, will the project representatives consult with close relatives, a guardian with powers over the person's welfare or professional carers? Will written consent be obtained from the patient's legal representative and his doctor? Will the person have an interview with a project representative in which he will be informed of the objectives, risks and inconveniences of the project or research activity and the conditions under which the project is to be conducted? Will the person be informed of his right to withdraw from the project or trial at any time, without being subject to any resulting detriment or the foreseeable consequences of declining to participate or withdrawing? Will the project ensure that persons involved in the project give their informed consent, not only in relation to the aims of the project, but also in relation to the process of the research, i.e., how data will be collected and by whom, where it will be collected, and what happens to the results? Are persons involved in or affected by the project able to withdraw from the project and to withdraw their data at any time right up until publication? Does the project or service collect information from children? How are their rights protected? Is consent given truly voluntary? For example, does the person need to give consent in order to get a service to which there is no alternative? Does the person have to deliberately and consciously opt out in order not to receive the "service"?
Nonmaleficence	Safety	Is there any risk that the technology or project may cause any physical or psychological harm to consumers? If so, what measures can be adopted to avoid or mitigate the risk?

		Have any independent studies already been carried out or, if not, are any planned which will address the safety of the technology or service or trials? If so, will they be made public? To what extent is scientific or other objective evidence used in making decisions about specific products, processes or trials? Does the technology or project affect consumer protection? Will the project take any measures to ensure that persons involved in or affected by the project will be protected from harm in the sense that they will not be exposed to any risks other than those they might meet in normal everyday life? Can the information generated by the project be used in such a way as to cause unwarranted harm or disadvantage to a person or a group? Does the project comply with the spirit of consumer legislation (e.g., Directive 93/13 on unfair terms in consumer contracts, Directive 97/7 on consumer protection in respect of distance contracts and the Directive on liability for defective products (85/374/EEC))?
	Social solidarity, inclusion and exclusion	Has the project taken any steps to reach out to the excluded (i.e., those excluded from use of the Internet)? If not, what steps (if any) could be taken? Does the project or policy have any effects on the inclusion or exclusion of any groups? Are there offline alternatives to online services? Is there a wide range of perspectives and expertise involved in decision-making for the project? How many and what kinds of opportunities do stakeholders and citizens have to bring up value concerns?
	Isolation and substitution of human contact	Will the project use a technology which could replace or substitute for human contact? What will be the impact on those affected? Is there a risk that a technology or service may lead to greater social isolation of individuals? If so, what measures could be adopted to avoid that? Is there a risk that use of the technology will be seen as stigmatising, e.g., in distinguishing the user from other people?

	Discrimination and social sorting	Does the project or service use profiling technologies? Does the project or service facilitate social sorting? Could the project be perceived as discriminating against any groups? If so, what measures could be taken to ensure this does not happen? Will some groups have to pay more for certain services (e.g., insurance) than other groups?
Beneficence	Utility	Will the project provide a benefit to individuals? If so, how will individuals benefit from the project (or use of the technology or service)? Who benefits from the project and in what way? Will the project improve personal safety, increase dignity, independence or a sense of freedom? Does the project serve broad community goals and/or values or only the goals of the data collector? What are these, and how are they served? Are there alternative, less privacy intrusive or less costly means of achieving the objectives of the project? What are the consequences of not proceeding with development of the project? Does the project or technology or service facilitate the self-expression of users?
	Universal service	Will the project or service be made available to all citizens? When and how will this be done? Will training be provided to those who do not (yet) have computer skills or knowledge of the Internet? Who should provide the training and under what conditions? Will the service cost the same for users who live in remote or rural areas as for users who live in urban areas? How should a cost differential be paid?
	Accessibility	Does the new technology or service or application expect a certain level of knowledge of computers and the Internet that some people may not have? Could the technology or service be designed in a way that makes it accessible and easy to use for more people, e.g., senior citizens and/or citizens with disabilities? Are some services being transferred to the Internet only, so that a service is effectively no longer available to people who do not (know how to) use computers or the Internet? What alternatives exist for such people?

	Value sensitive design	Is the project or technology or service being designed considering values such as human wellbeing, dignity, justice, welfare, human rights, trust, autonomy and privacy? Have the technologists and engineers discussed their project with ethicists and other experts from the social sciences to ensure value sensitive design? Does the new technology, service or application empower users?
	Sustainability	Is the project, technology or service economically or socially sustainable? If not, and if the technology or service or project appears to offer benefits, what could be done to make it sustainable? Should a service provided by means of a research project continue once the research funding comes to an end? Does the technology have obsolescence built in? If so, can it be justified? Has the project manager or technology developer discussed their products with environmentalists with a view to determining how their products can be recycled or how their products can be designed to minimise impact on the environment?
Justice	Distributive justice (fair, equitable, and appropriate distributions determined by justified norms that structure the terms of social cooperation)	Has the project identified all vulnerable groups that may be affected by its undertaking? Is the project equitable in its treatment of all groups in society? If not, how could it be made more equitable? Does the project confer benefits on some groups but not on others? If so, how is it justified in doing so? Do some groups have to pay more than other groups for the same service? Is there a fair and just system for addressing project or technology failures with appropriate compensation to affected stakeholders?
	Equality and fairness (social justice)	Will the service or technology be made widely available or will it be restricted to only the wealthy, powerful or technologically sophisticated? Does the project or policy apply to all people or only to those less powerful or unable to resist? If there are means of resisting the provision of personal information, are these means equally available or are they restricted to the most privileged? Are there negative effects on those beyond the person involved in the project or trials and, if so, can they be adequately mediated?

