rac.2023.112.1.org03
RACSG.2023;112(01):28-39
Recibido: 26/05/2023 | Aceptado: 07/06/2023

ORIGINAL

ARTIFICIAL INTELLIGENCE: PAST, PRESENT AND FUTURE

INTELIGENCIA ARTIFICIAL: PASADO, PRESENTE Y FUTURO

Verónica Bolón-Canedo[1]; Laura Morán-Fernández[2]

1. Department of Computer Science and Information Technologies, Universidade da Coruña, CITIC. Corresponding Academic of the Spanish Royal Academy of Science.
2. Department of Computer Science and Information Technologies, Universidade da Coruña, CITIC

ABSTRACT

This article will review the history of Artificial Intelligence (AI) and its transformative capacity on the world. Although its birth dates back to the mid-20th century, it is now that the factors have converged to make AI the key ingredient of the so-called Fourth Industrial Revolution: a significant increase in computer computing capabilities and the emergence of the Big Data phenomenon. In recent years, we have witnessed the most spectacular advances in AI, with notable applications in fields such as medicine, finance, industry, and law. However, AI still faces many challenges, and it is expected that the coming years will bring even more impressive developments. One of the major challenges is to ensure that AI is ethical and reliable. There are increasingly more intelligent systems that make decisions that impact our lives, so it is of utmost importance to ensure that these systems are robust, transparent, and fair. Furthermore, due to the trend of using increasingly larger and complex computational models, there is a growing concern about the energy resources consumed by AI algorithms. Therefore, another challenge for AI is to achieve increasingly sustainable and inclusive models.

Keywords: Artificial intelligence; Machine learning; Deep learning.

RESUMEN

Este artículo repasará la historia de la Inteligencia Artificial (IA) y su capacidad transformadora en el mundo. Aunque su nacimiento se remonta a mediados del siglo XX, es ahora cuando han confluido los factores que han hecho de la IA el ingrediente clave de la llamada Cuarta Revolución Industrial: un aumento significativo de las capacidades de computación de los ordenadores y la aparición del fenómeno Big Data. En los últimos años, hemos sido testigos de los avances más espectaculares en IA, con notables aplicaciones en campos como la medicina, las finanzas, la industria y el derecho. Sin embargo, la IA aún se enfrenta a muchos retos, y se espera que los próximos años traigan consigo avances aún más impresionantes. Uno de los principales retos es garantizar que la IA sea ética y fiable. Cada vez hay más sistemas inteligentes que toman decisiones que afectan a nuestras vidas, por lo que es de suma importancia garantizar que estos sistemas sean sólidos, transparentes y justos. Además, debido a la tendencia a utilizar modelos computacionales cada vez más grandes y complejos, existe una creciente preocupación por los recursos energéticos que consumen los algoritmos de IA. Por lo tanto, otro reto para la IA es conseguir modelos cada vez más sostenibles e inclusivos.

Palabras clave: Inteligencia artificial; Aprendizaje máquina; Aprendizaje profundo.

Correspondencia
Verónica Bolón-Canedo
Facultad de Informática. Campus de Elviña, s/n · 15071 A Coruña
E-mail: veronica.bolon@udc.es

INTRODUCTION

In this paper, we will review the history of Artificial Intelligence, an academic discipline which has more than half a century of history, but that nowadays has achieved its highest accomplishments. First, we will analyze the concept. According to the Royal Spanish Academy[1], the term “Artificial Intelligence” (AI) is defined as that scientific discipline that deals with

[1] ***https://dle.rae.es/inteligencia***

creating computer programs that perform operations comparable to those performed by the human mind, such as learning or logical reasoning. AI encompasses many sub-areas of work, such as robotics, natural language processing, automatic reasoning, machine learning, machine vision, intelligent agent-based modeling, ***Big Data***, etc. On the other hand, it is also a cross-discipline as it affects many fields of application such as healthcare, education, law, environment or industry, among others. And finally, AI is an increasingly broad and interdisciplinary field, as it has unquestionable implications in many aspects of society.

At regular intervals since the 1950s, experts predicted that it would take a few years for intelligent systems to exhibit behavior indistinguishable from humans in all respects and to have cognitive emotional and social intelligence. Only time will tell whether this will be the case. But to better understand what is feasible, we will look at AI from two angles: the road already traveled and the road still ahead. In this article, we will start by looking at AI's past to see how far it has evolved, as well as the social and economic change it has brought to our present. And finally, we will look at the challenges that AI poses for the future.

ORIGIN AND HISTORY OF ARTIFICIAL INTELLIGENCE

None of the advances achieved by AI were instantaneous, but all of them required considerable prior research and reflection, as well as reaching a certain maturity in their application. AI is no exception; it is not as recent a discipline as its current prominence might suggest. Thus, throughout this section we will analyze the history of AI from its birth to the present day, going through the so-called four seasons of AI (Haenlein & Kaplan, 2019).

1. AI Spring: the birth of AI

Although it is difficult to specify, perhaps the most remote historical antecedents of Artificial Intelligence are to be found in ancient Greece (Moret et al, 2005). Specifically, Archimedes, Demetrius of Phalerum, Archytas of Tarentum and Herron of Alexandria were the forerunners of the discipline now known as ***Automatics***. However, the real history of AI begins with Babbage's desire for his Analytical Engine to think, learn and create. Ada Lovelace, Babbage's collaborator, was surely the first programmer. She wrote programs for the unfinished Analytical Engine and even speculated about the possibility of the machine playing chess and composing music.

