


Artificial intelligence and gender bias

Inteligencia Artificial y sesgos de género

Amparo Alonso Betanzos

CITIC. Universidade da Coruña. amparo.alonso.betanzos@udc.es,  <https://orcid.org/0000-0003-0950-0012>

Cátedra de Feminismos 4.0 DEPO-UVigo



Contacto:

Amparo Alonso Betanzos.
CITIC. Universidade da Coruña.
15008 A Coruña, España.

Cátedra de Feminismos 4.0
DEPO - UVigo



Universidade de Vigo

Resumen

Estamos inmersos en una nueva revolución, una era de transformación impulsada por la Inteligencia Artificial (IA), que afecta significativamente al equilibrio geopolítico, la sociedad, la economía, el empleo y la educación, generando cambios constantes en estos ámbitos. La IA es una disciplina transversal que está presente en prácticamente cualquier campo, desde la Industria, la Salud o el Medioambiente hasta áreas relacionadas con las Ciencias Sociales y las Humanidades. Esta omnipresencia trae consigo innumerables oportunidades y abre nuevas perspectivas, pero también trae retos, algunos relacionados con ciertos desafíos éticos que pueden aparecer por el tratamiento de datos a gran escala que hace la tecnología. Uno de ellos es la posible aparición de sesgos de género, que pueden deberse a que el funcionamiento de los algoritmos no ha sido suficientemente examinado en este sentido, o a que el entrenamiento de los modelos se ha realizado con datos históricos cuya calidad no es la adecuada, entre otros. Además, es fundamental tener en cuenta que los sesgos no sólo están en los algoritmos, también pueden derivar de desigualdades en el acceso a la tecnología, o de falta de diversidad en los equipos de diseño, entre otros. Estos factores pueden limitar la perspectiva y comprensión holística de los problemas, perpetuando así prejuicios y desigualdades en la IA. Es imprescindible abordar con responsabilidad estos sesgos si queremos garantizar una IA ética, confiable y justa. No es menos importante fomentar la diversidad en los equipos de desarrollo y diseño de tecnologías de IA, para que se reflejen diferentes perspectivas que puedan conducir a soluciones más inclusivas y equitativas.

Palabras clave

Inteligencia Artificial ética, Sesgos de género, Discriminación algorítmica

Abstract

We are immersed in a new revolution, an era of transformation driven by Artificial Intelligence (AI), which significantly affects the geopolitical balance, societal norms and behaviour, the economy, employment, and education, generating constant changes in these areas. AI is a transversal discipline that is present in practically any field, from Industry, Health, or the Environment to areas related to the Social Sciences and Humanities. This omnipresence brings with it innumerable opportunities and opens new perspectives, but it also brings challenges, some related to certain ethical issues that can arise from the large-scale processing of data of AI algorithms. One of them is the possible appearance of gender biases, which may be because the operation of the algorithms has not been sufficiently examined in this sense, or that the training of the models has been carried out with historical data whose quality is not adequate, among other things. In addition, it is essential to consider that biases can also derive from inequalities in access to technology, or from a lack of diversity in design teams. These factors can limit the holistic perspective and understanding of the problems, thus perpetuating prejudices and inequities in AI. It is essential to responsibly address these biases to guarantee ethical, trustworthy and fair AI. It is no less important to encourage diversity in the development and design teams of AI technologies so that different perspectives are reflected which can lead to more inclusive and equitable solutions.

Keywords

Ethical Artificial Intelligence, Gender Bias, Algorithmic discrimination

1. INTRODUCTION

Artificial Intelligence (AI) is at the forefront of the current technological revolution and is a cross-disciplinary field present in any application area we can think of, from medicine, education or marketing to social sciences and humanities. The reasons for this are several. First, we have an immense availability of data (which serves as the fuel for current AI algorithms) derived from the ongoing and intensive process of digitization, within which social change plays a part, as connectivity is essential for all of us, and had revealed so during the recent CoVid-19 pandemic. We also have ever-increasing computing capabilities that enable cost-effective processing of the heterogeneous volumes of data we generate, which have multiplied by over 30 in the last decade, and are estimated to reach 180 zettabytes by 2025, with an average annual growth of almost 40% during the period 2020-25. Advances in software have been undeniable over the last few years, introducing new types of databases that allow us to store and process structured and unstructured data beyond classical scientific data. The emergence of new theoretical developments, primarily

mathematical, like those achieved in the field of Deep Learning, Reinforcement Learning, or Natural Language Processing, has led to high-precision results, as seen for example in the case of Chat GPT [1], establishing AI as a mature technology with significant success and impact. AI is present in many activities in our daily lives, from fingerprint identification in our mobile devices, the map services used to find our route, the suggestions to correct the texts we are typing or the recommendations that we receive in different web services, from purchasing goods to digital contents to consume.