		If persons are treated differently, is there a rationale for differential applications, which is clear and justifiable? Will any information gained be used in a way that could cause harm or disadvantage to the person to whom it pertains? For example, could an insurance company use the information to increase the premiums charged or to refuse cover?
Privacy and data protection (4 dimensions of privacy: the person, personal behaviour, personal communications, personal data)	Collection limitation (data minimisation) and retention	How will the project determine what constitutes the minimum amount of personal data to be collected? Who will determine what constitutes the minimum amount of personal data to be collected? Will any data be collected which is not necessary for fulfilling the stated purpose of the project? Is information collected in ways of which the data subject is unaware? Is information collected against the wishes of the person? For how long will the information be retained? Will the information be deleted when it is no longer needed for the purpose for which it was collected?
	Data quality	What measures will be put in place to ensure the quality of the information gathered? What assurances exist that the information collected is true and accurate? Has the information been collected from others than the person to whom it pertains? If the information collected is not accurate, what consequences might ensue?
	Purpose specification	Regarding the project, technology or service, are individuals aware that personal information is being (is to be) collected, who seeks it, and why? Has the purpose of collecting personal data been clearly specified? Has the project given individuals a full explanation of the purpose of the project or technology in a way that is clear and understandable? Has the person been informed of the purpose of the research, its expected duration and the procedures by means of which the data is being (will be) collected? Is there an appropriate balance between the importance of the project's objectives and the

		cost of the means? How have the goals of the data collection been legitimated? Is there a clear link between the information collected and the goal sought?
	Use limitation	Is the personal information used for the purposes given for its collection, and do the data stay with the original collector, or do they migrate elsewhere? Is the personal data collected used for profit without permission from or benefit to the person who provided it? Who will have access to or use of the data collected? Will the data be transferred to or shared with others?
	Confidentiality, security and protection of data	Has the project taken measures to ensure protection of personal data, e.g., by means of encryption and/or access control? If so, what are they? Who will have access to any personal data collected for the project or service? What safeguards will be put in place to ensure that those who have access to the information treat the information in confidence? Many service providers who provide service via the telephone say that conversations are monitored for training or quality control purposes. Will that happen in this project or service? What happens (will happen) to such recorded conversations?
	Transparency (openness)	If a new database is to be created or an existing database extended, has the data controller informed the data protection supervisory authority? Has the data controller made known publicly that he has or intends to develop a new database, the purpose of the database, how the database will be used and what opportunities exist for persons to rectify inaccurate personal information? If a database is breached or if the data controller has lost any data, has he informed the persons whose data have been compromised and/or the data protection authority? What activities will be carried out in order to promote awareness of the project, technology or service? Will such activities be targeted at those interested in or affected by the project, technology or service?

		Has an analysis been made of who are the relevant stakeholders? Are studies about the pros and cons of the project or technology available to the public?
	Individual participation and access to data	Have measures been put in place to facilitate the person's access to his or her personal data? Is there a charge for access to data and, if so, how has that charge been determined? Is the charge stated on the website of the project or service? Will the charge be perceived as reasonable by those whose data are collected and by the data protection supervisory authority? How long should it usually take to respond to requests for access to personal data and to provide such data? Can the person whose data are collected rectify easily errors in those data? What procedures are in place for doing so?
	Anonymity	Has the project taken steps to ensure that persons cannot be identified from the data to be collected? Have pseudonyms or codes been used to replace any data that could identify the individual? Is there a possibility that data from different sources could be aggregated or matched in a way that undermines the person's anonymity?
	Privacy of personal communications: monitoring and location tracking	Does the project monitor or record a person's communications? If so, is it with the person's consent? Does the project involve observation or monitoring of individuals or tracking their movements or whereabouts? If so, is it with their consent? If the project or other action involves interception of private communications, has such interception been properly authorised (e.g., has a warrant been obtained from a judge)?
	Privacy of the person	Does the project or the service or policy or program involve body searches or body scanning? Does the project involve biometrics, e.g., taking fingerprints or eye scans? Is the individual informed in advance of such requirements? How long will such data be retained and who will have access to such data?