More recently, the roots of AI probably go back to the 1940s, specifically to 1942, when American science fiction writer Isaac Asimov published his story "***Runaround***". The story's plot revolves around the Three Laws of Robotics: (1) a robot shall neither harm a human being nor, by inaction, allow a human being to be harmed; (2) a robot must carry out orders given by humans, except for those that conflict with the first law; and (3) a robot must protect its own existence to the extent that this protection does not conflict with the first or second law. Asimov's work inspired generations of scientists in the field of robotics, AI and computer science, including American cognitive scientist Marvin Minsky, who later co-founded the MIT AI lab.

At about the same time, but in the United Kingdom, the mathematician Alan Turing was working on much less fictitious subjects and developed a machine called "Bombe" for the British government, with the purpose of deciphering the Enigma code used by the German army in World War II and thus being able to locate them, thus anticipating their strategy. The way in which "Bombe" was able to decipher the Enigma code, a task hitherto impossible even for the best mathematicians, made Turing wonder about the intelligence of such machines. In 1950, he published the article "Computing Machinery and Intelligence" (Turing, 1950), in which he described how to create intelligent machines and, in particular, how to test their intelligence. This Turing test, as illustrated in Figure 1, is still considered today as a benchmark for identifying the intelligence of an artificial system: if a human person has a conversation with a machine and another person, but fails to distinguish which of the two conversationalists is actually a machine, then the machine is said to be intelligent.

However, the first work on AI is credited to (McCulloch & Pitts, 1943). They rely on three sources: information about the basic physiology and function of neurons in the brain, the formal analysis of propositional logic by Russell and Whitehead, and Turing's theory of computation. For example, they show that any computational function can be computed using some network of interconnected neurons and that all logical connectors can be implemented using a simple network structure.

The term Artificial Intelligence was officially coined in 1956, when Marvin Minsky and John McCarthy (Stanford computer scientist) organized the Dartmouth Summer Research Project on Artificial Intelligence[2], lasting approximately two months. This project, marking the beginning of the AI Spring, brought together those who would later be considered the founders of AI. Participants included computer scientist Nathaniel Rochester, who would later design

[2] ***http://raysolomonoff.com/dartmouth/***

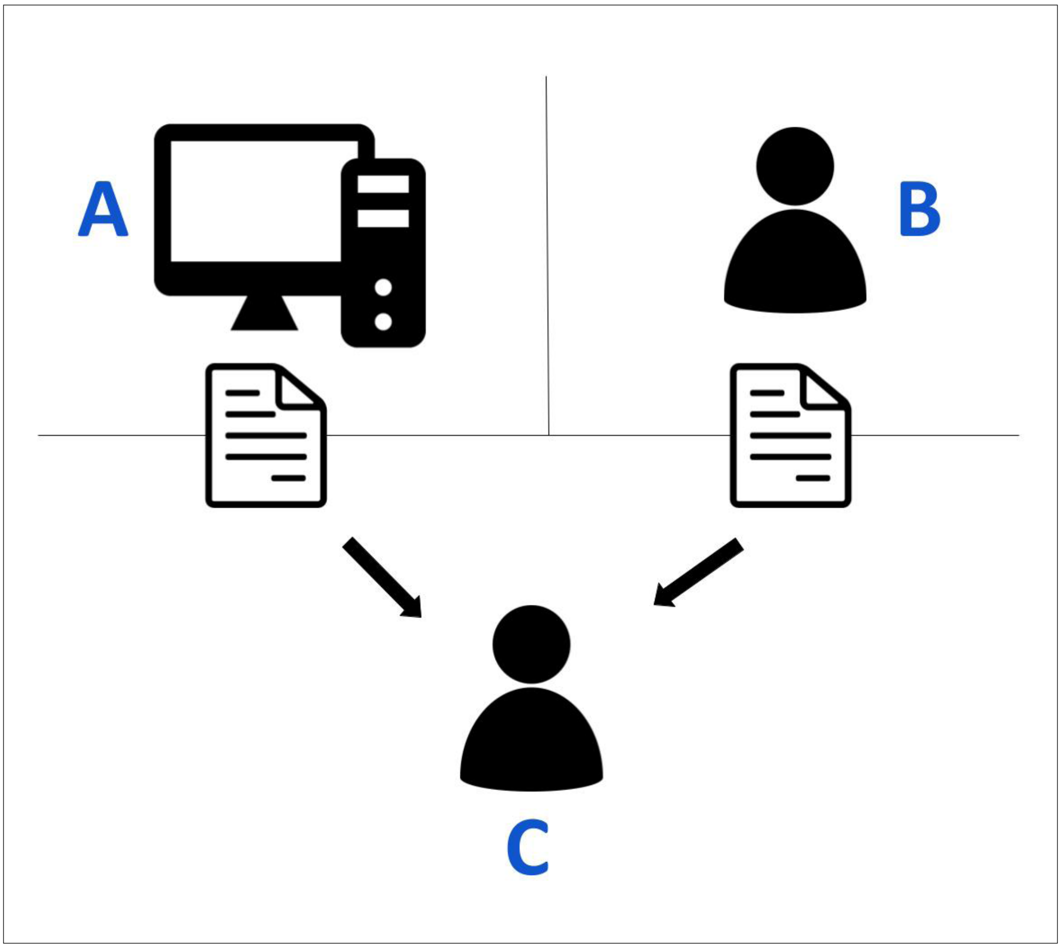

***Figure 1.** The standard interpretation of the Turing test, in which player C, the interrogator, is given the task of trying to determine which player (A or B) is a computer and which is a human.*

the IBM 701, the first commercial scientific computer, and mathematician Claude Shannon, who founded the field of Information Theory. The goal of the project was to bring together researchers from diverse areas to create a new field of research, aimed at building machines that use language, form abstractions and concepts, solve types of problems now reserved for humans, and improve themselves. To do this, they relied on the conjecture that every aspect of learning or any other feature of intelligence can, in principle, be described so precisely that a machine can be created to simulate it (Russell & Norvig, 2010).

