Artificial Intelligence: Fiction, reality and… dreams

Nuria Oliver (nuria@alum.mit.edu)

Preface

His Excellency Mr. President of the Spanish Royal Academy of Engineering, Your Excellencies Ladies and Gentlemen members of the Spanish Royal Academy of Engineering, Ladies and Gentlemen; dear friends,

I would like to start by thanking you all for attending this solemn and emotional ceremony for the public reception of my status as member of the Spanish Royal Academy of Engineering. Thank you for the effort you have made to be here today and sharing this memorable moment with me.

I must continue by expressing how honored I feel to receive a medal as special as No. 1 of this Royal Academy, which was deposited by the illustrious academic Mr. Eugenio Andrés Puente, to whom the Academy paid a posthumous tribute on September 25th this year. Professor Andrés Puente was a leading figure in the Industrial Engineers school at the Polytechnic University of Madrid and a pioneer of research in industrial engineering in Spain. In addition to an excellent teaching career and a large scientific production, he stood out for his dedication to the promotion of industrial innovation and his close collaboration with the industry, as is my case. For this reason, he received prestigious awards such as the Encomienda with plaque of the Civil Order of Alfonso X the Wise and the First Merit Cross of the Order of Merit of the Federal Republic of Germany, where he obtained his doctorate. So, I receive this medal, which was his, with a great sense of responsibility and a firm commitment to be worthy of it.

Next, I express my sincere and profound gratitude to all the members of this corporation for believing in me and for endorsing my candidacy with your vote. In particular, I am grateful for the support of the academics Dr. Javier Aracil, Dr. Josefina Gomez, Dr. Manuel Marquez and Dr. Elias Muñoz who promoted my candidacy and I am also grateful to those who supported it with their vote. I started my relationship with the Royal Academy of Engineering before being proposed as a candidate. For this reason, I would like to highlight the support received from Javier Pérez de Vargas and Prof. Sara Gómez in the first activities in which I have participated in this house. Finally, I thank the President, his Excellency Mr. Elías Fereres, for guiding me in these first steps, for his confidence, his valuable and constructive feedback with this speech and his great dose of patience. Thank you all for your interest in my profile and for your courage to promote and support a candidacy that is at least original: I am a woman (the fourth of sixty academics), the first telecommunications engineer and perhaps of insufficient age. Thank you for the opportunity.

Eleanor Roosevelt said that *the future belongs to those who believe in the beauty of their dreams*. And I share Anaïs Nin's words, *dreams are necessary for life*. In my life there have always been many dreams, dreams that have inspired me, encouraged me and helped me find my way. Therefore, today I would like to share with you some of my professional dreams, both those that over time have become, to my surprise, a reality, and others that I would love to see come true in the future.

Several months before the Spanish Royal Academy of Engineering was created, as I was studying telecommunications engineering at the UPM, I discovered Artificial Intelligence. It was love at first sight. At the time I was using SGI workstations which had less processing power than most peoples' phones today. Technology, with its speed of exponential progress, makes us feel even older than we are.

At that moment, my first dream became clear: to study my PhD in Artificial Intelligence in the United States.

I was lucky to be able to realize such a dream, thanks to a fellowship from the la Caixa Foundation and to the opportunity that my thesis director, Professor Sandy Pentland of MIT, gave me to do a PhD in its mythical Media Lab. With Sandy I started my scientific career, I published my first scientific articles and gave my first talks. I will always be grateful to Sandy for the opportunity, for his valuable support, and for his teachings and advice for over two decades. After my PhD I began my professional career as a researcher at Microsoft Research. Although I always wished to return to Spain, I never thought that one day I could do it. Therefore, for a long time a possible return was just another of my dreams.

However, eleven years ago, Fortune -well, it was actually Telefónica- again knocked on my door with an opportunity to see that dream fulfilled: return to Spain as Scientific Director -the first female director- at Telefónica R&D in Barcelona.

They say that life is cyclical (the circle of life) and certainly that has been my case, offering me the opportunity to fulfill another dream that I thought would never see fulfilled: being able to return to Alicante, my hometown. However, three years ago we moved to Alicante to be closer to my family. Thanks to technology and to the support of the organizations where I have worked and currently work at, I have been able to turn into reality what for decades seemed impossible.

I reiterate my gratefulness to all my collaborators and mentors during my professional career, and especially to Prof. Sandy Pentland, Eric Horvitz, Mary Czerwinski, Jennifer Chayes, Carlos Domingo, Emmanuel Letouzé and Katia Walsh, for their valuable advice and the full confidence that they have always given me.

This recognition is for all of you, since without your help, support and collaboration, I would never have been able to achieve what we have achieved together.

For approximately 25 years, I have focused my research activity on the computational modeling of human behavior using artificial intelligence techniques. I have worked on smart rooms, clothes, cars and phones. I have invented systems capable of recognizing human behaviors or characteristics such as: facial expressions [1], activities [2], human interactions [3], driving maneuvers [4], sleep apnea [5], credit scoring [6], crime hotspots in cities [7] or even boredom [8]…I have built interactive, intelligent systems on computers and mobile phones, *e.g.* [9] [10]. Thanks to having developed my scientific career in a business context, I have been able to feel, in the first person, the deep happiness that floods you when you see that what was no more than an idea -sometimes even a bit crazy- becomes a reality that can help millions of people.

I have been not only a witness but also an active participant in technological progress, in the increasingly relevant and ubiquitous presence of technology in our lives and in the co-dependency we have developed towards it.

Throughout my career, the inspiration and drive of my work has always been a question with a clear social application. In all my projects, the person -individually or collectively- has been and still is the central element: technology by and for society; Technology that understands us, as a necessary step to be able to help us. However, today, more than ever, I am concerned about the impact that this technology is having on our lives, an impact that is not necessarily always positive[1].

Therefore, in addition to describing some of my professional dreams -past and present- I have decided to devote this speech to Artificial Intelligence, making a brief tour of its history and sharing my vision from the future.

Brief History of Artificial Intelligence (AI): my personal account

Although Artificial Intelligence (AI) might seem to be something new, the idea of machines capable of thinking or endowed with some human capabilities has captivated our interest since ancient times. The first automatons -anthropomorphic robots- that imitated human movements were built millennia ago. According to the Iliad, Hephaestus -the Greek god of fire and forge- created two artificial golden women with *"sense in their entrails, strength and voice"* that freed him from part of his work, that is, robots to help him, which showed he was a real visionary. Other examples include automatons with religious purposes -for example, the mechanical figures of the gods in Ancient Egypt to surprise the crowd with gestures operated by priests- or for entertainment -such as the famous "*talking heads*" and automata of the Middle Ages, the Renaissance and the eighteenth century.

Beyond automation, human beings have always been interested in explaining and understanding how the human brain works, among other reasons, as a necessary requirement to build an artificial mind. More than 700 years ago, Ramon Llull -a religious philosopher from Majorca- describes in his Ars Magna (1315) the creation of the Ars Generalis Ultima, a mechanical device capable of analyzing and validating or invalidating theories using logic; that is, he describes an Artificial Intelligence system.

In the 19th century, Ada Byron was the first person the use of Babbage's analytical machine to solve complex problems beyond mathematical calculations. However, the myth and literary fiction regarding Artificial Intelligence began to become a reality in the 1940s with the development of the first computers.

Alan Turing, the great English mathematician and computer scientist, is considered to be the father of Artificial Intelligence, of which he spoke in his legendary article *"Computing Machinery and Intelligence"*, published in 1950. In that article he proposed the famous Turing test, designed to determine if an artificial system is intelligent or not, and illustrated in Figure 1. This test consists of a human (C in the figure), known as the "interrogator", who interacts via text with a system to which

[1] http://humanetech.com/

(s)he can ask questions. The system is considered to pass the Turing test if the interrogator is unable to discern when the answers to his/her questions are answered by a machine (A in the figure) vs by a human (B in the figure).

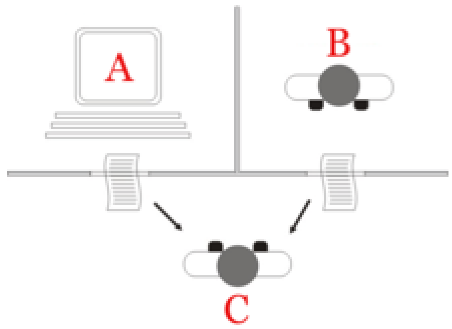


Fig 1. The Turing test (source: Wikipedia)

The 1950s were a foundational decade for Artificial Intelligence. In 1951, Professor Marvin Minsky - whom I had the honor of meeting during my time at MIT- built the first computational neural network as part of his PhD at Princeton. In 1956, the legendary Dartmouth conference took place with the participation of key figures, such as John McCarthy, Marvin Minsky, Claude Shannon, Herbert Simon and Allen Newell. It was in this conference where the term Artificial Intelligence was defined and the foundations for its development were established.

In this conference, McCarthy defined Artificial Intelligence as *"the discipline within Computer Science or Engineering that deals with the design of intelligent systems"*, that is, systems with the ability to perform functions associated with human intelligence, such as learning, understanding, adaptation, reasoning and interacting in a similar way to human beings. McCarthy coined the term to differentiate it from the concept of cybernetics, promoted by Norbert Wiener (a professor of MIT at the time) whose vision of intelligent systems was based on pattern recognition, statistics, and control and information theories. McCarthy wanted to emphasize the connection of Artificial Intelligence with logic. In a wink of fate, Wiener's intellectual proposal has become the dominant approximation to Artificial Intelligence but using McCarthy's terminology.

Fig. 2. Participants of the Darmouth conference, including Marvin Minksky, Claude Shannon and Ray Solomonoff (source: achievement.org)

In 1957 Rosenblatt proposed the perceptron at the Cornell Aeronautical Laboratory. One of its first implementations was called the Mark 1 Perceptron (illustrated on the right part of Figure 3), whose application was image analysis. Rosenblatt, who was a psychologist, was convinced that the Perceptron was a simplified version of how neurons worked. As can be seen in Figure 3 (left), the Perceptron receives a set of input values (which represent the activations of adjacent neurons), multiplies each input value by one weight (representing the strength of the synapse with each adjacent neuron), and produces an output which is a 1 if the sum of the inputs modulated by their weights is greater than a certain value, and 0 if it is lower (representing therefore whether that neuron is activated or not). This model is based on previous work by McCulloch and Pitts who demonstrated that a neuron's model like the one described above can represent OR/AND/NOT functions. This result was important because at the time it was thought that we would achieve Artificial Intelligence when computers were able to carry out formal logical reasoning operations.

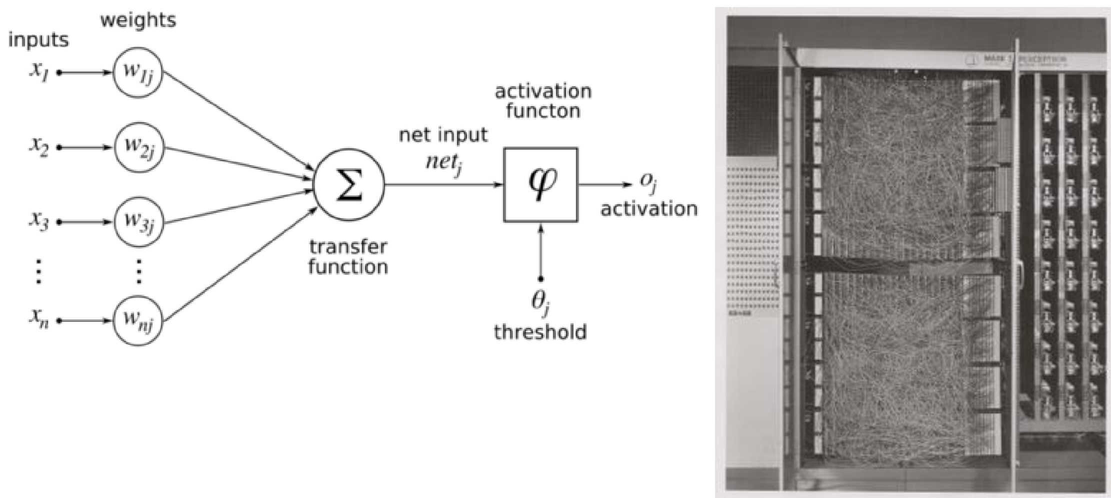


Fig. 3. Left: Rosenblatt's Perceptron (source: Wikipedia); Right: Mark 1 Perceptron from Cornell's Aeronautics Laboratory (source: Wikipedia)

Throughout history there has been some level of tension between two schools of thought with respect to AI: the *symbolic-logical* approach (originally called the *neats*) or *top-down* and the *data-driven* approach (originally known as the *scruffies*) or *bottom-up*. The symbolic approach postulated that one would need to follow a set of predefined rules and the principles of logic to develop machines that reasoned. On the other hand, the bottom-up school proposed that intelligence should be inspired by biology, so that computer systems would have to learn from observation and experience, that is, from data.