		Have third parties been consulted with regard to the necessity of such data collection? Have less privacy-intrusive alternatives been considered?
	Privacy of personal behaviour	Does the project involve surveillance of individuals or groups of people? If so, what is the legal basis of such surveillance? Have any signs or other notifications been made to alert people to the presence of CCTV cameras or other surveillance devices? How long will images or data be retained? How will such images or data be used or erased? Who will authorise the surveillance practice, whether in public places such as city streets or banks or in assisted living residences? What measures will be put in place to avoid abuses where, for example, overseers watch others engaged in behaviour that generally accepted social norms would regard as intimate or private?

The limitations of *principlism*: dealing with "tensions"

The four principles of beneficence, non-maleficence, autonomy, and justice have played a prominent role in bioethics, a field with decades of experience in managing the challenges posed by new technologies. These principles aim to articulate general values on which everyone can agree, and to function as practical guidelines. But they have spurred substantial debate: some argue that we should put no weight on principles and focus entirely on the elements of specific cases, while others have advocated a more moderate view, whereby principles should be considered in close conjunction with analysis of paradigm cases. However, even the strongest advocates of *principlism* in bioethics acknowledge that principles alone are not enough – they should be taken as guidelines, which need to be made specific for use in policy and clinical decision-making, accompanied by an account of how they apply in specific situations, and how to balance them when they conflict.[lxxxiv]

Referring to the word "tension" to represent "any conflict, whether apparent, contingent or fundamental, between important values or goals, where it appears necessary to give up one in order to realise the other", Whittlestone, Nyrup, Alexandrova and Cave believe that in order to the gap between principles and practice, acknowledge differences in values, highlight areas where new solutions are needed and identify ambiguities and knowledge gaps, four tensions are particularly central to thinking about the ethical issues arising from the applications of AI systems in society today:

- Tension 1: Using data to improve the quality and efficiency of services vs. respecting privacy and autonomy of individuals.

- Tension 2: Using algorithms to make decisions and predictions more accurate vs ensuring fair and equal treatment. (These algorithms may improve accuracy overall, but discriminate against specific subgroups for whom representative data is not available.)
- Tension 3: Reaping the benefits of increased personalisation in the digital sphere vs enhancing solidarity and citizenship. (Personalisation can make it easier for people to find the right products and services for them, but differentiating between people in such fine-grained ways may threaten societal ideals of citizenship and solidarity.)
- Tension 4: Using automation to make people's lives more convenient and empowered vs promoting self-actualization and dignity. (With automation we may see the gifts of arts, languages and science become more accessible to those who were excluded in the past, but we may also see widespread deskilling, atrophy, ossification of practices, homogenisation and cultural diversity.)

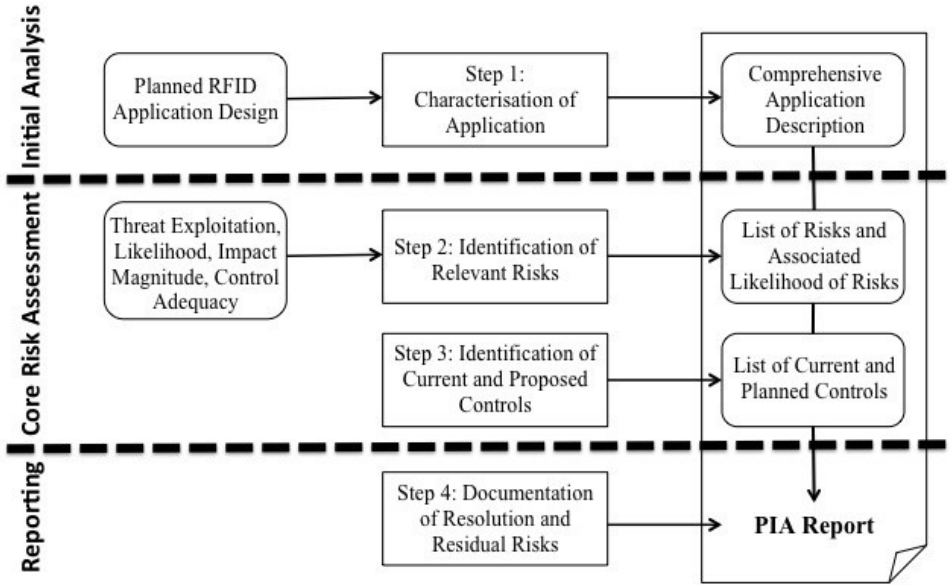
These tensions, and probably some others, deserve to be acknowledged and articulated more clearly and explicitly. On one hand, to be useful in practice principles need to be formalized in standards, codes and ultimately regulation (or self-regulation by technology and/or application domain). On the other hand, to be effective, these must acknowledge that tensions exist between the different high-level goals of AI ethics and provide guidance on how they should be resolved in different scenarios. Therefore, besides regulation (hard/self-/co-regulation), research priorities in AI ethics, in particular under the future Digital Europe programme, should include work on the main ambiguities and gaps that may blur our understanding of how AI is currently being applied in society.

AI Ethical Impact Assessment: towards a European model?