However, the first work on AI is credited to (McCulloch & Pitts, 1943). They rely on three sources: information about the basic physiology and function of neurons in the brain, the formal analysis of propositional logic by Russell and Whitehead, and Turing's theory of computation. For example, they show that any computational function can be computed using some network of interconnected neurons and that all logical connectors can be implemented using a simple network structure.

The term Artificial Intelligence was officially coined in 1956, when Marvin Minsky and John McCarthy (Stanford computer scientist) organized the Dartmouth Summer Research Project on Artificial Intelligence[3], lasting approximately two months. This project, marking the beginning of the AI Spring, brought together those who would later be considered the founders of AI. Participants included computer scientist Nathaniel Rochester, who would later design the IBM 701, the first commercial scientific computer, and mathematician Claude Shannon, who founded the field of Information Theory. The goal of the project was to bring together researchers from diverse areas to create a new field of research, aimed at building machines that use language, form abstractions and concepts, solve types of problems now reserved for humans, and improve themselves. To do this, they relied on the conjecture that every aspect of learning or any other feature of intelligence can, in principle, be described so precisely that a machine can be created to simulate it (Russell & Norvig, 2010).

[3] ***http://raysolomonoff.com/dartmouth/***

However, as often happens, there was no consensus and two main groups emerged: those who thought that symbolic representation was mainly based on logic (i.e., syntax and predicate calculus), and those who thought it was mainly based on semantics. The former opted for elegant, clear and probably correct solutions, while the latter thought that the intelligence is too complicated, probably computationally intractable, and therefore cannot be solved with the kind of homogeneous system that clear requirements demand.

2. AI summer and winter: The ups and downs of AI

The Dartmouth Project was followed by a period of nearly two decades that saw significant successes in the field of AI. An early example is the famous ELIZA computer program, created between 1964 and 1966 by Joseph Weizenbaum at MIT. ELIZA was a natural language processing tool capable of simulating a conversation with a human. Another early AI success was the "General Problem Solver" program, developed by Nobel laureate and RAND Corporation scientists Cliff Shaw and Allen Newell, which was capable of automatically solving certain types of simple problems, such as the Towers of Hanoi. As a result of these achievements, funds were poured into AI research, leading to more and more projects. In 1970, Marvin Minsky gave an interview in which he stated that within three to eight years, a machine with the general intelligence of an average human being could be developed.

However, this was not the case. Only three years later, in 1973, the U.S. Congress began to strongly criticize the high spending on AI research. That same year, the British Council for Scientific Research asked Professor James Lighthill to assess the state of AI research in the UK. His report, which criticized the failure of AI to achieve its "grandiose goals," ended support for AI research in virtually all British universities; a response that was soon followed by the U.S. government. It was during this period that the AI winter began. And although the Japanese government began to heavily fund AI research in the 1980s, to which the U.S. DARPA responded with increased funding as well, no further progress was made in the following years.

3. The autumn of AI: the harvest

One of the reasons for the lack of initial progress in the field of AI and the fact that reality fell back sharply relative to expectations lies mainly in the specific way in which early systems, such as ELIZA or the "General Problem Solver", attempted to replicate human intelligence. Specifically, they were all expert systems, i.e., collections of rules that assume that human intelligence can be formalized and reconstructed in a top-down approach as a series of "if-then" statements. Expert systems can perform impressively well in areas that lend themselves to such formalization. For example, IBM's Deep Blue chess program, which in 1997 was able to defeat the then world champion Garri Kasparov, is such an expert system. Deep Blue is said to have been able to process 200 million possible moves per second and to determine the optimal next move by analyzing 20 moves ahead by using a method called tree search (Campbell et al, 2002).

However, expert systems do not perform well in areas that do not conform to this formalization. For example, an expert system cannot be easily trained to recognize faces or even to distinguish an image showing a cupcake from one showing a chihuahua. To perform these tasks requires a system capable of correctly interpreting external data, learning from that data and using those learnings to achieve goals through flexible adaptation. Statistical methods to achieve more powerful AI had been discussed as early as the 1940s, when Canadian psychologist Donald Hebb developed a learning theory known as Hebbian Learning, which replicates the process of neurons in the human brain. This led to the creation of research on artificial neural networks, most notably the Perceptron, developed by psychologist Frank Rosenblatt in 1958. However, this work stalled in 1969 when Marvin Minsky and Seymour Papert showed that computers did not have sufficient processing power to do the work required by these artificial neurons.

Artificial neural networks made a comeback when in 2015 AlphaGo, a program developed by Google DeepMind, was able to beat the world champion in the board game Go. Go is substantially more complex than chess (for example, there are 20 possible moves in the opening in chess, compared to 361 in Go) and, for a long time, it was believed that computers would never be able to beat humans at this game. Another milestone in the history of AI proved wrong one of James Lighthill's assertions that machines would only reach the level of an "experienced amateur" at board games.