In Artificial Intelligence systems, human intelligence is usually taken as the reference. Since human intelligence is diverse and multiple, Artificial Intelligence is also a discipline with many branches of knowledge. The symbolic-logical school includes, among others, game theory, logic, optimization, reasoning and knowledge representation, automatic planning and learning theory. The bottom-up school includes computational perception (one of my areas of expertise, which includes image, video, text, audio and other sensor data processing), statistical machine learning (another of my areas of expertise), reinforcement learning, search methods (for *e.g.* text, images, videos …), agent systems, robotics, reasoning with uncertainty, human-AI collaboration, recommendation and personalization systems, computational social sciences and affective computing.

At the beginning, the bottom-up approach -to which I belong- did not have much practical success since we did not have the necessary data and computing capabilities to train our models. For this reason, in the 1960s the first successful practical application of Artificial Intelligence systems was with expert systems, which belong to the symbolic-logical approach. The period between 1956 and 1974 is usually known as the first golden age of Artificial Intelligence. Feigenbaum -one of the founders of the computer science department at Stanford University- led the team which built the first expert system, implemented in LISP, the computer program developed by McCarthy.

Until 1974 there was a period of optimism regarding Artificial Intelligence and its impact. In 1956, after the famous Dartmouth conference, Herbert Simon predicted that *"in twenty years, machines will be able to do a person's work",* and in 1970, Marvin Minsky stated in Life magazine that *"within three to eight years we will have a machine with the general intelligence of a human being"*.

However, these expectations were not fulfilled. Between 1974 and 1980 Artificial Intelligence experienced what is known as its first winter, which was characterized by a decline in the interest and funding of the field. In addition, after 1969 the interest towards bottom-up models and specifically neural networks fell significantly, particularly after the publication of the book *Perceptrons* by Minsky and Papert, whom I also had the honor of meeting during my years at MIT. In this book they showed that the perceptrons were very limited since they could not learn the XOR function -because it was not linearly separable- and therefore they would not have the capacity to solve real-life problems. Artificial Intelligence researchers found limitations and insurmountable difficulties in the 1970s, including very limited computing capabilities that made it impossible to process large amounts of data to train the complex models which were necessary to address real-life problems. It has not been until about a decade ago that we have been able to address this limitation.

In the early 1980s, the interest and funding towards Artificial Intelligence began to increase. In fact, the first expert systems were adopted commercially in the 1980s, such that in 1985 the market for AI systems in companies reached billions of dollars. In 1984 Cyc was born, the first scientific effort to implement common reasoning in a machine, creating a massive database that was to contain all the knowledge about the world that an average person would have.

However, in the same year (1984) and during the conference of the American Association of Artificial Intelligence (AAAI), Minsky and Schank warned that the enthusiasm and investment towards Artificial Intelligence would lead to disappointment, and that the field of AI began its second winter in 1987, reaching its darkest moment in 1990.

However, the scientific community continued to make progress in both schools of thought. One of the most important milestones of the *bottom-up* school (and particularly of *connectionism*, the school of AI based on the use of neural networks) was the proposal to use the *backpropagation* algorithm by Rumelhart, Hinton and Williams [11] in 1986 to generate useful internal representations of the input data in the hidden layers of neural networks (see an example in Figure 4). Although Rumelhart *et al.* were not the first to publish an article about backpropagation, their article stands out for the clarity with which they presented the idea, such that they managed to impact the scientific community. In fact, this algorithm is the basis for the vast majority of deep neural network models today and allows to train multi-layer networks from data. Other important work included that of Judea Pearl (corresponding member of this Academy) at the end of the 80s in which he incorporated probability theory and decision theory in Artificial Intelligence. Some of the proposed new methods include Bayesian networks, hidden Markov models (both key methods in my research), information theory, stochastic modeling and optimization. Evolutionary algorithms were also developed in this time period.

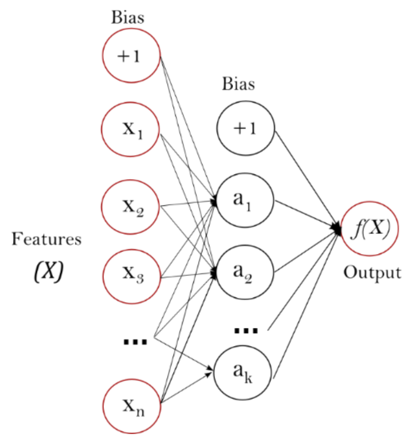


Fig. 2. Multi-layer perceptron with one hidden layer (source: scikit-learn)

Since the mid-90s, precisely when I started my PhD at MIT, until today, and especially for the last 10 years, there has been a very significant advance in data-driven statistical machine learning techniques that belong to the *bottom-up* approach. My scientific contributions in this field consist of the development and use of new dynamic graphical models to model different human behaviors, as illustrated in Figure 5. I started using hidden Markov models (HMMs) to recognize real-time facial expressions in a system called LAFTER which was named an outstanding publication by the Pattern Recognition magazine in the year 2000 [12] [1]. I proposed a new architecture of hidden Markov models called *coupled hidden Markov models* (coupled HMMs or CHMMs) to automatically recognize human interactions from videos [3] [13]. This work has been cited extensively (with more than 4000 citations) and included in university bibliographies of computer vision and machine learning. Another innovative graphical model that I developed are *layered hidden Markov models* (layered HMMs), capable of learning hierarchical models from data from different sensors, that is, in multi-sensory systems. I illustrated this new method in SEER, a multimodal system of real-time recognition of office activities from video, microphone and keyboard and computer mouse data. One of the publications resulting from this project received the 10-year impact award at the ACM ICMI conference [2] and we demonstrated SEER together with Bill Gates during his keynote at IJCAI 2001 in Seattle to thousands of people.

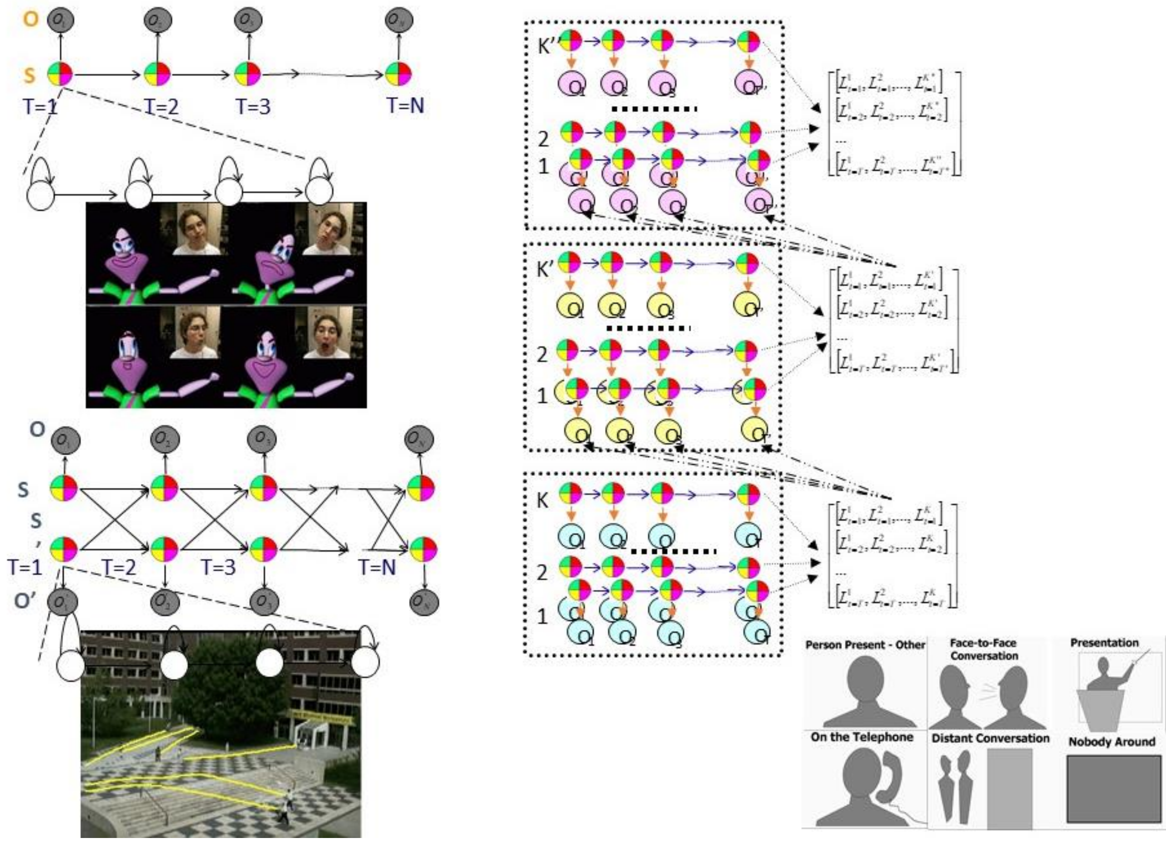


Fig. 5. Different dynamic graphical models to recognize facial expressions (top left), human interactions (bottom left) and office activities (right).

The availability of massive amounts of data -Big Data- as a result of human activity through the use of digital services and digitalization processes of the physical world and very powerful processors at low cost, together with the development of deep and complex neural network architectures -deep learning models [14], which are an evolution of the basic Perceptron with many hidden layers and different architectures, *e.g.* figures 6 y 7- trained with such data, has led to Artificial Intelligence living today a *"perpetual spring"* –as Stanford's Professor Andrew Ng –whom I also met at MIT-- likes to refer to. In the past 6-8 years there has been a strong come back of the bottom-up approach and particularly connectionist approaches within AI.

The interest towards Artificial Intelligence has experienced exorbitant levels, partly fueled by the achievement of important milestones that only three decades ago seemed unattainable and that have received considerable media attention, among which I would highlight: the victory of Deep Blue to Gary Kasparov, world chess champion, in 1997; the automatic driving of 131 miles on a desert road in 2005 by the autonomous vehicle developed by Stanford as part of the DARPA Grand Challenge; Watson's victory in the game of Jeopardy! Against two world Jeopardy! champions, Brad Rutter and Ken Jennings, in 2011; the development since 2011 of personal assistants in mobile phones (*e.g.* Siri, Cortana and Google Now) and since 2015 for the home (*e.g.* Alexa, Google Home) that allow their users to use voice and natural language to ask questions and request them to carry out actions automatically; the 2016 victory of DeepMind's AlphaGo against Lee Sedol, one of the best Go players in the world; the victory playing poker by the AI Libratus program against his 4 human opponents -placed among the best players in the world- with an impressive statistic of victories; AlphaZero's ability not only to beat the best chess program in the world, but to learn how to play chess by itself and the superiority of an system developed by Alibaba over human performance in Stanford’s reading comprehension test on a set of one hundred thousand questions, both in 2018.