Europe should take the leadership in discussing such a framework. The 2016 General Data Protection Regulation (GDPR)[lxxxv] includes a Data Protection Impact Assessment (Article 35) that makes the European Commission and the European Data Protection Board (EDPB) relevant and credible organisations to trigger the process in consultation of all stakeholders within the European Union while remaining open to discussions with non-European countries.

Once a common set of principles, values and questions have been agreed upon, it will be possible to work on developing a comprehensive Ethics Impact Assessment Framework for Artificial Intelligence, based for example on what the European Commission did in 2011 regarding Radio Frequency Identification[lxxxvi].

RFID PIA Process Reference Model (EC, 2011)



Using the RFID PIA case as a metaphor and fundament for AI EIA, discussions in the European Commission, and beyond, should contemplate a process that includes at least the following steps:

- describing the system;
- identifying and listing how the system under review could threaten ethics and evaluation of the magnitude and likelihood of those risks;
- Documenting current and proposed technical, managerial and organisational controls to mitigate identified risks; and
- documenting the resolution (results of the analysis) regarding the system.

It is urgent to start working on such an AI EIA Framework/Methodology since the discussions within the EU will take a minimum of three years and meanwhile Artificial Intelligence will continue to thrive on scientific and technological advances and eventually pervade all recesses of our societies and individual intimacies.

---

[i] Geneviève Fieux-Castagnet is deontologist at French Railways (SNCF).

[ii] Gérald Santucci is retired from the European Commission where he worked over more than thirty years at Directorate-General Communications Networks, Content and Technology (DG_CNECT), in particular with responsibility for Radio-Frequency Identification and the Internet Things.

[iii] Council of Europe Commissioner for Human Rights, "unboxing Artificial Intelligence: 10 steps to protect human rights", Council of Europe, May 2019.

[iv] Sources: report on "AI in the UK: ready, willing and able?", House of Lords Select Committee on Artificial Intelligence, 16 April 2018; Council of Europe Commissioner for Human Rights, op.cit., Annex, Definitions, pp. 24-25.

[v] SIEBEL Thomas M., Digital Transformation: Survive and Thrive in an Era of Mass Extinction, RosettaBooks, New York, 2019, page 4.

[vi] Source: Innosight.

[vii] WANG Ray, founder and principal analyst at Constellation Research, Forbes, 19 December 2014.

[viii] Source: LEE Kai-Fu, interview on "the Power of AI to Transform Humanity", by ZUCKERMAN Andrew, 25 April 2019.

[ix] WALCH Kathleen, Principal Analyst, Cognilytica, "The Seven Patterns of AI", Forbes, 17 September 2019.

[x] For example, "The European Unicorns Reunion", France Digital Day (#FDDay), 18 September 2019.

[xi] Source: https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

[xii] "Pan-Canadian Artificial Intelligence Strategy", 12 April 2017.

[xiii] According to Gavin Menzies, an amateur historian and a former submarine commanding officer who has spent 14 years charting the movements of a Chinese expeditionary fleet between 1421 and 1423, the eunuch admiral, Zheng He, was in America 72 years before Colombus and, in his colossal multi-masted ships stuffed with treasure, silks and porcelain, made the first circumnavigation of the world, beating the Portuguese navigator Ferdinand Magellan by a century.

[xiv] "Artificial Intelligence Strategy", report of the Federal Government, November 2018.

[xv] Report on "Automated and Connected Driving", Ethics Commission, Federal Ministry of Transport and Digital Infrastructure (BMVI), June 2017.

[xvi] "Artificial Intelligence Technology Strategy", Report of Strategic Council for AI Technology, 31 March 2017.

[xvii] This was not the first time South Korea had made a big commitment to AI. In 2016, after DeepMind's AlphaGo defeated Korean Go Master Lee Sedol, which shocked Koreans, the country announced it would invest ₩1 trillion (US$863 million) in AI research over the next five years.

[xviii] https://government.ae/en/about-the-uae/strategies-initiatives-and-awards/federal-governments-strategies-and-plans/uae-strategy-for-artificial-intelligence

[xix] https://uaecabinet.ae/en/details/news/mohammed-bin-rashid-launches-five-decade-government-plan-uae-centennial-2071

[xx] Policy paper, "Artificial Intelligence Sector Deal", Sector Deal between government and the Artificial Intelligence (AI) sector, 26 April 2018 (last updated 21 May 2019) https://www.gov.uk/government/publications/artificial-intelligence-sector-deal

[xxi] "AI in the UK: ready, willing and able?", Select Committee on Artificial Intelligence, House of Lords, 16 April 2018.