This harvest of the fruits of past improvements and advances seen in recent years constitutes the autumn period of AI, where we find ourselves today, thanks mostly to techniques such as machine learning. Machine learning is often confused with artificial intelligence, but it is only one part of it. It involves processes in which the machines themselves create their own rules (algorithms) and predictions based on data provided by humans. An example is the qualitative leap made by translation engines such as Google Translate or DeepL. From translating based on syntactic rules, they started to do so based on millions of examples of real translations.

Deep learning is a subdomain or type of machine learning. Its power is based on what are known as neural networks, i.e. layers and layers of information processing. Unlike machine learning, here it is the systems, with little supervision, that are capable of learning to improve themselves as they gain experience. Much of the most recent innovation in AI is linked to this form of learning and the advent of *Big Data*.

THE PRESENT: ARTIFICIAL INTELLIGENCE AND ITS WORLD-TRANSFORMING CAPACITY

Modern economic history intersperses periods of incremental improvement with others in which certain discoveries and developments have fundamentally changed the way we relate to each other and organize ourselves as a society. In the midst of the 21st century, AI seems to be joining advances that were once considered revolutionary, such as navigation, steam engines or electricity.

This field of computing is going through its most glorious era today, not by accident, but by a combination of several factors:

- The availability of increasingly large data sets. Due to the gradual and intense digitization process underway, almost every human experience is being digitized, from travel to medical care. Not only human activity, we also have an increasing number of sensors recording data from natural processes that allow us to understand the behavior and evolution of our environment.

- The required computational capacity. Advances in high-performance computing technologies, the cheapening of cloud computing, and the availability of new parallel and distributed computing platforms enable fast and cost-effective processing of large amounts of heterogeneous data.

- The unprecedented social and technological change in which we find ourselves. In an increasingly interconnected world where people interact with our mobile devices, connectivity is critical. This connectivity creates both market opportunities and social adaptation challenges.

Progress has also been made on the software side. New types of databases allow us to store and process structured and unstructured data beyond classical scientific data. At the same time, the emergence of new theoretical developments (mainly mathematical) has been disruptive, such as those obtained in the fields of deep learning, reinforcement learning or natural language processing that have yielded highly accurate results, positioning AI as a successful and transformative mature technology.

Currently, what we have is specific AI. Specific (or weak) AI is understood as a system oriented to solve specific and delimited problems, something that machines learn to do through repetitive patterns and trends thanks to algorithms programmed by humans. An example would be a virtual assistant that, although it is capable of "holding" a conversation, does not fail to respond to specific commands with the results of a search on the Internet or among its databases. But it does not have a general intelligence, like that of humans, capable of addressing decisions in a proactive, deductive and self-aware manner. This type of AI exists only in the realm of science fiction. It is the notion of AI shown by the film industry in productions such as A Space Odyssey (1969), Blade Runner (1982), Matrix (1999), Eva (2011), Her (2014) or Ex Machina (2015).

Still, there are many domains in which AI outperforms human intelligence, such as in specific areas of medicine, finding solutions to logical formulas with many variables, recommender systems, etc. Robots, autonomous vehicles, personal assistants or automatic translators are all successes of specific AI.

One of the most recent advances has been the launch of ChatGPT. ChatGPT is a chat system based on the Artificial Intelligence language model GPT-3 (later upgraded to GPT-4), developed by OpenAI. It is a model with over 175 million parameters (1 billion in the case of GPT-4), trained on large amounts of text to perform language-related tasks, from translation to text generation. It is capable of providing very accurate and comprehensive responses, even spanning multiple paragraphs. Moreover, in these responses, it is able to express itself naturally and with highly precise information.

Today, technological maturity has pushed AI out of the realm of research and into everyday life and the business world, where the following applications stand out:

- Personalization and segmentation tools. One of the most widespread uses of AI is the analysis of large volumes of data in order to extract patterns of human behavior that make it possible to segment the product or service offerings of a given company, for example. Consumers, when interacting online, generate data that can be collected, classified and used. Given that, as mentioned above, the number and volume of data sets is growing all the time, the assistance of AI systems is becoming indispensable. In fact, AI-based computing techniques are expected to become an important factor in decision making. AI systems not only enable the analysis of behavioral patterns and thus the design of user profiles, but can also be used to suggest personalized advertising campaigns and create satisfying consumer experiences for customers.

- Virtual assistants. The main function of virtual assistants, which are already a member of our homes, is to help users resolve their doubts and perform certain actions, such as purchasing products or consulting information of any kind. Today, this is possible thanks to the advances that have taken place in recent years related to machine learning and natural language processing. Natural language processing is a branch of AI that makes it possible to understand the meaning of sentences uttered by a user. This has some limitations mainly due to the inherent ambiguity of human language.

- Robotics. Robotics is the science or branch of technology that studies the design and construction of machines, also known as robots, capable of performing tasks performed by humans or that require the use of intelligence. These robots operate in the real world, either virtually, such as a web ***chatbot*** for example, or physically, such as a machine that can be touched. Depending on what they are designed to do, physical robots can move, understand messages and communicate, or manipulate objects precisely, to give a few examples. All these tasks make robotics one of the most complex branches of AI, as it involves interaction with other branches and disciplines. A physical robot needs to recognize objects (machine vision), understand meanings (natural language processing) and make decisions while communicating effectively (automatic speech recognition). Moreover, for all this to be possible, the intervention of other branches, such as mechanics, engineering or electronics, is required.