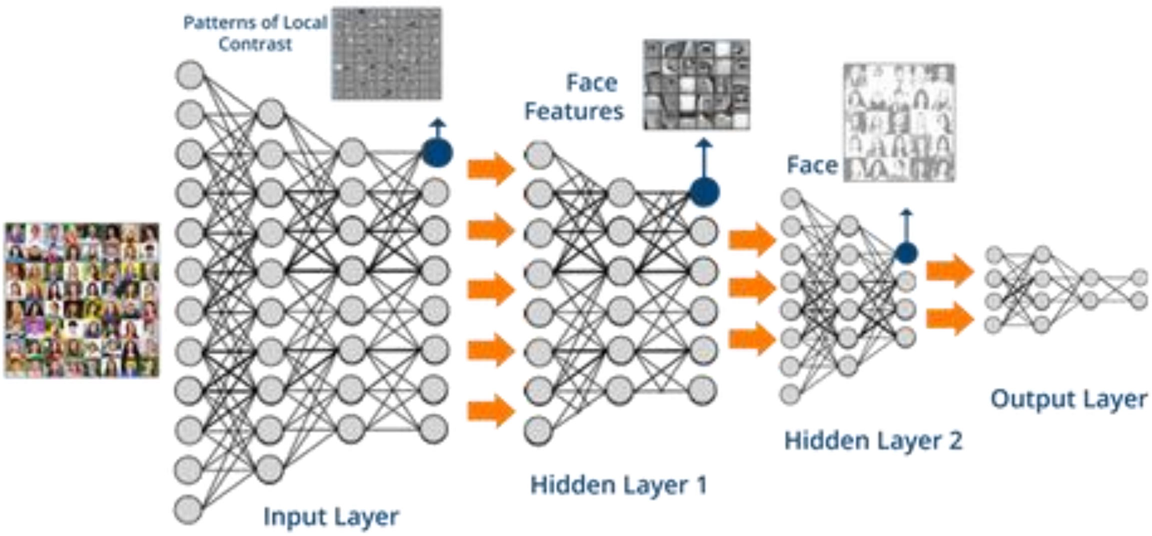


Fig. 6. Example of deep neural network to analyze faces in images (source: https://cdn.edureka.co/blog/wp-content/uploads/2017/05/Deep-Neural-Network-What-is-Deep-Learning-Edureka.png)

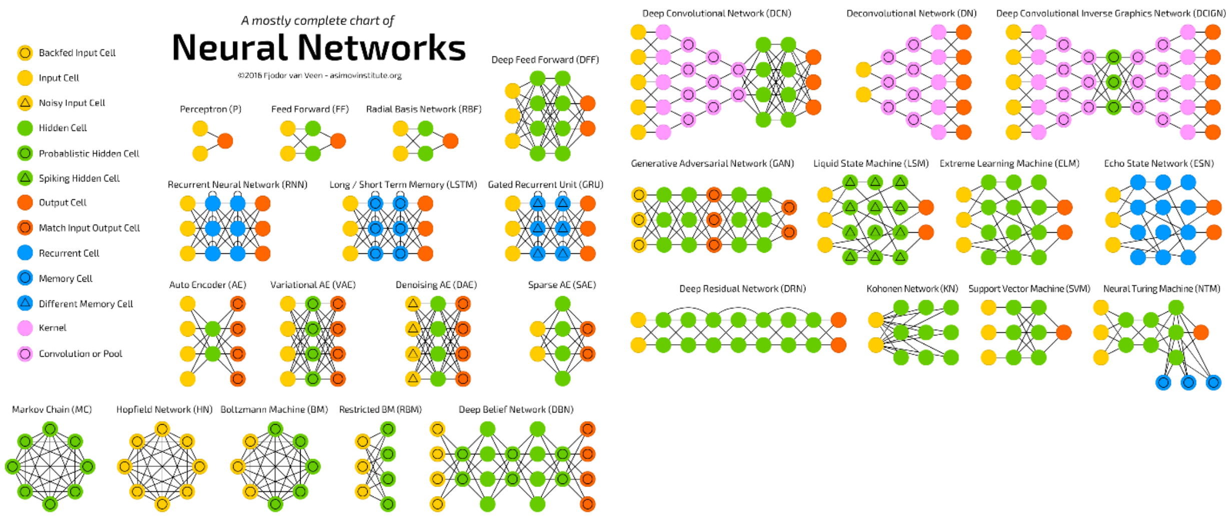


Fig. 7. Examples of different architectures of deep neural networks (source: asimovinstitute.org Fjodor van Veen)

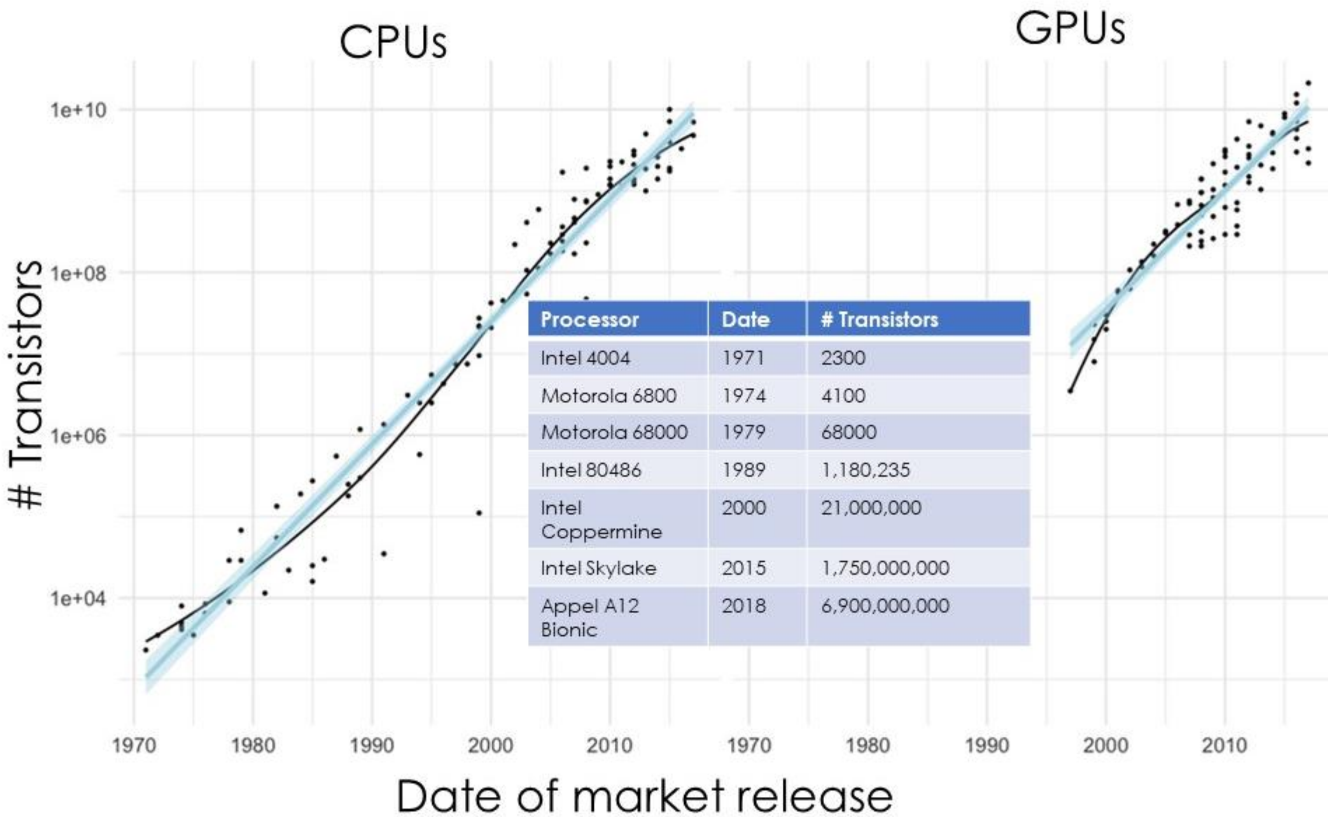
Beyond these milestones, Artificial Intelligence already occupies an important place in many spheres of our life. It is present in the information/content/products/friends search and recommendation systems that we use in our day to day (*e.g.* Netflix, Spotify, Facebook, Twitter, news services), in the mobile camera applications that automatically detect the faces in photos, in our personal assistants (*e.g.* Siri, Cortana, Google Now, Alexa or Google Home, conversational chatbots) and in smart cities to, for example, predict traffic. It is also part of decision-making systems in financial markets and in companies (*e.g.*, stock trading, credit allocation, insurance, rate setting), in medical solutions (*e.g.*, automatic diagnostic systems, automatic processing of clinical histories and DNA and automatic analysis of medical images), in decision-making systems of Public Administrations (*e.g.*, surveillance systems, support for judicial decisions or automatic classification and prioritization of students), in production and manufacture processes (*e.g.*, industrial robots, planning systems and prediction of demand/production ...), in autonomous vehicles, in weather forecasting and physical models, and in security and defense systems (*e.g.*, autonomous weapons, border control and granting of visas).

Throughout my professional career, I have contributed to personalization and recommender systems with models that include context [15], that optimize a ranked list of results [16] or that include expert opinions [17]. I have also developed projects such as *MobiScore*, a system to automatically infer credit scoring in developing economies from mobile phone usage data [6], projects to automatically detect people [18] or crime [7] *hotspots* in cities through the analysis of aggregate mobile data, and a pioneering method to automatically segment the land in a city by analyzing the patterns of use of shared bicycling systems [19].

We cannot ignore the contribution of the development of novel hardware systems -and in particular large-scale, distributed and parallel storage and processing systems- for the development and practical application of AI. We would not be able to train complex deep learning models using massive amounts of data with a reasonable energy consumption and in a reasonable amount of time without the existence of new processors, which in many cases are optimized for deep learning algorithms. In the last years we have moved from using general purpose processors (CPUs and GPUs or graphics processing units) to the use of specialized processors which are optimized for this type of models (FPGAs or field-programmable gate arrays and ASICs or application-specific integrated circuits like the TPU developed by Google).

Figures 8 and 9 illustrate the evolution in computing capacity. Figure 8 shows the famous Moore's Law since 1971, according to which the number of transistors that we can integrate into a circuit doubles every 1-1.5 years. On the left side of the graph, we have Moore's Law for CPUs and on the right for GPUs. Figure 8 illustrates where the different types of processors are located according to their flexibility and ease of use versus their energy efficiency and performance. The interested reader can find a summary of the main processors used for machine learning in [20].

Finally, we have to consider the energy consumption necessary to transmit and process the huge amounts of data that exist today. In a recent study [21], Andrae estimates that if we do not achieve a dramatic improvement in the energy efficiency of the processors used, the IT industry could consume 20% of all the planet's electricity and emit up to 5.5% of the planet's $CO2$ emissions by the year 2025. Prediction of electricity demand would go from 200-300 TWh of electricity per year to 1,200-3,000 TWh in 2025. Data centers could produce 1.9Gt (or 3.2% of the total) of the planet's carbon emissions.