[xxii] Artificial Intelligence for the American People https://www.whitehouse.gov/ai/

[xxiii] NITRD, Supplement to the President's 2020 Budget, https://www.whitehouse.gov/wp-content/uploads/2019/09/FY2020-NITRD-AI-RD-Budget-September-2019.pdf

[xxiv] Trustworthy Artificial Intelligence should respect all applicable laws and regulations, as well as the following requirements: **Human agency and oversight**: AI systems should enable equitable societies by supporting human agency and fundamental rights, and not decrease, limit or misguide human autonomy. **Robustness and safety**: Trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of AI systems. **Privacy and data governance:** Citizens should have full control over their own data, while data concerning them will not be used to harm or discriminate against them. **Transparency:** The traceability of AI systems should be ensured. **Diversity, non-discrimination and fairness:** AI systems should consider the whole range of human abilities, skills and requirements, and

ensure accessibility. **Societal and environmental well-being:** AI systems should be used to enhance positive social change and enhance sustainability and ecological responsibility. **Accountability:** Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.

[xxv] The recommendations call on EU and national policymakers to: **Empower and protect humans and society:** ensure individuals understand the capabilities, limitations and impacts of AI; protect them from any harm; and provide them with the necessary skills to use and benefit from AI. **Take up a tailored approach to the AI market:** assess the different needs and sensitivities raised by AI systems used in Business-to-consumers (B2C), Business-to-business (B2B) and Public-to-Citizens (P2C) contexts, and address these accordingly. **Secure a Single European Market for Trustworthy AI:** remove barriers to procure lawful, ethical and robust AI-enabled goods and services from all over Europe, while enabling a competitive global position through large integrated markets. **Enable AI ecosystems through sectoral multi-stakeholder alliances:** boost stakeholder cooperation across civil society, industry, the public sector and research and academia, while understanding the different impacts and enablers for different sectors. **Foster the European data economy:** further advance policy actions in data access, sharing, reusing and interoperability, while ensuring high privacy and data protection, and putting in place the necessary physical infrastructures. **Exploit the multi-faceted role of the public sector:** ensure the public sector leads by example by delivering human-centric public services, making strategic use of innovation-driven public procurement, and fostering cooperation with stakeholders. **Strengthen and unite Europe's research capabilities:** establish and demonstrate intellectual and commercial leadership in AI by bringing together European research capacity in a multidisciplinary manner. **Nurture education to the Fourth Power:** ensure a wide skills base through primary, secondary and tertiary education, as well as enabling continuous learning and strive towards a work-life-train balance. **Adopt a risk-based governance approach to AI and ensure an appropriate regulatory framework:** map relevant laws, assess to which extent these are still fit for purpose in an AI-driven world, and adopt new measures where needed to protect individuals from harm, thus contributing to an appropriate governance and regulatory framework for AI. **Stimulate an open and lucrative investment environment:** enhance investment levels in AI with both public and private support. **Embrace a holistic way of working, combining a 10-year vision with a rolling action plan:** look at AI's overall opportunities and challenges for the next 10 years, while continuously monitoring the AI landscape and adapting actions on a rolling basis as needed; join forces with all stakeholders for the concrete implementation of the ethics guidelines and policy recommendations.

[xxvi] Source: EC Study on "identification and quantification of key socio-economic data for the strategic planning of 5G introduction in Europe", SMART 2014/0008.

[xxvii] Source: "Cutting through the 5G hype: Survey shows telcos' nuanced views", McKinsey, February 2019.

[xxviii] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Radio Frequency Identification (RFID) in Europe: steps towards a policy framework – EU-COM(2007) 96 final.

[xxix] The International Telecommunication Union (ITU) published in 2005 a report on the Internet of Things, the seventh report in the series of ITU Internet Reports (1997-2005), which was written by a team of analysts from ITU's Strategy and Policy Unit (SPU) led by Lara Srivastava.

[xxx] Commission Recommendation of 12 May 2009 on the implementation of privacy and data protection principles in applications supported by radio- frequency identification (COM)2009) 387 EC; Communication of 18 June 2009 from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Internet of Things: an action plan for Europe (COM(2009) 0278 final; DG CONNECT Internal Report on the implementation of the Commission Recommendation on the implementation of privacy and data protection principles in applications supported by radio-frequency identification (27/08/2014).

[xxxi] The scope of the work of the IoT-EG was broad – Identification, Architectures, Privacy, Ethics, Standards, Governance. Because of the complexity of these topics and the variety of opinions among the stakeholders, strong 'concertation' was needed to reach a minimum consensus in the perspective of a Recommendation. But DG CONNECT considered that another year of discussions to complete the work would not be an

effective approach, especially because the question of whether the Internet of Things really was substantially new, or if it was just an extension of the existing Internet had not been settled yet. DG CONNECT followed the opinion of some of the experts according to whom a secondary forum for political questions regarding the Internet of Things should not be instantiated and relevant issues should be dealt with within existing multi-stakeholder structures of Internet Governance such as the Internet Governance Forum (IGF).

[xxxii] During the implementation of the Telematics Applications Programme (TAP, 1994-1998), part of the EU Framework Research Programme 4 (FP4), Michel RICHONNIER, Director of the TAP, had explicitly set as one of the programme's objectives the emergence of a "European Google".