- Autonomous vehicles. The arrival of these vehicles will bring about a revolutionary transformation not only in mobility, but also in legislation, in the appearance of cities and even in the concept of ownership. Autonomous vehicles will be a reality in the short term for the transport of goods, but perhaps also in the medium term for our journeys. According to a study by Tony Seba (Seba, 2014), an economist at Stanford University, autonomous vehicles will be used ten times more than owned cars, which spend less than 4% of their useful life in operation. As a result, by 2030, travel in electric and autonomous vehicles will be four to ten times cheaper than owning a new vehicle. To achieve autonomous driving, driverless cars must be equipped with a large number of cameras, sensors, radars and GPS, which generate a large amount of information and, channeled by an intelligent system, allow the vehicle itself to "make decisions" on its own.

- The Internet of Things. The Internet of Things (IoT) refers to the entire network of physical objects (vehicles, machines, household appliances, etc.) that are connected and exchange data via the Internet through sensors. In theory, the goal of IoT is to have most objects permanently connected to the Internet to improve their functionality, including those beyond specific uses. Most AI systems have the ability to collect data about their users and their respective environments. There are many examples of IoT devices equipped with sensors: smart watches that monitor the health of their users, refrigerators that monitor food consumption, cars that collect data on how their owners drive, and so on. All these devices generate a lot of data, and if there is one technology that can take advantage of it, it is AI.

But the expansion of this successful AI does not end here; new areas of AI application are continuously emerging. Sectors such as healthcare and education will change dramatically through the use of AI. Possibilities are opening up for the development of new drugs and more personalized treatments, as well as the identification of genetic factors susceptible to developing a disease thanks to data collection. In the field of education, it would allow more personalized attention to students in order to optimize learning. AI is also being studied to help sectors of the population and geographic areas at risk of exclusion or depressed areas. The incorporation of healthcare personnel or virtual educators could bring quality services, at low cost, to places where they had not previously penetrated, making it possible to recover our welfare society in an economically viable way.

Thanks also to AI, other sectors such as banking can create end products that bring value to both the entity and the customers themselves, allowing them to extract relevant data from ***Big Data***, search for patterns that contribute to more personalized offers or detect bank fraud processes. In industry, robotics and the implementation of AI-based systems continue what other automation technologies have already done in the past. However, these systems are not limited to replacing tasks that require human power; today, AI-driven technologies can retrieve information, coordinate logistics, manage inventories or provide certain services.

The objective of AI applied to law is to complement the activity of legal professionals. For example, providing all the documentation and jurisprudence related to the case to the tool so that it can process all the information and provide legal support. The ultimate goal is to achieve faster and more efficient law enforcement, in addition to providing legal professionals with the necessary tools to free themselves from more mechanical tasks to focus on more specialized work.

Even the environment could benefit from AI, with fleets of drones capable of planting a billion trees a year to combat deforestation, unmanned underwater vehicles to detect pipeline leaks, or smart buildings designed to reduce energy consumption.

THE FUTURE OF ARTIFICIAL INTELLIGENCE: CHALLENGES AND OPPORTUNITIES

With the advent of ***Big Data***, we have witnessed an overwhelming increase in the volume of data, along with other aspects such as variety, variability, veracity, etc. Initially, this situation made some classical machine learning algorithms unusable, so it soon became essential to make the algorithms scalable, viable, practical and able to adequately deal with ***Big Data*** problems.

Interest in data science and artificial intelligence will continue to increase in the future, as we will be working in more and more areas (e.g., law, finance, marketing) or unexpected topics such as the fight against COVID-19, which require ever greater interdisciplinary skills. From a more technical point of view, although less infrastructure work is to be expected as the market is settled, faster and more powerful solutions will be needed, so there is still room for improvement. Other challenges to be addressed due to the increasing volume of data will be the need for preprocessing to obtain quality data, visualization to explain the conclusions reached and privacy in the handling of the data within our algorithms.

From a regulatory point of view, the entry into force in the European Union of the General Data Protection Regulation (GDPR) on May 24, 2016 and applicable since May 25, 2018, definitely introduced a new scenario, establishing the lawful processing and use of data through a framework of principles and a set of rights aiming at a transparent, accurate and fair Artificial Intelligence to prevent discrimination based on race, health, sex, etc. The GDPR will certainly not be the only one. Stricter data privacy legislations with extraterritorial applicability are appearing in more and more economies around the world, such as the US, Australia, or Canada[4], where sector-specific data protection and privacy regulations apply. The expansion of data protection legislation puts pressure on the field, with the need for ethical and legal standards that ensure privacy, lawful use of data, transparency in decisions, etc., to be constantly updated.

In April 2021, the European Commission published a Communication to announce the European approach to Artificial Intelligence[5], clearly defining the two sides, opportunities and risks, of AI, with reliability being the combination of all of them. Another European initiative focused on AI was the creation of the Artificial Intelligence High Level Expert Group (AI HLEG)[6], composed of representatives from academia, civil society and industry. Its objective is to support the implementation of the European strategy on AI, including the elaboration of recommendations on the future development of AI-related policies and on ethical, legal and societal issues related to AI (including socio-economic challenges). During the first year of its mandate, AI HLEG worked on two main deliverables: “Ethical Guidelines for Trusted AI”[7], which proposes a human-centered approach and lists seven key requirements that AI systems must meet to be trusted, and “Policy and Investment Recommendations for Trusted AI”[8], which presents 33 recommendations to guide trusted AI towards sustainability, growth, competitiveness and inclusion, while at the same time, the recommendations will empower, benefit and protect European citizens.