Processor	Date	# Transistors
Intel 4004	1971	2300
Motorola 6800	1974	4100
Motorola 68000	1979	68000
Intel 80486	1989	1,180,235
Intel Coppermine	2000	21,000,000
Intel Skylake	2015	1,750,000,000
Appel A12 Bionic	2018	6,900,000,000

Fig 8. Moore's law since 1971 for CPUs (left) and GPUs (right).
(source: Graphs made by http://coulmont.com from Wikipedia data)

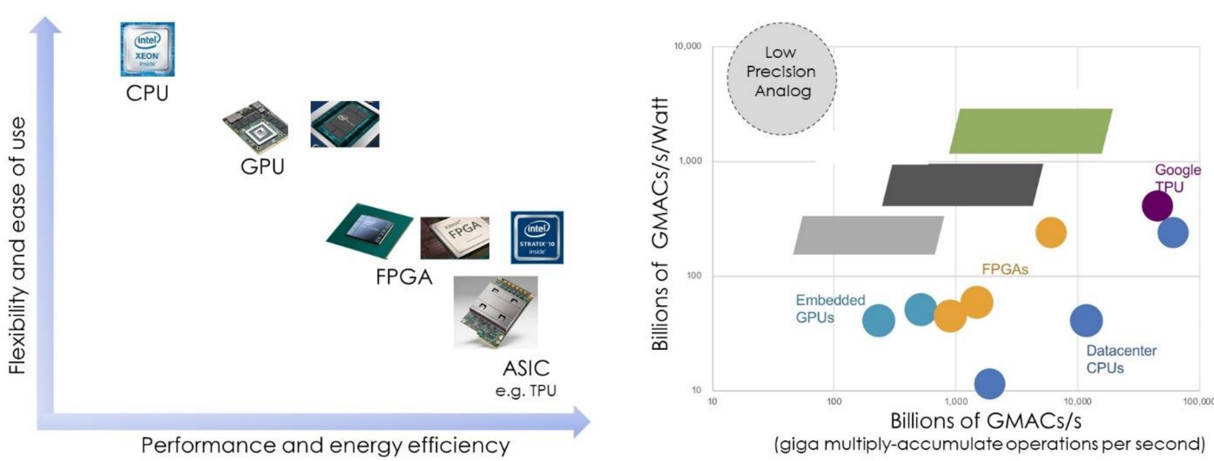


Fig 9. Left. Different types of processors used in deep learning systems; Right, Performance vs efficiency of different types of processors (source: Chris Rowen, Cognite Ventures)

Types of Artificial Intelligence

Regarding their level of competence, Artificial Intelligence systems are usually divided into three types:

1. *Narrow AI* systems, which is the type of Artificial Intelligence we have today. These are systems capable of performing a specific task (for example, recognizing speech, recognizing images, processing text ...) even better than a human, but only that task;

2. *General AI* systems, which would be the aspiration of Artificial Intelligence since these are systems that exhibit an intelligence like human intelligence: multiple, adaptable, flexible, efficient, incremental ... We are far from reaching this type of systems;

3. *Super-intelligence*, which is a somewhat controversial term referring to the development of systems that would have an intelligence superior to that of a human, as proposed by the British philosopher Bostrom [22].

Today we have narrow AI systems, that is, systems that are able of automatically and autonomously performing a specific task, but only that task. For example, a chess playing algorithm might play chess better than the best of humans, but such a system is unable to do any other task. In fact, he does not "know" what chess is and it would have difficulties to play if we made a variation to the rules of the game. Therefore, current systems manifest a limited type of intelligence because they are unable, among other things, to generalize and extend their levels of competence in a certain task to other areas automatically, as a human would.

Artificial Intelligence vs Human Intelligence

The recent success of Artificial Intelligence systems -as I have illustrated previously- may be diverting our attention from foundational problems of current Artificial Intelligence methods that are still to be solved. Among others: (a) the need to incorporate semantics and reasoning in natural language processing systems; (b) the need to develop representations of uncertainty that are computationally tractable; (c) the importance of developing systems that can formulate, adapt and pursue long-term objectives; (d) the importance of representing and inferring causality; (e) the need to incorporate context information in the models; (f) the capacity to learn constantly, incrementally and associatively; and (g) the importance that the models are endowed with robustness, so that they do not fail miserably when certain characteristics are changed in *e.g.* the input data, as it happens today.

Some of these problems derive from three basic limitations in today's machine learning mechanisms that differentiate them from intelligent biological systems, as Jeff Hawkins explains in an article for IEEE Spectrum [23] and which I summarize next.

First, biological learning systems (brains) are capable of learning quickly -a couple of observations or tactile experiences are usually enough to learn something new, unlike the millions of examples that are needed by today's Artificial Intelligence systems. Also, biological systems learn incrementally, *i.e.*, we can add new knowledge without having to relearn everything from scratch or lose previous knowledge, and we can learn continuously, that is, we never stop learning as we interact with the physical world that we live in. Rapid, incremental and constant learning is an essential element that enables intelligent biological systems to adapt to a changing environment and survive. The neuron is a key element in biological learning and the complexity of biological neurons and their connections is what gives the brain the ability to learn. Today we know that the brain has plasticity and that new neurons (neurogenesis) and synapses (synaptogenesis) are constantly being created. In fact, up to 40% of the synapses in a neuron could be replaced every day with new synapses, which give rise to new connections between neurons. Artificial learning systems would not have to reproduce exactly how biological neurons work, but this ability to exhibit rapid, incremental and constant learning, characterized by the destruction and creation of synapses is essential.

Second, the brain uses what are known as sparse distributed representations (SDRs) [24]. They are called sparse because only a small set of neurons is active at each moment of time. Which neurons are active changes from one instant to another, depending on what the living being is doing, but the set of neurons that is active is small. This type of representation is robust to errors and noise. In addition, they have two interesting properties: the overlap property, which allows to quickly detect if two perceptions are identical or different; and the union property, which allows the brain to maintain several representations in parallel. For example, if we feel that there is an animal hiding in some bushes, but we have not been able to see it clearly, it could be a rabbit, a squirrel or a rat. Since the representations in the brain are not very dense, our brain can activate three SDRs at the same time (that of the rabbit, the squirrel and the rat) without having any interference between them. This property of being able to unite SDRs is what allows the brain to manage, operate and make decisions with uncertainty.

Third, learning is embodied, so that our brain receives information from different senses, information that changes as we move and act in our environment. An Artificial Intelligence system does not have to have a physical body, but it does have the ability to act on its environment (physical and virtual) and receive feedback based on its actions. Reinforcement learning systems do something similar and they have been instrumental in the achievement of some of the milestones previously mentioned, such as in the case of AlphaZero. In addition, the brain can integrate the information captured by the different senses and the motor system, in order to not only process, recognize and decide based on what is perceived, but also act. This sensory-motor integration is basic in the functioning of the brain and probably should also be of the Artificial Intelligence systems.

Finally, in the brain, sensory information is processed by a hierarchical system so that as information is passed from one level to another, increasingly complex characteristics are being computed about what is being perceived. Deep-learning models also use hierarchies, but much deeper, with tens or hundreds of levels and billions of parameters, while the brain needs just a few levels. Also, as I said, deep neural networks need millions of observations to learn a pattern, while the brain needs few examples. The brain has a capacity to learn that is much more efficient than that of today's computer models.

However, despite these limitations, Artificial Intelligence is already having a great impact on society -as I have described previously- and is an integral part of the Fourth Industrial Revolution, in which we are immersed, as I explain below.

The Fourth Industrial Revolution

In the last three centuries, we have experienced four industrial revolutions, illustrated in Figure 10. The First Industrial Revolution took place between the eighteenth and nineteenth centuries in Europe and North America and corresponds to the historical moment in which societies that were mostly agrarian and rural they became industrial and urban. The main driver of this revolution was the invention of the steam engine, together with the development of the textile and metallurgical industries.

The Second Industrial Revolution -known as the Technological Revolution- occurred just before the First World War, between 1870 and 1914, and corresponds to a growth of the previous industries and the development of new industries such as the steel industry, oil and electricity. The most important technological advances of this revolution include the telephone, the light bulb, the phonograph and the internal combustion engine.

The Third Industrial Revolution is known as the Digital Revolution and refers to the transition from mechanical and analogue devices to the use of digital technologies. It started in the 80s and continues today. The key technological advances in this Third Industrial Revolution include personal computers, internet and the development of other information and communication technologies (ICTs).

Finally, the Fourth Industrial Revolution is based on advances in the Digital Revolution but incorporates the ubiquity of digital technology both in our society and in our bodies and the growing union between the physical and the digital worlds. The technological advances that make this new revolution possible include robotics, Artificial Intelligence, the availability of Big Data, nanotechnology, biotechnology, the internet of things, autonomous vehicles, 3D printers and quantum computing. The term was coined for the first time during the World Economic Forum in 2016 by its founder, Klaus Schwab.

Artificial Intelligence has a set of characteristics that contribute to make it one of the key elements in this Fourth Industrial Revolution, including:

(1) *Transversality and invisibility:* Artificial Intelligence techniques can be used in a myriad of use cases, including biology, physics, medicine, chemistry, energy, transport, education, production systems, logistics and transport, digital services and the provision of public and private services. In addition, the vast majority of Artificial Intelligence systems that are used today are invisible, that is, they consist of software at the heart of the intelligent systems and services that we use in our day to day life. These two properties -transversality and invisibility- position Artificial Intelligence at the heart of the Fourth Industrial Revolution, with a role like that played by electricity in the Second Industrial Revolution.

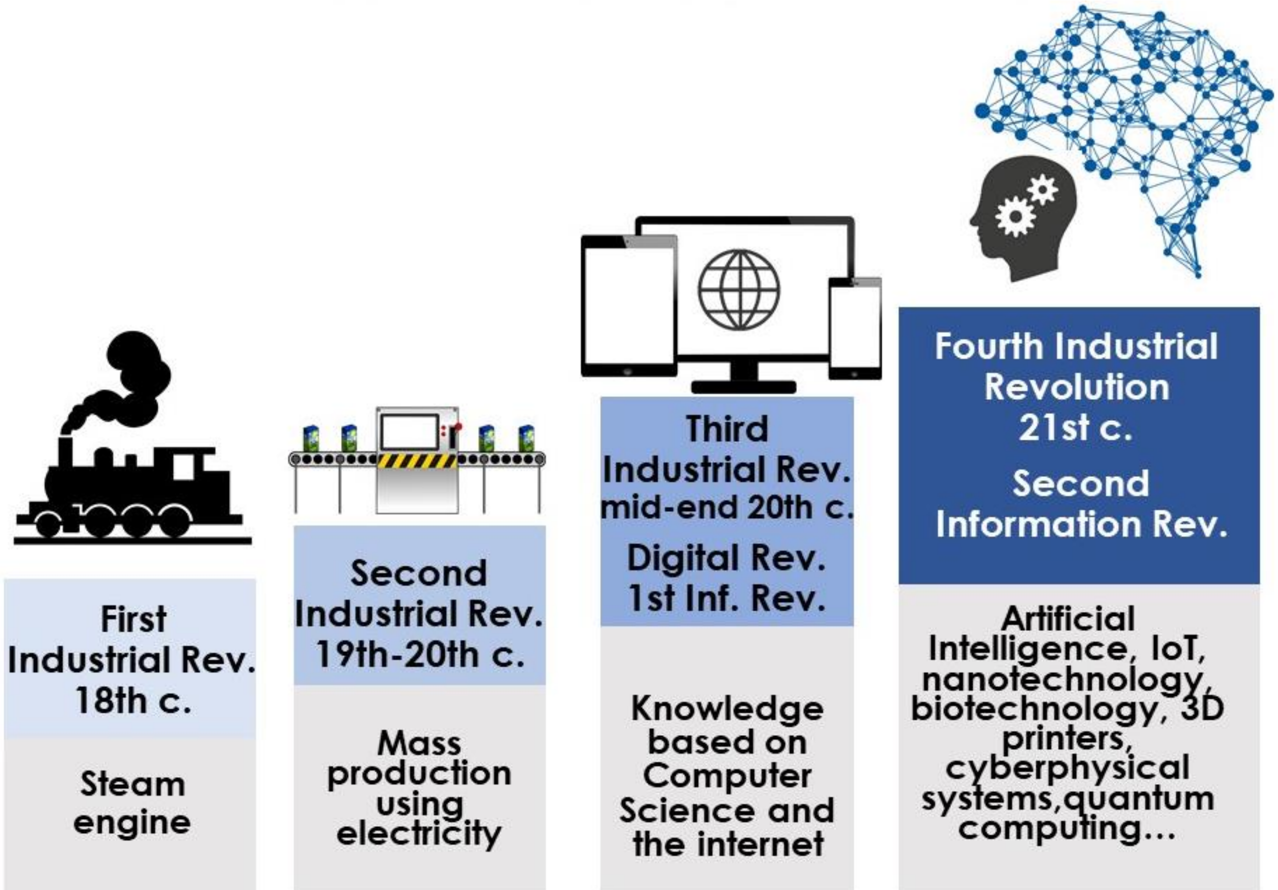


Fig 10. The four industrial revolutions since the 18th century

(2) *Complexity, scalability and constant actualization:* Current AI-based systems which use deep learning models are complex, with hundreds of layers and billions of parameters. This complexity hinders the ability to interpret the models which in certain use cases -such as in medicine or in education- might be a necessary condition to be able to use an AI system in the first place. At the same time, this complexity allows AI-based systems to process huge amounts of data - which would otherwise be unviable- and perform tasks with higher levels of competence than humans. That is, they give these systems great scalability. In most Big Data use cases we can only extract knowledge and value from such amounts of data by using Artificial Intelligence methods, since traditional methods do not scale to such large volumes of data which is also varied, unstructured and generated at high speed. Likewise, Artificial Intelligence systems are highly scalable as they consist of software, which, in addition, can be connected to thousands or millions of other AI systems, giving rise to a collective AI network. This scalability, combined with the ability to update the software in a massive way, would allow such a system to reach and have impact on hundreds or thousands of millions of people in a short period of time. We might imagine unprecedented scenarios in our history. Scenarios where, for example, a network of AI systems could very quickly incorporate the latest computational methods in the diagnosis of a disease and deploy them to the entire population of the planet. The analog equivalent, which would consist of an almost instantaneous incorporation of such knowledge in the brains of all the doctors of the planet, is unfeasible.