[xxxiii] As many as 28 of 30 DSM legislative proposals were agreed between Parliament, Council and Commission before the end of the five-year mandate in April 2019.

[xxxiv] For more insights, read: DITTRICH Paul-Jasper, Research Fellow at Jacques Delors Institut Berlin, "New Beginnings: Challenges for EU Digital and Innovation Policy", 2 September 2019. https://www.delorsinstitut.de/en/all-publications/new-beginnings-digital-europe/

[xxxv] See https://www.theinternetofthings.eu/sophie-le-pallec-rick-bouter-gérald-santucci-metamorphosis-objects-and-human-subjects-internet

[xxxvi] "The Jobs That Artificial Intelligence Will Create", MIT Sloan Management Review, Vol. 58, No. 4, Summer 2017.

[xxxvii] https://iapp.org/news/a/study-gdprs-global-reach-to-require-at-least-75000-dpos-worldwide/

[xxxviii] SIEBEL (Thomas M.), Digital Transformation: Survive and Thrive in an Era of Mass Extinction, RosettaBooks, New York, 2019, page 104.

[xxxix] FREY (Carl Benedikt) and OSBORNE (Michael A.), "The future of employment: How susceptible are jobs to computerization?", study released at the "Machines and Employment" Workshop hosted by the Oxford University Engineering Sciences Department and the Oxford Martin Programme on the Impact of Future Technology, September 17, 2013. The authors modelled the characteristics of 702 occupations and classified them according to their "susceptibility to computerization".
https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

[xl] To be honest, the study does not say that half of all jobs are going to be automated in a decade or two, leaving half the population unemployed. It only suggests that the occupations concerned are the most vulnerable to automation. It makes no attempt to estimate how many jobs will actually be automated since that will depend on other things, such as cost, regulation, political pressure, and social resistance.

[xli] "Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages", McKinsey report, November 2017, https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

[xlii] IBM Study: The Skills Gap is Not a Myth, But Can Be Addressed with Real Solutions", IBM News Room, 6 September 2019. The study shows that new skills requirements are rapidly emerging, while other skills are becoming obsolete. In 2016, executives ranked technical core capabilities for STEM and basic computer and software/application skills as the top two most critical skills for employees. In 2018, the top two skills sought were behavioural skills – willingness to be flexible, agile, and adaptable to change and time management skills and ability to prioritize. Ethics and integrity is also a skill often named most critical.

[xliii] Virtual assistants can be contrasted with another type of consumer-facing AI programming, called "smart advisers". Smart adviser programs are subject-oriented, while virtual assistants are task-oriented.

[xliv] Project Maven: an initiative aimed at developing better AI for the US military. The Pentagon offered Google a defence contract to help military drones gain the ability to track objects.

[xlv] In 2018, Google recognized it had been working on a censored search engine for China. Facing pressure from lawmakers, stockholders, and human rights groups alike, Google confirmed to US officials in July 2019 that this project was over.

[xlvi] Precise voice-recognition processing is part of Amazon's push to bring its AI expertise and automation to just about every layer of its business, including its warehouse robots, cashierless retail stores, and its virtual assistant Alexa. Alexa, first launched in 2014, along with the Amazon Echo smart speaker, can provide results for web searches, order products from Amazon, and act as a hub for compatible IoT devices all via voice command. In 2018, Amazon made Alexa's API available to developers, allowing for integration in non-Amazon devices.

[xlvii] The "belt" refers to an overland push across Eurasia and the "road" to a maritime route to South Asia and beyond.

[xlviii] Although the original trading routes were established more than 2,000 years ago, the "Silk Road" name was coined in 1877 by a German geographer (Ferdinand von RICHTOFEN). As US involvement in international trade agreements is scaled back, President XI Jinping is using the Belt and Road Initiative to position himself as a champion of global cooperation and development as well as free trade.

[xlix] Source: Freedom House research director Adrian SHAHBAZ, report on "Freedom on the Net 2018: The Rise of Digital Authoritarianism" – https://freedomhouse.org/report/freedom-net/freedom-net-2018/rise-digital-authoritarianism. Freedom House is a democracy watchdog group whose main financier is the US Government.

[l] The CARIN Alliance is a non-partisan, multi-sector alliance which is led by distinguished risk-bearing providers, payers, consumers, pharmaceutical companies, consumer platform companies, health IT companies, and consumer-advocates who are working collaboratively with other stakeholders in government to overcome barriers in advancing consumer-directed exchange across the U.S.

[li] See "*Les GAFA s'immiscent dans les systèmes de santé publique*", Les Échos, 5/09/2019, page 24.

[lii] Source: "*Scalable and accurate deep learning with electronic health records*", Nature, npj Digital Medicine 1, Article Number: 18 (2018).