[4] ***https://insights.comforte.com/countries-with-gdpr-like-data-privacy-laws***

[5] ***https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence***

[6] ***https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence***

[7] ***https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai***

[8] ***https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence***

The European Union (EU) is currently deliberating on a fresh legal framework intended to enhance regulations pertaining to the advancement and utilization of artificial intelligence. Known as the Artificial Intelligence (AI) Act[9], the proposed legislation primarily concentrates on fortifying regulations concerning data quality, transparency, human supervision, and accountability. It also seeks to tackle ethical concerns and implementation hurdles across a broad spectrum of sectors, including healthcare, education, finance, and energy.

The foundation of the AI Act resides in a classification system designed to ascertain the potential risk an AI technology may pose to an individual's health and safety, as well as their fundamental rights. This framework encompasses four distinct tiers of risk: unacceptable, high, limited, and minimal.

In the following, we will outline some challenges for trustworthy artificial intelligence.

1. Robustness

One of the requirements for reliable AI is to verify the robustness of the developed algorithms. Society demands new algorithms that are safe, reliable and also robust to cope with errors or attacks. A robust model is a combination of many aspects, from accurate classification or prediction to efficient optimization techniques.

Adversarial learning is a branch of AI that has gained importance in recent years. Its goal is to try to prevent possible attacks, or the introduction of false data that could fool a machine, but not a person. There are famous cases in which minimal noise is added to an image, imperceptible to the human eye. However, this noise is enough for the AI algorithm to believe that it is a different image. In some cases, the deception is as impressive as the case of a 3D print of a turtle that Google's artificial intelligence incorrectly classified as a rifle[10].

Another important property for achieving robustness is that the accuracy of the results can be confirmed and reproduced by independent evaluation, resulting in the study of the stability of the methods. It must be ensured that the result of an algorithm is consistent with the output and is not sensitive to small variations in the training data.

2. Transparency, explainability and interpretability

Transparency of AI systems is one of the principles of responsible AI (Dignum, 2017). Dormant terms, such as Explainable Artificial Intelligence (XAI), resurfaced in Google search trends and in the literature. The definition of XAI has not yet reached a consensus (Arrieta et al, 2020), although authors agree on aspects such as: explainable models, allowing human users to understand properly and trust. Thanks to this new XAI, we were able to implement a right to explanation, which is one of the pillars of the Ethical Guidelines for Trusted AI of the European AI expert group and is closely related to the principles and rights of the European RGPD law.

Like XAI, transparency does not have a consensus definition, and the concept is even a bit ambiguous, giving rise to different types of transparency (Weller, 2019), which obeys different motivations. That is, whether transparency is from the developer's perspective (to understand what works correctly and why), from the user's perspective (to understand what the system is doing and why), or whether it is to facilitate the monitoring and testing of security standards. Understanding the purpose for which it is necessary to develop more transparent systems is important, as different types of transparency require different types of explainability, with all that implies, such as different measures of effectiveness. It is important to note that, from the user's perspective, an intelligent system can be transparent even if not all parts included in the architecture are, which opens the door to complex but transparent models (Chen et al, 2018).

In the scientific literature to address explainability one can find some work related to the relationship between model complexity and its interpretability and explainability (Bai et al, 2021), work related to the fact that interpretability implies understanding, i.e. of global behavior or local behavior (Ribeiro et al, 2018), work related to the applicability of interpretability techniques to different types of algorithms. In the latter type of work, the literature identifies several techniques, and one of them is influence methods, which include feature selection methods (Guyon et al, 2008). Feature selection is crucial for a correct interpretability of the data, as it indicates which features are relevant for a given task and which are irrelevant and/or redundant.

Another aspect with different approaches in the literature is the evaluation of explanations. One of the two possible ways of evaluating an explanation, mentioned in the literature, is interpretability (Gilpin et al, 2018). Interpretability aims at producing descriptions simple enough to be understood by a human being (Vassiliades et al, 2021). Typical approaches to interpretability refer to: creating explanations for

[9] ***https://artificialintelligenceact.eu***

[10] ***https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed***

system responses, creating explanations for data representation, and creating explanations that produce systems. Works such as LIME (Ribeiro et al, 2016) produce explainability of the first type mentioned above. An example of the performance of LIME can be seen in Figure 2.