(3) *Ability to predict:* Artificial Intelligence systems can be used for automatic decision making and to predict future situations. In fact, the aspiration is that algorithmic decisions based on AI trained with data will overcome the limitations of human decisions (such as conflicts of interest, biases, self-interest, and corruption) and therefore will be more just and objective. However, this is not necessarily the case if we do not consider the limitations of data-driven algorithmic decisions, as I will explain later.

State-of-the-art machine learning methods need access to large amounts of data and computing, as well as experts in the field to design and train such systems. There is, unfortunately, a situation of asymmetry such that only a minority has access to said data, computing capacity, knowledge and experience, and therefore can fully benefit from Artificial Intelligence. Unfortunately, a majority lacks these capacities and therefore is a mere user of such technology. This characteristic poses an important challenge, since we should minimize the asymmetry to guarantee that the impact of Artificial Intelligence is positive in all of society and not only in a subset of it.

In addition, we are able today to generate synthetic content (photos, text, audio and videos) which indistinguishable from real content using machine learning methods. This ability to invent content is transforming human communication, information dissemination and public opinion. It gives unprecedented power to those who can generate and disseminate such artificial content [25]. Again, this characteristic of today's AI methods poses important challenges that we should address.

Despite these limitations, the presence of AI in our lives and its ability to have a positive impact on our society are undeniable. For this reason, the great world powers -both companies and governments- have understood that the leadership in Artificial Intelligence will lead not only to an economic but also a political and social leadership, given AI's central in the Fourth Industrial Revolution. In fact, in the last 2 years, the governments of around twenty countries[2] --including United States, China, Canada, France, Taiwan, Singapore, Mexico, Sweden, India, Australia and Finland-- have developed national strategies on Artificial Intelligence, as shown in Figure 11. At a European level, the European Commission published in April 2018 a communication on Artificial Intelligence[3] and it is expected that before the end of 2018 they publish a document drafting the European strategy on Artificial Intelligence. At a national level, I am a member of the committee of 9 experts appointed by the Secretary of State for Digital Advancement[4] with the task of writing a white paper on Big Data and AI, including strategic recommendations for the Spanish Government. The white paper should be finished before the summer of 2019.

[2] https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

[3] https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[4] Anteriormente llamada Secretaria de Estado de Agenda Digital

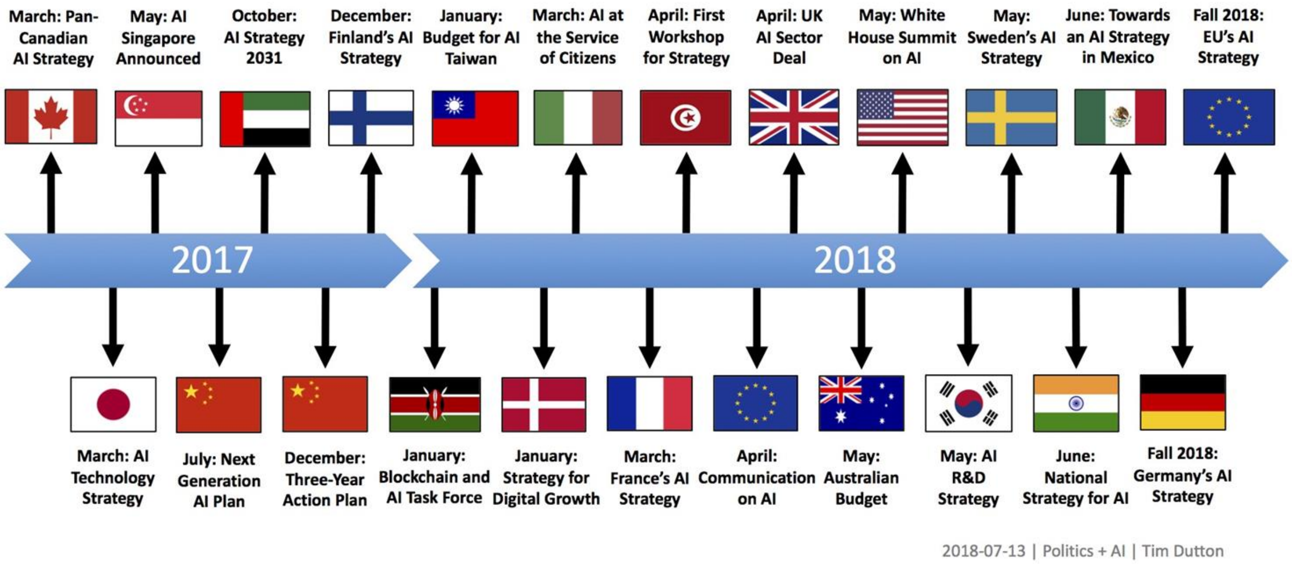


Fig. 11. Timeline of national strategies on Artificial Intelligence (source: Tim Dutton)

Because of some of the above characteristics, the impact of AI will not necessarily be distributed homogeneously or fairly in society. Consequently, we should urgently address the limitations of existing AI systems to ensure that it has a positive social impact. In the following sections, I will briefly describe four fundamental dimensions in the context of AI development: the labor, economic, social, and ethical dimensions.

Labor and economic impact of AI

The technological progress associated with the Fourth Industrial Revolution is polarizing the labor market. On the one hand, new high-paying jobs appear and require specialization in technological areas (such as, data science, ...). On the other hand, other types of work will disappear as they will be partially or totally automated (for example, taxi drivers / carriers, cashiers, travel agents ...). In general terms, any job that could be automated or replaced with the use of technology will be automated. As a result, labor demand is experiencing a bias in favor of specialized professional skills, and to the detriment of routine and mechanical skills or occupations. This trend predicts a complete change in the occupational structure that probably entails risks for society if we are not able to adapt to this change.

In Spain, the Organization for Economic Cooperation and Development (OECD) accounts for 12% of jobs that may be at risk due to automation [26].

At the same time, the development of disruptive technologies with the capacity to transform society has historically led to the generation of employment. According to a McKinsey study[5], one third of the new jobs created in the US in the last 25 years belong to disciplines that did not exist before, in areas such as information technology, hardware manufacturing, mobile applications or the

[5] https://www.mckinsey.com/featured-insights/employment-and-growth/technology-jobs-and-the-future-of-work#section 1

management of technological systems. In the next decade, employment will be concentrated in qualified functions with greater contribution of value, reaching unemployment rates of less than 3.5% for these profiles, compared to a 20% unemployment rate for professions that require a low qualification. Worldwide, a recent study by the World Economic Forum[6] estimates that there will be a net growth of 58 million jobs in 2011 as a result of the IA.

The European Commission[7] anticipates a need for more than 900,000 new technological jobs in the short term that we will be unable to cover if we do not transform our educational programs.

In the national context, the EPYCE 2017[8] report shows that the most demanded profile in Spain today is computer engineer, followed by profiles related to Big Data. In addition, the report predicts that 49% of the professions which will be in highest demand in the future in Spain will be in the fields of engineering and technology. However, are we prepared as a society to supply such labor demand for new jobs derived from AI? I do not think so.

Therefore, it is of vital importance that we invest in the training of professionals whose work will be affected by the development of Artificial Intelligence so that they can continue to contribute to society in new jobs. We are progressing towards a model of lifelong learning and the disappearance of a single professional career during a person's working life. Thanks to technology we have online training systems (for example, mass online open courses or MOOCs) that offer, in a scalable and economical way, learning opportunities to anyone from anywhere. We should ensure that professionals have the opportunity to learn emerging technologies in their fields of activity and thus continue to be relevant in their areas of competence, including -and especially- if these areas are affected by automation.

This need for constant learning, because of the constant change due to technological progress, might be difficult to manage from an emotional point of view. As we all know, humans tend to be reluctant to change, especially as we get older. Therefore, we should contemplate the possibility of the existence of a social collective which will be unable to adapt to this constant need for learning and therefore will lack the necessary tools to contribute to the society of tomorrow. In the words of the historian Yuval Noah Harari, a "useless class" [27]. Some of the potential solutions that have been proposed to address this challenge are the creation of a universal basic salary or the universal provision by governments of basic needs free of charge.

From an economic point of view, in 2018 the market for products, hardware and software related to Artificial Intelligence is expected to exceed 4 billion dollars, according to a study by Statista[9] (see Figure 12). With a sustained growth over time, a macroeconomic study by PwC[10] estimates that Artificial Intelligence will generate more than 15 billion dollars worldwide in the year 2030, with impact on all areas of activity, both in the public and private sectors. This economic impact is due both to the increase in productivity -as a result of automation processes and assistance to the

[6] http://reports.weforum.org/future-of-jobs-2018/?doing_wp_cron=1537554858.4296801090240478515625
[7] https://ec.europa.eu/commission/commissioners/2014-2019/ansip/blog/digital-skills-jobs-and-need-get-more-europeans-online_en
[8] http://marketing.eae.es/prensa/_EPyCE2017.pdf
[9] https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/
[10] https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf

workforce with Artificial Intelligence systems- and to a growing demand from consumers for products and services enriched with IA. Geographically, North America and China will experience the greatest benefits given the concentration of research, innovation and development of Artificial Intelligence in these two regions.