[liii] China's Song Dynasty military forces as early as 904 A.D. used gunpowder devices against their primary enemy, the Mongols. Military applications of gunpowder included flying fire arrows, primitive hand grenades, poisonous gas shells, flamethrowers and landmines. By the mid- to late-eleventh century, the Song government had become concerned about gunpowder technology spreading to other countries. The sale of saltpetre to foreigners was banned in 1076. Nonetheless, knowledge of the miraculous substance was carried along the Silk Road to India, the Middle East, and Europe. In 1267, a European writer referred to gunpowder, and by 1280 the first recipes for the explosive mixture were published in the West. China's secret was out.

[liv] Future of Life Institute, *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, 2015.

[lv] ROTHBLATT Martine, Virtually Human: The Promise – and the Peril – of Digital Immortality, St. Martin's Press, New York, 2014, foreword by Ray KURZWEIL, pp. 3-7.

[lvi] SCHNEIDER Susan, Associate Professor of Philosophy and Cognitive Science at University of Connecticut, *The problem of AI Consciousness*, Kurzweil accelerating intelligence, daily blog, 18 March 2016.

[lvii] ROBITZSKI Dan, Staff Reporter Futurism, and Freelance Writer: "*Will we be da Vinci, painting a self-amused woman who will be admired for centuries, or will we be Uranus, creating gods who will overthrow us? Right now, AI will do exactly what we tell AI to do, for better or worse. But if we move towards algorithms that begin to, at the very least, present as sentient, we must figure out what that means.*"

[lviii] BLOCH Emmanuel, Director Corporate Responsibility at Thales, "Les algorithmes ont-ils une âme?", Matinée de l'Éthique, Saint-Denis (France), 6 June 2019.

[lix] See for example: "AI and Consciousness: Theoretical foundations and current approaches", AAAI Symposium, Washington, DC, 8-11/11/2007; Susan Schneider, Associate Professor of Philosophy and Cognitive Science at the University of Connecticut and faculty member in the technology and ethics group at Yale's Interdisciplinary Center for Bioethics, "The problem of AI consciousness", 18/03/2016; "Will artificial intelligence become conscious?", The Conversation, 08/12/2017.

[lx] LEVANDOWSKI Anthony, interviewed by Mark HARRIS, a freelance journalist reporting on technology from Seattle, "Inside the First Church of Artificial Intelligence", WIRED, 15/11/2017.

[lxi] HARARI Yuval Noah, Homo Deus: A Brief History of Tomorrow, HarperCollins, 2017, page 367. "Dataism declares that the universe consists of data flows, and the value of any phenomenon or entity is determined by its contribution to data processing."

[lxii] DUGAIN Marc, Transparence, Gallimard, 25/04/2019.

[lxiii] PAUL-CHOUDHURY Sumit, "Tomorrow's Gods: What is the future of Religion?", BBC, 2 August 2019.

[lxiv] MÜLLER Vincent C. (ed.), Eindhoven University of Technology, "Fundamental Issues of Artificial Intelligence", Springer, Berlin, 2016.

[lxv] SPIEKERMANN Sarah, Vienna University of Economics and Business, Institute for Management Information Systems, Ethical IT Innovation: A Value-Based System Design Approach, CRC Press, Taylor & Francis Group, 2016, page 14.

[lxvi] KLUCKHOHN Clyde, "Values and Value Orientations in the Theory of Action: An Exploration in Definition and Classification", in *Toward a General Theory of Action*, edited by Talcott Parsons, Edward Albert Shils, and Neil J. Smelser, 388-433, Cambridge, MA: Transaction Publishers, 1962, page 395.

[lxvii] In March 2018, The New York Times, working with *The Observer* of London and *The Guardian*, obtained a cache of documents from inside Cambridge Analytica, the data firm principally owned by the right-wing donor Robert Mercer. The documents proved that the firm, where the former Trump aide Stephen Bannon a board member, used data improperly obtained from Facebook to build voter profiles. The news put Cambridge Analytica under investigation and thrust Facebook into its biggest crisis ever.

[lxviii] "Your chosen target will be exposed to 10 articles on major social networks and news sites. The articles are picked by The Spinner's editors' team in order to support the desired narrative. You will receive an innocent looking link. (You can choose the link address). You will send that link to your target via email or text message. After clicking the link, the target will start seeing the articles chosen for him or her. For any language other than English – you'll have to select the articles yourself. Send us the articles you want to deliver, and we will make sure the articles fit ads platforms' criteria. When you are ready to set and launch you Tailor-Made microtargeted campaign, please contact us TailorMade@TheSpinner.net."

[lxix] In 2018, China had an estimated 200 million surveillance cameras (300 million cameras installed by 2020), i.e. four times as many as the United States. Source: The New York Times, 8 July 2018.

[lxx] RIO Edern, "Algorithmes de risques de récidive: quand l'IA determine votre peine de prison", Opportunités Technos, 24 janvier 2019, https://opportunites-technos.com/algorithmes-de-risques-de-recidive-quand-ia-determine-votre-peine-de-prison/

[lxxi] Source: VILLASENOR John, Governance Studies, Centre for Technology Innovation, "Artificial Intelligence and bias: Four key challenges", Brookings, 3 January 2019.