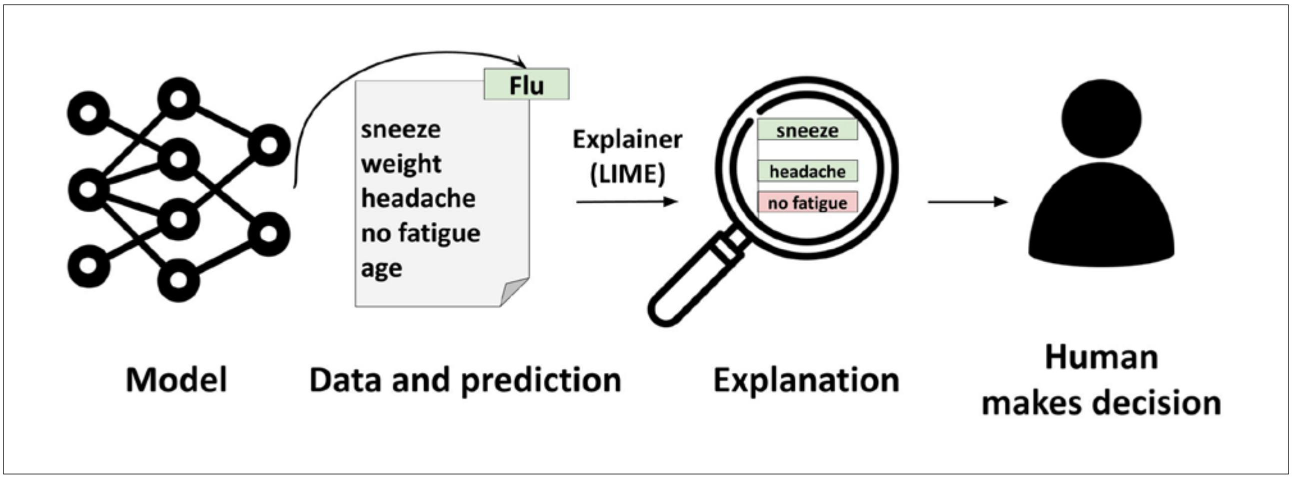

Figure 2. Explanation of individual predictions. A model predicts that a patient has the flu, and LIME detects the symptoms in the patient's history that led to the prediction (sneezing and headache). Source: Ribeiro et al.

3. Privacy and ethics

As noted in the Ethical Guidelines for Trusted AI, "AI systems must ensure privacy and data protection throughout the life cycle of a system." This aspect is based on the relationship between AI algorithms and society, as it is necessary to address the effect of the algorithm on society, its social implications, and what the algorithm does with the data it can access. Our multimedia consumption, Internet interaction, clinical records, and financial transactions are some of the many types of information that are periodically collected and stored. This data collection takes place on our mobile devices and personal computers, making it inherently distributed and sensitive in nature. Such private data is used for a variety of AI applications. Most of the time, this data is uploaded to centralized locations in raw format for AI algorithms to extract patterns and create models from it. This way of working is an obvious threat to users' privacy, so AI methods that take data privacy into consideration are of utmost importance (Azmoodeh et al, 2019).

One possible solution to respect the privacy of data collected on different devices is to develop algorithms *on the edge* (Shi et at, 2016), avoiding the transfer of sensitive data to the cloud. Even in the case of data that still needs to be processed remotely, *on the edge* devices can be used to discard personally identifiable information prior to data transfer, thus improving user privacy and security. In addition, decentralization can make systems more robust by providing transient services during a network failure or cyberattack (also related to system robustness). Learning *on the edge* may be particularly interesting for autonomous vehicles, because given the highly dynamic and real-time nature of driving, these vehicles must react to changes in the environment around them, regardless of the connectivity the car has at any given time. This is why learning *on the edge* can provide a more robust and adaptive paradigm.

There is other recent work focusing on privacy-preserving machine learning, including significant advances in secure multiparty computation (Riazi et al, 2018), homomorphic encryption (Sun et al, 2018), differential privacy (Yang et al, 2018), and federated learning (Kairouz et al, 2021). Efforts are being made to combine these approaches to achieve more robust models in terms of security, faster runtime or improved generalization performance. Another approach that allows data sharing while preserving privacy is to generate artificial data using original data as seeds. Data generation can be used to produce synthetic or semi-synthetic datasets that inherit features from the original. These methods allow maintaining the information of interest with high correlation to the original and eliminating sensitive information to be protected.

Recent research has shown that various private attributes of users (such as political affiliation, sexual orientation, and gender) can be inferred from user data to predict user preferences, make recommendations, and place targeted advertisements. It is also interesting to address the problem of algorithmic ***fairness*** in machine learning (Corbett-Davies et al, 2017). It has been documented that machine learning algorithms can behave very differently for different groups, e.g., minorities, race, or gender, depending on how these groups are represented in the training set. Many definitions of fairness have been introduced in the literature, most of which require the final model to behave statistically similarly; e.g., having an equal true positive rate for all protected groups, equal accuracy, etc. Although the general concept of fairness is relevant to any machine learning algorithm, only classification and regression tasks have been analyzed theoretically. It is still necessary to work on a general fairness framework and a set of algorithms to (a) prevent unfair treatment in different machine learning tasks such as classification, recommendation algorithms, clustering and (b) techniques to diagnose possible unfair biases in existing systems.

4. Sustainability

In recent years, the use of cloud computing has become popular for advanced computations that often require the most sophisticated deep learning models. However, according to Greenpeace, cloud computing sites can consume up to 622.6 billion kilowatts per hour, consuming an estimated 1-2% of the world's electrical resources (Cook, 2011). As a result, the field of green AI is gaining interest (Schwartz et al, 2020). This term refers to AI research that is more environmentally friendly and inclusive, not only by producing novel results without increasing computational cost, but also by ensuring that any researcher with a laptop has the opportunity to perform high quality research without the need to use cloud servers (with the economic cost involved). More classical AI research (sometimes referred to as red AI) aims to obtain state-of-the-art results at the expense of using massive computational resources, usually through a huge amount of training data and numerous experiments.

Efficient machine learning approaches (especially deep learning) are starting to receive attention from the scientific community. The problem is that, most of the time, the motivation for these works is not to achieve green AI, but to make models that can work on very small devices. Therefore, the AI community needs to be encouraged to recognize the value of the work of researchers who take a different path, optimizing for efficiency rather than just accuracy.