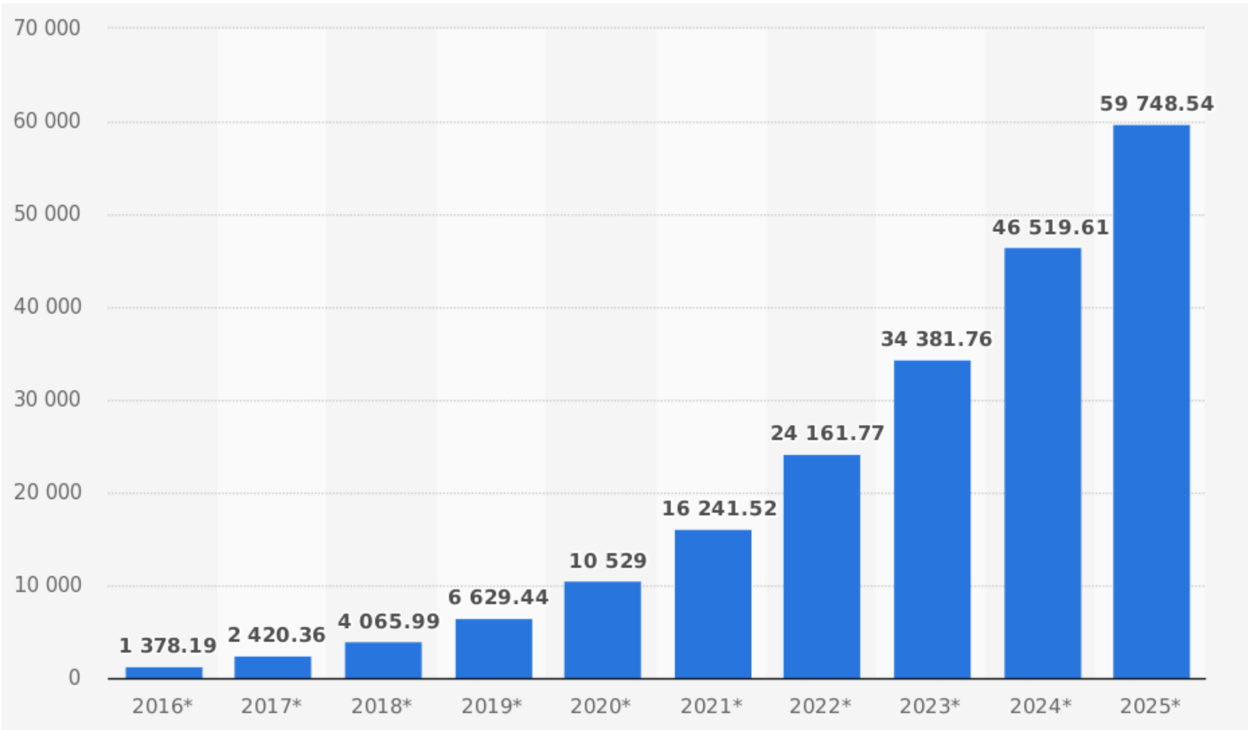


Fig 12. AI market size in millions of US dollars (source: Statista)

With respect to AI innovation, a recent study by Asgard and Roland Berger[11] analyzes the distribution of AI startups worldwide. USA is the world leader with 40% of all IA startups analyzed by this report (~1400 AI startups, of which ~600 are in San Francisco), followed by China and Israel. At a European level, London is the second city in the world in number of AI startups, followed by Paris in tenth position. Finally, a report by Accenture and Frontier Economics[12] estimates the economic growth that different countries (including Spain) could have if they invested in AI and managed to incorporate their benefits into the economy, compared to the basic economic growth, without the use of AI. In the case of Spain, the prediction is for a GDP growth of 0.8% in the year 2035 thanks to AI.

In the Spanish context, I would highlight the opportunities that exist in health, transportation, energy, agriculture, tourism, e-commerce, banking and public administration.

[11] "The global AI landscape" https://asgard.vc/global-ai/

[12] https://www.accenture.com/us-en/insight-ai-industry-growth

The potential of AI to improve society

Society should be enriched by the development and deployment of Artificial Intelligence techniques. Undoubtedly, Artificial Intelligence is contributing and will contribute to economic growth, as I previously explained. It will allow us to have a precision medicine (personalized, preventive and predictive medicine), a personalized and permanent education, smart cities, a more efficient management of resources and a fairer, more transparent and evidence-based decision making. Likewise, this impact will not be exempt from profound social changes, including a transformation of the labor market, as described in the previous section.

The development and implementation of human-centric Artificial Intelligence should result in an empowerment of society. A necessary condition for this empowerment is knowledge.

Hence, we would need to invest in both formal and informal education. Otherwise, it will be very difficult, if not impossible, for us as a society to make decisions about technologies that we do not understand and that, consequently, we often fear. I fully agree with Marie Curie's words, *"nothing in life should be feared, but understood. Now is the time to understand more in order to fear less".*

From a formal point of view, many countries in the world -among which unfortunately is not Spain-have included a core subject in Computational Thinking[13] in their compulsory education (primary and secondary) curricula. Computational Thinking [28] includes five core competencies: algorithms, data, networks, programming and hardware.

In the book entitled "*Los nativos digitales no existen*" (Digital Natives Don't Exist) [29] I wrote a chapter called "*Eruditos digitales*" (digital erudites) which emphasizes the need to teach Computational Thinking in as part of the normal curriculum in primary and secondary education. It also highlights the need to develop critical thinking and social, emotional and creative skills that we are not currently developing and that will increasingly be important for our mental health and our peaceful and harmonious coexistence with technology, other humans and our planet. Among others, I believe in the importance of developing the ability to know how to make decisions in a changing environment and with uncertainty; to nurture the necessary mental and emotional stabilities to re-invent, learn and adapt to change; the develop the abilities to be bored, thus fostering creativity; to concentrate on a single task, without interruptions, during sustained periods of time; and to discern between what is important and what is not, verifying the veracity of the sources and know how to form an opinion from a multitude of data sources; to culture a tolerance towards other points of view; and the necessary discipline to endow our biological body with sufficient rest hours, without stimuli or interruptions.

From my experience, the knowledge of the average person in relation to technological concepts is very limited. Therefore, I believe we should do more science and technology outreach efforts. It could be effective to develop communication and awareness campaigns in various media (social networks, newspapers, radio, TV etc.) to give educate in technological concepts and inspire new generations, and especially girls, to study technological careers. Unfortunately, the percentage of women who study computer science or who work in technical positions within technological

[13] https://ec.europa.eu/jrc/en/computational-thinking

companies is between 10 and 20%[14]. This lack of gender diversity is alarming, especially because the percentages of women in these disciplines have decreased progressively since the 1980s. This decline in the percentage of female students in computer science is the result of a set of factors, among which I would highlight:

(1) an erroneous and strongly stereotyped image about who works in technology and what the work consists of. The image -reinforced by movies, television shows and the media- is that of a geek, a guy with glasses, with very low levels of social and emotional intelligence and little personal hygiene, working in a windowless basement, surrounded of "junk food", typing in front of one or more computer screens. This image is, without doubt, very unattractive for girls and also far from reality. In addition, we live in a world with strong gender stereotyping in the toys, books, clothes and films / series consumed by our children and adolescents;

(2) gender biases -conscious and subconscious- that both men and women have and that imply a systematic undervaluation of women against male counterparts with identical qualifications [30];

(3) a systematic lack of recognition towards women in all contexts and particularly in technological careers, illustrated by the salary gap or the poor presence of women in prizes, distinctions and positions of power in this field -for example, the Turing Award (equivalent to the Nobel Prize in Computer Science) has been awarded only to 3 women since its creation in 1966. In addition, women founders of startups received only 2% of US venture capital investments in 2017[15] and the female founder teams of startups attract an average of $82 of investment for every $100 of investment obtained by men's teams [16];

(4) the lack of feminine references in this field that can inspire girls and adolescents to study these careers; and

(5) a markedly sexist and misogynist culture, called *brogrammer*[17] culture, which in a recent study published by *Fortune*[18] magazine was the second most frequent reason -cited by 68% of women- causing them to leave their work in the technology sector.

In order to encourage the presence of women in this field, dozens of support, training and awareness initiatives have emerged -both nationally and internationally. In this house, we have the Women and Engineering initiative, led by Prof. Sara Gómez, whose objective is to support girls who are studying engineering through tutoring networks. Other national examples include the *Wisibizalas* initiative of the UPF, a contest aimed at schools/highschools to spur a dialogue about the (lack of) visibility of women, and to give visibility to women working in technology; the Association of Women Researchers and Technologists (AMIT), whose objective is to promote gender equality; *MujeresTech*, a non-profit association that offers resources and knowledge with the aim of increasing the presence of women in the digital sector; the *Top100*, an initiative dedicated to

---

[14] https://smallbiztrends.com/2018/03/women-in-technology-statistics.html
[15] http://fortune.com/2018/01/31/female-founders-venture-capital-2017/
[16] https://techcrunch.com/2018/01/15/the-portion-of-vc-backed-startups-founded-by-women-stays-stubbornly-stagnant/?guccounter=1
[17] https://en.wikipedia.org/wiki/Brogrammer
[18] http://fortune.com/2014/10/02/women-leave-tech-culture/

identifying the 10 most influential women in Spain in 10 categories. Although it does not have a technological category per se, it includes a category of entrepreneurs and another category of academics/researchers; and the University of Deusto's Ada Byron Prize to the Technological Woman of the Year, which I had the honor of receiving in 2016, whose objective is to reward and make visible excellent trajectories of women in various technological fields.

These initiatives are necessary to encourage greater diversity in technology. However, it may be necessary to join efforts so that the initiatives have greater impact and visibility. The introduction of Computational Thinking as a core subject would also contribute to reducing the gender gap in technological careers since all students (boys and girls alike) would naturally acquire technological skills, breaking down the stereotype of gender associated with technology.

Another dimension for social empowerment through Artificial Intelligence entails the generation of collaboration spaces (urban laboratories or living labs) based on data (sharing data), technology (sharing infrastructure) and skills (sharing talent) to contribute to the progress of Artificial Intelligence and democratize its access, balancing inequities and the asymmetry to which I referred earlier. Some examples include the Mobile Territorial Lab [31] in Trento -with whom I collaborated when I was Scientific Director in Telefonica- and the Urban Laboratory in Bogotá -supported by the NGO Data-Pop Alliance, where I am the chief data scientist. In the Mobile Territorial Lab I led a research project whose objective was to understand, from a human-centric perspective, the monetary value that users assign to the data captured by their mobile phones. This project received a best paper award in the *ACM Ubicomp 2014* international conference [32], as well as great interest by the national and international media, being the first research work that explored this issue.

Social inequality comes hand in hand with the existence of property. When this property is not evenly distributed in society, inequality arises. The twenty-first century is characterized by an increase in inequality. Today, according to a recent study by Credit Suisse[19], the richest 1% of the planet owns half of the world's wealth and the 100 richest people in the world have more than the 4 billion poorest people in the world. This accumulation of wealth in the hands of very few has been partially attributed to technological development and the Fourth Industrial Revolution. With the Agrarian Revolution in the Neolithic and for thousands of years, ownership of the land entailed wealth. In the Industrial Revolution, wealth became linked to the ownership of factories and machines. Nowadays, we could argue that data and the ability to make sense of it are the assets that generate the most wealth, giving rise to what is known as the data economy. Therefore, if we want to maximize the positive impact of AI on society, given that a significant number of AI systems need data, new models of ownership, management and regulation of data are needed. The European General Regulation for the Protection of Data (RGDP) is an example in this direction. However, the complexity in its implementation and practical application shows the difficulty to define and implement the concept of "property" when we speak of an intangible, distributed, varied, growing, dynamic and replicable infinite times at virtually zero cost.

Despite the difficulties, in the last 3-5 years different initiatives have emerged worldwide to promote Big Data and Artificial Intelligence for Social Good, one of my research areas for

[19] https://www.credit-suisse.com/corporate/en/research/research-institute/global-wealth-report.html

approximately the past 10 years, and thus contribute to democratization in the access and use of data, including:

--*New Deal on Data*, led by Professor Sandy Pentland - my thesis director - from the World Economic Forum, focused on consensus policies and initiatives for citizens have control over the possession, use and distribution of their personal data.

--Flowminder[20], is an NGO based in Sweden -with whom I have collaborated for years- with a decade of experience in Big Data analysis projects for the social good, with success cases in the use of Big Data to, for example, quantify the displacements and human settlements following the earthquakes in Haiti and Nepal.

--United Nations: the *World Data Forum* and the *Global Partnership for Sustainable Development Data*. It is an alliance -led by United Nations- for the achievement of the 17 Sustainable Development Goals (SDGs) through data analysis, with more than 150 collaborators representing a broad spectrum of data producers and users, including governments, companies, universities, NGOs, civil society groups, foundations, statistical offices, etc. Data and Artificial Intelligence can be used both to measure compliance with the 17 SDGs and to help achieve them and improve decision-making, for example, in public policies. The collaboration between the different organizations involved is key to the development of the established goals. Such collaboration will only be possible if all participating entities (governments, international organizations, companies, universities, civil society, etc.) commit to working together taking advantage of the potential provided by the data and involving citizens. In United Nations there is also a unit since 2009 dedicated to the analysis of data using Artificial Intelligence techniques for Social Good called *United Nations Global Pulse*[21], with whom I have also collaborated. In 2014 we carried out a study where we demonstrated the value of the aggregated and anonymized data of the mobile phone network, combined with satellite image data, for the detection of areas affected by floods in Mexico [33].