[lxxii] The first comprehensive report on the potential of Artificial Intelligence for the future of humans is the so-called "NBIC Report": Converging Technologies for Improving Human Performance: Nanotechnology,

Biotechnology, Information Technology and Cognitive Science, US National Science Foundation (NSF), June 2002.

[lxxiii] Recital 71: The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measure should not concern a child. In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject, and prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.

[lxxiv] **Definitions of AI ethics principles**: Accountability. People and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended. AI Code. UK House of Lords Select Committee on Artificial Intelligence recommends to establish an AI Code, which can be adopted nationally and internationally, based on five principles: 1. Artificial intelligence should be developed for the common good and benefit of humanity. 2. Artificial intelligence should operate on principles of intelligibility and fairness. 3. Artificial intelligence should not be used to diminish the data rights or privacy of individuals, families or communities. 4. All citizens should have the right to be educated to enable them to flourish mentally, emotionally and economically alongside artificial intelligence. 5. The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence. Contestability. When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm. Do not harm. Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes. The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence. Flourishing alongside AI. All people should have the right to be educated as well as be enabled to flourish mentally, emotionally and economically alongside artificial intelligence. For children, this means learning about using and working alongside AI from an early age. For adults, government should invest in skills and training to negate the disruption caused by AI in the jobs market. Intelligibility, diversity, non-discrimination and fairness. The development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly. Generates net-profit. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being; it should generate benefits for people that are greater than the costs. Human agency and oversight. Human dignity. Privacy and Data governance. Any system, including AI systems, must ensure people's private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm. Regulatory and

legal compliance. AI systems must comply with all relevant government obligations, regulations and laws. They should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society. Societal and environmental well-being. Technical robustness and safety. AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed. Transparency & Explainability. People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions. In other words, there should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.

[lxxv] FAN Shelly, "What does Ethical AI Look Like? Here's What the New Global Consensus Says", SingularityHub, 10 September 2019.

[lxxvi] Source: https://www.cnil.fr/fr/comment-permettre-lhomme-de-garder-la-main-rapport-sur-les-enjeux-ethiques-des-algorithmes-et-de

[lxxvii] Source: https://www.montrealdeclaration-responsibleai.com

[lxxviii] Source: https://www.smartdubai.ae/initiatives/ai-principles-ethics

[lxxix] Beijing AI Principles, https://www.baai.ac.cn/blog/beijing-ai-principles

[lxxx] Source: https://www.oecd.org/going-digital/ai/principles/

[lxxxi] Source: "EU plans sweeping regulation of facial recognition", Financial Times, 22 August 2019.

[lxxxii] TransAlgo: Plateforme scientifique pour le développement de la transparence et de la redevabilité des algorithmes et des données, https://www.inria.fr/actualite/actualites-inria/transalgo

[lxxxiii] WRIGHT David, Director of Trilateral Research & Consulting (a London-based research company), "A framework for the ethical impact assessment of information technology", article published online: 8 July 2010.

[lxxxiv] According to some scholars, (i) principles are highly general, (ii) principles come into conflict in practice, (iii) different groups may interpret principles differently. WHITTLESTONE Jess, NYRUP Rune, ALEXANDROVA Anna, CAVE Stephen, Leverhulme Centre for the Future of Intelligence, University of Cambridge, "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions", Association for the Advancement of Artificial Intelligence (www.aaai.org), 2019.

[lxxxv] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

[lxxxvi] The European Commission achieved an important landmark in impact assessment when Vice-President Neelie KROES signed on 6th April, 2011, the Privacy Impact Assessment Framework (PIAF) for RFID applications after 18 months of extensive work among all stakeholders from industry, academia, and representatives of society (like BEUC and ANEC). The RFID PIA Framework was the first foray of Europe into the PIA activity, whereas the concept was already frequently used in Canada, the United States, Australia, New Zealand and other countries around the world. For Sarah Spiekermann, who took part in the RFID Expert Group that was established by European Commission DG Information Society and Media (DG INFSO, now DG CNECT) in the aftermath of the 2009 RFID Recommendation, "the only PIA guideline with a valid process model is the PIA Framework for RFID, which was endorsed by the Article 29 Working Party and signed by the European Commission (...) This framework encourages European RFID application operators to run through a four-step PIA process: (1) describe their system landscape, (2) identify privacy risks, (3) mitigate those risks through appropriate controls, and (4) document the analysis and residual risks in a PIA report. This four-step methodology has been called a "landmark for privacy-by-design" by Ontario's data protection authorities, who invented the concept of privacy-by-design." OETZEL Marie, SPIEKERMANN Sarah, "A systematic

methodology for privacy impact assessments – a design science approach", European Journal of Information Systems (EJIS), Vol. 23, pp. 126-150, July 2013.