Edge computing is a promising approach to address the sustainability of AI models, which refers to computations being performed as close as possible to data sources, rather than in remote and distant locations (such as cloud servers). Recent studies have shown that a fully distributed architecture consumes 14% to 25% less energy than fully centralized and partially distributed architectures (Ahvar et al, 2019), as a result of not using cloud servers and large cooling systems. According to a report by information and communications technology research firm Gartner[11], while most data today is created and processed within centralized data centers, by 2025 about 75% of data will need analysis and action ***on the edge***.

The sustainability of AI is directly related to the Sustainable Development Goals (SDGs) defined by the United Nations[12] and the European Union's Green Deal principles[13]. In particular, the sustainability of algorithms is in line with SDG 12: achieving more responsible production and consumption.

CONCLUSIONS

This article has traced the history of Artificial Intelligence and its ability to transform the world. Although its birth dates back to the middle of the last century, it is only now that the factors that make AI the key factor in the so-called fourth Industrial Revolution have come together: a significant increase in the computational capabilities of computers and the emergence of the ***Big Data*** phenomenon.

In recent years we have witnessed the most spectacular advances in AI, with outstanding applications in fields such as medicine, finance, industry and law.

However, AI still has many challenges to meet, and the next few years are expected to bring even more impressive developments. One of the biggest challenges is making AI ethical and trustworthy. There are more and more intelligent systems making decisions that affect our lives, so it is of utmost importance to ensure that these systems are robust, transparent and fair. On the other hand, and due to the trend of using increasingly large and complex computational models, there is growing concern about the energy resources consumed by AI algorithms. This is why another challenge for AI is to achieve increasingly sustainable and inclusive models.

[11] https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders/

[12] https://sdgs.un.org/goals

[13] https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

DECLARATION OF CONFLICTING INTERESTS

The authors declare no conflicts of interest in preparing this article.

REFERENCES

1. Ahvar, E., Orgerie, A. C., & Lebre, A. (2019). Estimating Energy Consumption of Cloud, Fog, and Edge Computing Infrastructures. IEEE Transactions on Sustainable Computing, 7(2), 277-288.

2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

3. Azmoodeh, A., Dehghantanha, A., & Choo, K. K. R. (2019). Big data and internet of things security and forensics: Challenges and opportunities. Handbook of Big Data and IoT Security, 1-4.

4. Bai, X., Wang, X., Liu, X., Liu, Q., Song, J., Sebe, N., & Kim, B. (2021). Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. Pattern Recognition, 120, 108102.

5. Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep blue. Artificial intelligence, 134(1-2), 57-83.

6. Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018, July). Learning to explain: An information-theoretic perspective on model interpretation. In International Conference on Machine Learning (pp. 883-892). PMLR.

7. Cook, G. (2011). How dirty is your data: A look at the energy choices that power cloud computing. Greenpeace.

8. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining (pp. 797-806).

9. Dignum, V. (2017). Responsible Artificial Intelligence: Designing AI for Human Values, ITU Journal: ICT Discoveries, Special Issue, no. 1.

10. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.

11. Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). Feature extraction: foundations and applications (Vol. 207). Springer.

12. Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California management review, 61(4), 5-14.

13. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1-210.

14. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5, 115-133.

15. Moret, V., Alonso, A., Cabrero, M., Guijarro, B., & Mosqueira, E. (2000). Fundamentos de Inteligencia Artificial. Universidade da Coruña, Servicio de Publicacións, A Coruña.

16. Riazi, M. S., Weinert, C., Tkachenko, O., Songhori, E. M., Schneider, T., & Koushanfar, F. (2018, May). Chameleon: A hybrid secure computation framework for machine learning applications. In Proceedings of the 2018 on Asia conference on computer and communications security (pp. 707-721).

17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

18. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

19. Russell, S. J. & Norvig P. (2010). Artificial Intelligence: A Modern Approach. Englewood Cliffs, N.J., Prentice Hall.

20. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green ai. Communications of the ACM, 63(12), 54-63.

21. Seba, T. (2014). Clean disruption of energy and transportation: how silicon valley will make oil, nuclear, natural gas, coal, electric utilities and conventional cars obsolete by 2030, Tony Seba.

22. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE internet of things journal, 3(5), 637-646.

23. Sun, X., Zhang, P., Liu, J. K., Yu, J., & Xie, W. (2018). Private machine learning classification based on fully homomorphic encryption. IEEE Transactions on Emerging Topics in Computing, 8(2), 352-364.

24. Turing, A. (1950), Computing Machinery and Intelligence, Mind, 433-460

25. Vassiliades, A., Bassiliades, N., & Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. The Knowledge Engineering Review, 36, e5.

26. Weller, A. (2019). Transparency: motivations and challenges. In Explainable AI: interpreting, explaining and visualizing deep learning (pp. 23-40). Cham: Springer International Publishing.

27. Yang, M., Zhu, T., Liu, B., Xiang, Y., & Zhou, W. (2018). Machine learning differential privacy with multifunctional aggregation in a fog computing architecture. IEEE Access, 6, 17119-17129.

Si desea citar nuestro artículo:

Bolón-Canedo V, Morán-Fernández L. Artificial Intelligence: past, present and future. RACSG.2023;112(01): 28-39. rac.2023.112.1.org03