--OPAL[22] is a project led by Data-Pop Alliance -where I am Chief Data Scientist- in collaboration with various public-private stakeholders, with the goal of taking advantage of Big Data and Artificial Intelligence for social good while preserving the privacy of people, in a sustainable, scalable, stable and commercially viable way. The OPAL project proposes that algorithms are executed where the data is held and not vice versa, democratizing the access to said data and the knowledge derived from them.

-- There are numerous initiatives in private companies whose objective is to promote, apply and develop projects that use Big Data and Artificial Intelligence for social good. Among others, Telefónica's LUCA-Big Data for Social Good - which I had the honor of creating and promoting - whose objective is to use Big Data to have positive social impact; *BBVA Data & Analytics* for social good, by BBVA; Vodafone Big Data and AI for social good -which I have had the honor of creating and promoting- with projects in public health, financial inclusion, transport and official statistics in

[20] http://www.flowminder.org/
[21] https://www.unglobalpulse.org/
[22] https://www.opalproject.org/

Africa and Europe; Telenor *Big Data for Social Good*, with public health projects in Bangladesh and Pakistan; *Orange Data for Development*, two pioneering challenges where Orange shared aggregate and anonymized mobile data from Senegal and Côte d'Ivoire with hundreds of international teams, focused on use cases for social good; and *Turkcell Data for Refugees Challenge*, in which the telephone operator Turkcell has shared aggregate and anonymous data to help address the refugee crisis.

-- *Partnership on AI*[23] – It is a non-profit organization created in 2016 by Microsoft, Facebook, Amazon, IBM, Google and Apple with more than 50 public and private organizations affiliated to it. Its goal is to study and establish best practices regarding Artificial Intelligence, advance knowledge about Artificial Intelligence by citizens, and act as an open platform for the debate on Artificial Intelligence and its impact on people and society.

--GSMA – *Big Mobile Data for Social Good*[24], led by the GSMA and the United Nations Foundation in which 20 mobile operators –including Vodafone-- participate to contribute through the analysis of aggregate and anonymous mobile data to solve problems in the areas of public health and climate change /natural disasters.

--*AI for Good Global Summit de la ITU*[25], international summit of United Nations for the dialogue on Artificial Intelligence, aimed at identifying the practical applications of AI for the improvement of the sustainability of the planet. It is managed by the ITU (International Telecommunications Union) as a specialized agency of United Nations for Information and Communication Technologies.

--The *Center for Humane Technology*[26] is a multidisciplinary center recently created in California that advocates for the development of *humane* technology, that is, technology that has been designed taking into account human values, needs and interests above all.

Ethical principles needed for the development of a human(ity)-centered Artificial Intelligence

Throughout history, the development of disruptive technologies has had a profound impact on human life and relationships. Agriculture, printing, the steam engine, electricity or the internet are examples of technologies that have deeply affected the way we live, work and relate.

Artificial Intelligence is also transforming the world we live in. Today we have massive amounts of data that we can use to train Artificial Intelligence algorithms that allow researchers, companies, governments and other public sector actors to address complex problems.

Decisions with both individual and collective impact that were previously taken by humans -often experts- are nowadays taken by Artificial Intelligence systems (*i.e.*, algorithms) including decisions regarding the hiring of people, the granting of credits and loans, judicial judgments, medical treatments and diagnoses or the purchase-sale of shares in the stock market. Data-driven

[23] https://www.partnershiponai.org/
[24] https://www.gsma.com/betterfuture/bd4sg
[25] https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx
[26] http://humanetech.com/

algorithmic decisions have the potential to improve our decision making. In fact, history has shown that human decisions are not perfect [34]: they are subject to conflicts of interest, corruption, selfishness/greed and cognitive biases, which has resulted in unfair and /or inefficient processes and outcomes. Therefore, the interest towards the use of algorithms can be interpreted as the result of a demand for greater objectivity in decision making[27].

However, data-driven decision making is not exempt of limitations. Interestingly, Plato's words 2,400 years ago are surprisingly current today, *"A good decision is based on knowledge, not on numbers (data)".*

When these decisions affect thousands or millions of people, important ethical dilemmas arise. For example, does this mean that automatic decisions are beyond our control? What levels of security do these systems have to protect themselves from cyberattacks or malicious use? How can we guarantee that such decisions and/or actions do not have negative consequences for people? Who is responsible for these decisions? What will happen when an algorithm knows each one of us better than ourselves and can take advantage of that knowledge to manipulate our behavior subliminally?

Recently there have been numerous organizations and national or supra-national -for example, te European Union- movements for the definition of principles to establish global standards and regulation of this area. For example, in the Future of Life Institute AI conference in 2017 --with the participation of more than 1,200 international figures related to technological and scientific innovation, the Asilomar Principles for the development of Artificial Intelligence were defined with a total of 23 recommendations. Recently, the European Commission has appointed a High-Level Expert Group on Artificial Intelligence, with a focus on the ethical, legal and social implications of AI[28], where I am a reserve member. In Spain, I would highlight the Barcelona[29] declaration, promoted by the Artificial Intelligence Research Center of the Spanish Research Council, led by Dr. Ramon López de Mántaras, where they define 6 basic principles to develop an ethical AI.

Beyond preserving human rights[30], the existing literature has proposed a set of ethical principles and working dimensions that I believe will be necessary to address in order to get Artificial Intelligence to have a positive impact on society. The interested reader can find an extended version of some of these ethical principles in a recent article written with my collaborators of MIT, Data-Pop Alliance, Harvard and FBK, and published in the magazine "Philosophy and Technology" [34]. Next, I summarize the most relevant principles, grouped in 5 dimensions as per [35]:

Fairness

• *Justice and solidarity. Non discrimination.* Justice should be a central element in the development of automatic decision (and action) systems based on Artificial Intelligence. The decisions based on algorithms can discriminate because the data used to train said algorithms has biases that give rise to discriminatory decisions; because of the use of a certain algorithm; or by the misuse of certain models in different contexts. In addition, data-driven algorithmic decision making processes

[27] https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms
[28] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
[29] http://www.iiia.csic.es/barcelonadeclaration/
[30] https://en.wikipedia.org/wiki/Fundamental_rights

may imply that opportunities are denied to people not by their own actions, but by the actions of others with whom they share certain characteristics. For example, some credit card companies have reduced the credit limit of their customers not because of the customer's payment history, but as a result of analyzing the behavior of other customers with a history of poor payments who had purchased in the same establishments where the customer had purchased. Different solutions have been proposed in the literature to face algorithmic discrimination and maximize justice.[31]. However, I would like to underline the urgency for experts from different fields (including law, economics, ethics, computer science, philosophy and political science) to invent, evaluate and validate different metrics of justice for different tasks in the real world. In addition to this empirical research, it is necessary to propose a framework of theoretical modeling -supported by empirical evidence- that helps the users of these algorithms to make sure that the decisions made are as fairly as possible.

• *Cooperation.* Due to the transversal nature of Artificial Intelligence and its potential application to all areas, we should encourage and develop a constructive exchange of resources and knowledge between the private, public and social sectors to achieve its maximum potential of application and competitiveness. This need for cooperation not only between different sectors but also between nations -given today's globalization- has been emphasized by the well-known Israeli historian and thinker, Professor Yuval Noah Harari [27].

Autonomy and Accountability

• *Autonomy and dignity.* Autonomy is a central value in Western ethics, according to which each person should have the ability to decide their own thoughts and actions, thus ensuring free choice, and freedom of thought and action. However, today we can build computational models of our desires, needs, personality and behavior with the ability to influence our decisions and behaviors subliminally.

Therefore, we should ensure that autonomous intelligent systems always preserve human autonomy and dignity. For this, these systems need to behave in accordance with accepted ethical principles in the society where they are used. There are numerous examples of ethical principles as well as institutes and research centers created for this purpose, such as the AI Now Institute at New York University and the Digital Ethics Lab at the University of Oxford. However, this is an active area of research and there is no single method to embed ethical principles into data-driven algorithmic decision processes. It is also important to highlight that all developers and professionals working in the development of Artificial Intelligence systems that affect or interact with people (for example, algorithms for decision making, recommendation and personalization systems, chatbots ...) should behave in accordance with a clear Code of Conduct and Ethics defined by the organizations where they work. As Disney wisely said *"It is not difficult to make decisions when you know your values"*.

• *Accountability and the human role*. It is also important to be clear about the attribution of responsibility for the consequences of the actions or decisions of autonomous systems, in the same way that happens with the rest of the products used in society. Transparency is often considered a fundamental factor in contributing to accountability. However, transparency and audits are not enough to guarantee clear accountability. In fact, in a recent article, Kroll [36] has

[31] "Fairness in machine learning" https://fairmlclass.github.io/

proposed the use of computational methods to provide clarity regarding the attribution of responsibility, even when part of the information is hidden.

Finally, I think it is constructive to have a synergistic vision between humans and Artificial Intelligence systems. This view is often called *"intelligence augmentation"* so that Artificial Intelligence systems are used to increase or complement human intelligence. For example, an internet search engine can be considered a system to increase our intelligence, since it expands our knowledge with the capacity to process billions of documents and find the most relevant ones; or an automatic simultaneous translation system, since it allows people who do not speak the same language to be able to communicate.

Transparency

•*Transparency.* Transparency refers to the quality of being able to understand a computational model and therefore it can be a mechanism that contributes to the attribution of responsibility for the consequences of the use of said model. A model is transparent if a person can observe and understand it easily. Burrell [35] proposes three different types of opacity -*i.e.* lack of transparency- in algorithmic decisions: (1) *intentional opacity*, whose objective is the protection of the intellectual property of the inventors of the algorithms. This type of opacity could be mitigated with legislation that would force the use of open software systems. The new European General Data Protection Regulation (GDPR) with its right to an explanation is an example of this type of legislation. However, powerful commercial and governmental interests can make it difficult to eliminate this type of opacity; (2) *illiterate opacity*, due to the fact that the vast majority of people lack the technical skills to understand how algorithms and data-driven computational models work. This type of opacity would be attenuated with educational programs in digital competences -as I have explained previously- and by allowing independent experts to advise those affected by data-driven algorithmic decision-making processes; and (3) *intrinsic opacity*, which arises from the nature of certain machine learning methods (for example, deep learning models). This opacity is well known in the machine learning research community and is also known as the problem of interpretability.

Likewise, it is essential that Artificial Intelligence systems be transparent not only in relation to what data they capture and analyze about human behavior and for what purposes -which is contemplated in the GDPR at a European level- but also in what situations humans are interacting with artificial systems (*e.g.* chatbots) vs with other humans.

Beneficence and Education

•*Sustainability.* Technological progress in general and in particular of Artificial Intelligence systems requires significant energy consumption, with negative impact on the environment, as previously described. Deep learning techniques require high computing capabilities with prohibitive energy costs, especially if we consider the deployment of this type of systems on a large scale. Therefore, it is increasingly important that technological development is aligned with the human responsibility to guarantee the basic conditions for life on our planet and to preserve the environment for future generations. Likewise, Artificial Intelligence will be key to enable us to address some of the most important challenges in the context of the environment (for example, climate change, the scarcity of resources ...) as well as to allow us to develop sustainable transportation (for example, autonomous electric cars) and more efficient and sustainable energy models (for example, smart grids).

**Veracity.* Today we can create synthetic content (text, photos, videos), invented by algorithms, that is indistinguishable from truthful content. This ability led to the emergence of *fake news*, which can define public opinion on important issues -as who should be the next president of a country or whether a country should belong or not to the European Union- to favor the interests of a minority that has the ability to generate and disseminate such fake news. Therefore, the principle of veracity of both the data used to train AI algorithms and the content we consume is of utmost importance.

**Diversity.* Given the variety of use cases in which Artificial Intelligence can be applied, it is important to reflect on the frequent lack of diversity in the teams that create the Artificial Intelligence systems, teams until now mostly composed of homogeneous groups of IT professionals. In the future, we should make sure that the teams are diverse both in relation to their areas of knowledge and in relation to demographic factors - and in particular gender, given that the percentage of professional women in IT is less than 20% in most of technology companies, as previously explained.

Likewise, personalization and recommendation algorithms often suffer from lack of diversity in their results, tending to pigeonhole their users in certain patterns of tastes, which gives rise to the appearance of what Pariser has called the *filter bubble* [38]. This lack of diversity in personalization / recommendation is not desirable, as it limits the opportunities of technology to help us discover content (for example, movies, books, music, news or friends) different to our own tastes and therefore content that would help us understand other points of view and encourages openness of mind.

Non-maleficence

• *Reliability and security*. The vast majority, if not all, of the systems, products and goods we use (for example, food, household appliances, vehicles, clothing, toys, medicines, medical devices, industrial machinery ...) are subject to strict quality, safety and reliability controls. to minimize the potential negative impact that they may have on society. Similarly, it is expected that Artificial Intelligence systems are also subject to similar processes. Beyond the theoretical processes of security, verification and reliability, perhaps it would make sense to create a European-level authority that would certify the quality, security and reliability of Artificial Intelligence systems before they were commercialized or implemented in society. Also, autonomous systems should ensure the safety and integrity of both the people who use them or are affected by their action and their own security against manipulation and cyberattacks.

• *Reproducibility.* To generate confidence, systems should have consistency in their operation so that their behavior is not only understandable by a human, but also reproducible, that is, replicable when subjected to the same input data or the same situation/context

• *Prudence*. The application of Artificial Intelligence requires professionals to meet strict requirements for their development, such as ensuring the availability of enough (high quality) data, the analysis of working hypotheses from different perspectives and the availability of experts and resources to be able to make an analysis and interpretation of the models and their results. The principle of prudence emphasizes the importance of considering different options in the initial phases of design of any system to maximize its positive impact and minimize the potential risks and negative consequences derived from its application.

• *Data protection and privacy.* In a world of data, in which we generate and consume data in a ubiquitous and massive way, the rights to personal data protection and respect for privacy are constantly questioned and even pushed to the limit. Numerous studies have focused on the misuse of personal data provided by users of services and the aggregation of data from different sources by entities such as data brokers with direct implications on people's privacy. An element that is often ignored is that advances in machine learning algorithms, combined with the availability of new data sources of human behavior (for example, social media data), allow the inference of private information. (for example, sexual orientation, political inclinations, level of education or levels of emotional stability) that has never been explicitly revealed by people. In a recent research project we showed that we were able to infer attributes as personal as some dimensions of personality, level of education or interests from non-personal data [37]. This element is essential to understand the implications of the use of algorithms to model -or even influence- human behavior at the individual level, as has been clear in the recent Facebook/Cambridge Analytica scandal[32]. Therefore, I believe that there should be attributes and characteristics of people who would have to remain in the private sphere (for example, sexual orientation, religion ...) -unless the person decides that it is expressly so- and should not be used or inferred by AI systems.

Europe has assumed some leadership worldwide with the recent implementation of the General Data Protection Regulation or GDPR, which adds rights such as the right to establish and develop relationships with other human beings, the right to technological disconnection and the right to be free of vigilance. In this context, other rights that we could/should add include the right to meaningful human contact -for example, in services operated exclusively by chatbots- and the right not to be measured, analyzed, profiled, oriented or subliminally influenced through algorithms.

Five of the above principles can be described with the acronym FATEN shown by the frame colors above and referring to: F of *fairness*, solidarity and cooperation; A of *accountability*, *autonomy* and *augmentation*; T of *transparency*; E of *bEneficience* (promoting progress, preserving human dignity, with sustainability and diversity) and education; and N of *non-maleficence* (preserving privacy, with security, veracity, reliability and prudence). Many of these principles are described in the literature, such as in [39]

In addition, we should always focus the development of AI systems on people and encourage the creation of collaborative environments to experiment and co-create policies and solutions based on AI, agreed upon by humans.

I believe that it will only be when we respect these principles that we will be able to move forward and achieve another one of my dreams: a model of democratic governance based on data and artificial intelligence, by and for the people.

Towards a better future thanks to Artificial Intelligence

I do not know what the future will bring and where we will be in twenty years, but I can dream how I would like it to be. Therefore, I conclude my speech by sharing three dreams for the future with you.

First, I would like it to be a future where technology in general -and Artificial Intelligence in particular- is an integral part of our lives, where we co-exist synergistically and harmoniously with technology that helps us not only to live longer, but also to everything to live better, to all. The potential to have a positive impact is immense and I think we should not miss this opportunity.

---

[32] https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

However, it is not a certain future as I have articulated previously. Therefore, we must seriously face the challenges and limitations presented by the current Artificial Intelligence systems –including the ones previously described- to make this dream a reality. I hope not only to be part of that future but also and very especially to be able to contribute to it with my work.

Secondly, I dream that Spain will invest much more in Artificial Intelligence than it does now, fulfilling the principles explained above, to become a leader in Europe and a bridge with Latin America and Africa. Ideally, we would stand out for an ambitious commitment to achieve the adoption of AI in our companies and public administrations, with a remarkable reinforcement of our scientific excellence in this area, investing significantly in the nurturing, attraction and retention of talent, in the updating of our educational system and in the promotion of creativity and innovation. We have the opportunity to elevate not only business and economic growth, but above all our quality of life. I hope, and I wish that we do not miss this opportunity.

Finally, I hope, I wish -and I dream- that there will be many more female researchers, engineers, inventors, innovators in technology and members of this Royal Academy, who are exceptional not because of their condition of being women, but because of the brilliance of their ideas and the impact of their work.

Thank you for contributing with this recognition that at least one of my dreams becomes a reality.

Bibliography

[1] N. . Oliver, A. . Pentland and F. . Bérard, "LAFTER: a real-time face and lips tracker with facial expression recognition," *Pattern Recognition,* vol. 33, no. 8, pp. 1369-1382, 2000.

[2] N. Oliver, E. Horvitz and A. Garg, "Layered representations for human activity recognition," in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, Pittsburgh, PA, USA , 2002.

[3] N. Oliver, B. Rosario and S. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 22, no. 8, pp. 831-843, 2000.

[4] N. Oliver and A. Pentland, "Graphical models for driver behavior recognition in a SmartCar," in *Proceedings of the IEEE Intelligent Vehicles Symposium 2000*, Dearborn, MI, USA , 2000.

[5] N. Oliver and F. Flores-Mangas, "HealthGear: a real-time wearable system for monitoring and analyzing physiological signals," in *International Workshop on Wearable and Implantable Body Sensor Networks*, Cambridge, MA, USA , 2006.

[6] J. San Pedro, D. Proserpio and N. Oliver, "MobiScore: towards universal credit scoring from mobile phone data," in *International Conference on User Modeling, Adaptation, and Personalization*, 2015.

[7] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, Pianesi and A. F. and Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," in *ACM Int Conf on Multimodal Interaction (ICMI),* 2014.

[8] M. Pielot, T. Dingler, J. San Pedro and N. Oliver, "When attention is not scarce-detecting boredom from mobile phone usage," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Osaka, Japan, 2015.

[9] R. de Oliveira, M. Cherubini and N. Oliver, "MoviPill: improving medication compliance for elders using a mobile persuasive social game," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*,

Copenhagen, Denmark, 2010.

[10] N. Oliver and F. Flores-Mangas, "MPTrain: a mobile, music and physiology-based personal trainer," in *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, Helsinki, Finland, 2006.

[11] D. Rumelhart, G. E. E. Hinton and R. J. Williams, "Learning representations by back-propagating errors.," *Nature ,* vol. 323, pp. 533--536, 1986.

[12] N. Oliver, A. Pentland and F. Berard, "LAFTER: lips and face real time tracker," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, USA , 1997.

[13] M. Brand, N. Oliver and S. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, USA , 1997.

[14] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning," *Nature,* vol. 521, pp. 436--444, 2015.

[15] A. Karatzoglou, X. Amatriain, L. Baltrunas and N. Oliver, "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering," in *Proceedings of the fourth ACM conference on Recommender systems*, Barcelona, Spain , 2010.

[16] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, N. Oliver and A. Hanjalic, "CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering," in *ACM conference on Recommender systems*, New York , 2012.

[17] X. Amatriain, N. Lathia, J. Pujol, H. Kwak and N. Oliver, "The wisdom of the few: a collaborative filtering approach based on expert opinions from the web.," in *32nd international ACM SIGIR conference on Research and Development in Information Retrieval* , 2009.

[18] M. Vieira, V. Frias-Martinez, N. Oliver and E. Frias-Martinez, "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics," in *IEEE Second International Conference on Social Computing*, Minneapolis, MN, USA, 2010.

[19] J. Froelich, J. Neumann and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," in *Proceedings of Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[20] P. Jawandhiya, "Hardware design for machine learning," *International Journal of Artificial Intelligence and Applications (IJAIA),* vol. 9, no. 1, pp. 63 -- 84, 2018.

[21] A. Andrae, "Total Consumer Power Consumption Forecast," in *Nordic Digital Business Summit*, 2017.

[22] N. Bostrom, SuperIntelligence: Paths, dangers, strategies, Oxford, UK : Oxford University Press, 2014.

[23] J. Hawkins, "What intelligent machines need to learn from the neocortex," *IEEE Spectrum ,* 2017.

[24] S. Ahmad and J. Hawkins, "Properties of Sparse Distributed Representations and their Application to Hierarchical Temporary Memory," Arxiv, 2015.

[25] D. Lazer, M. Baum, Y. Benkler, A. J. Berinsky, K. Greenhill, F. Menczer and e. al, "The science of fake news," *Science,* pp. 1094--1096, 2018.

[26] M. Arntz, T. Gregory and U. Zierahn, "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis," *OECD Social, Employment and Migration Working Papers,* vol. 189, no. OECD Publishing , 2016.

[27] Y. N. Harari, 21 lessons for the 21st century, London: Penguin Random House, UK , 2018.

[28] S. Bocconi, S. Chioccariello, G. Dettori, A. Ferrari and K. Engelhardt, "Developing computational thinking in compulsory education," JCR Science for Policy Report, 2016.

[29] S. Lluna and J. Pedreira, Los nativos digitales no existen, Deusto Editorial , 2017.

[30] R. Steinpreis, K. Anders and D. Ritzke, "The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study," *Sex Roles,* vol. 41, no. 7--8, pp. 509-528, 1999.

[31] S. Centellegher, M. De Nadai, M. Caraviello, C. Leonardi, M. Vescovi, Y. Ramadian, N. Oliver, F. Pianesi, A. Pentland, F. Antonelli and B. Lepri, The Mobile Territorial Lab: a multilayered and dynamic view on parents' daily lives, EPJ Data Science, 2016.

[32] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello and N. Sebe, "Money walks: a human-centric study on the economics of personal mobile data," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle, WA, USA , 2014.

[33] Y. Torres Fernández, D. Pastor Escuredo, A. Morales Guzmán, J. Baue, A. Wadhw, C. Castro Correa, L. Romanoff, J. Lee, A. Rutherford, V. Frias Martínez, N. Oliver, E. Frias-Martinez and M. Luengo-Oroz, "Flooding through the lens of mobile phone activity," in *Proceedings of IEEE Global Humanitarian Technology Conference*, San Jose, CA, USA , 2014.

[34] B. Lepri, N. Oliver, E. Letouzé, A. Pentland and P. Vinck, "Fair, Transparent, and Accountable Algorithmic Decision-making Processes," *Philosophy & Technology,* pp. 1-17, 2017.

[35] J. Burrell, " How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data and Society ,* 2016 .

[36] J. Kroll, "Accountable Algorithms," PhD Dissertation in the Computer Science Department of Princeton, Princeton, US, 2015.

[37] S. Park, A. Matic, K. Garg and N. Oliver, "When Simpler Data Does Not Imply Less Information: A Study of User Profiling Scenarios With Constrained View of Mobile HTTP (S) Traffic," *ACM Transactions on the Web (TWEB),* vol. 12, no. 9, 2018.

[38] E. Pariser, The filter bubble: how the personalized web is changing what we read and how we think, New York: Penguin Books, 2012.

[39] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Shafer, P. Valcke and E. Vayena, "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles and Recommendations," *Minds and Machines ,* no. December , 2018.