Biases are inherent to human civilization and have existed in every era of our history. As individuals and society, we tend to interpret and judge phenomena based on the standards of our culture, race, gender, religion, etc. [2]. These prejudices lead us to discriminate against other human beings and undeniably have a social and economic impact on certain groups. Technology may appear innocuous in this sense, but unfortunately, it's not the case. Certainly, it can be used to eradicate these biases, but for this, ethical and responsible oversight of algorithms and the data they use is essential to prevent the perpetuation of discrimination, whether intentional or not in the process. Thus, although the design and implementation of AI and other technical systems is not often seen as a technical challenge, in fact, there are decisions to make that imply them, such as their functionalities, the target users, the business model, etc. Although legal requirements, such as those involved in the General Data Protection Regulation (GDPR) [3], are currently checked, ethical principles behind the law are not often required. It might seem that AI can remove subjectivity in favor of decisions that rely more on data, but although this is true, it is also true that preassigned gender roles might be emphasized by AI in several ways [4], from the opportunities to access to the technology, appearance or voices of the applications, the selection of the data used to train the machine learning algorithms or the, even unconscious, prejudices in their coding, among others. For example, virtual assistants such as Alexa, Cortana, or Siri¹ had default feminine voices at their launching, and a design with "submissive" personalities (helpful, intuitive, and cheerful besides intelligent). In contrast, other devices, such as Watson use a masculine voice performing tasks such as teaching and instruction, with authoritarian and assertive personalities. Other stereotypes that can be perpetuated by AI include occupational roles. For example, male robots are commonly employed for security-related jobs while female ones are common for hotel receptionists. Affective labor chatbots (with activities such as caring, listening or comforting) are performed by feminized apps. Thus, the humanization of these virtual assistants might also allow for the dehumanization and objectification of women. Another concern lies in tools that, right from their inception, raise significant ethical questions due to their intended purpose or design. One such example is applications designed for the invasive act of undressing individuals, primarily used with women. These apps, initially affecting

¹ Mark West, Rebecca Kraut, Ei Chew Han. I'd blush if I could. UNESCO-Equals Skill Coalition, 2019. <https://en.unesco.org/ld-blush-if-i-could>

notorious persons (actresses, singers), are now making headlines in Spain, as they have been wielded by teenage men to exploit and blackmail underage girls.

To address the policy on bias, specifically on gender bias, we will need to check the data that is collected, how it is stored and processed, the guidelines followed by the designers and programmers of the algorithms (the great majority of AI engineers are male, and thus the applications featuring female characteristics might reflect their own ideas about women), etc. Algorithms should be audited for discrimination, and the UE has taken a special interest in ethical AI since several years ago. In 2019, the Guidelines for Trustworthy AI were published [5]. In this publication, seven basic requirements for trustworthy AI were enumerated, among which were diversity, non-discrimination and fairness. In 2021, the UE proposed the first regulation on the field, the AI Act [6]. Although the UE has among its priorities equality and non-discrimination, the impact of algorithms [7] on gender equality is a challenge that has started to receive the attention of researchers only recently. In this article, we will describe not only the situations in which AI algorithms can present gender bias or the perpetuation of stereotypes, but also the several barriers that still slow down women in their incorporation into technology.

2. ARTIFICIAL INTELLIGENCE: A BRIEF HISTORY AND CONTEXTUALIZATION

Throughout the history of our civilization, the innate human drive to control our environment has fueled ongoing technological progress. While much of this progress has been gradual, there have been pivotal moments of transformation, such as the advent of navigation, the steam engine, and electricity. In the 21st century, it's becoming increasingly clear that technology, particularly artificial intelligence (AI), is poised to become the "new electricity," as articulated by British researcher Andrew Ng. AI has carried this name since a small group of scientists gathered in 1956 at Dartmouth College (USA) during a summer school to discuss a big question, "Can machines think?". This question had been posed half a dozen years earlier by the British scientist Alan Turing, considered the father of the AI discipline, in an article titled "Computing Machinery and Intelligence"[8]. In this work, largely philosophical because the capabilities of computers in that decade were far from enabling the implementation of his projects, Turing argued the idea that digital computers could exhibit intelligent behaviors and learn. He advocated the possibility that machines could compete with humans in purely intellectual fields. Turing was not only a precursor to the discipline but also a visionary, proposing models of machine learning that are still relevant today, along with other aspects that remain current, such as the possibility of computational creativity, which seeks to model, simulate, or replicate human creativity using computers, and humanized interfaces that enable complex interactions between humans and machines using natural language. These aspects connect Artificial

Intelligence with disciplines in the so-called "Humanities" such as Cognitive Psychology, Language, Philosophy, and Art. Therefore, Artificial Intelligence is an interdisciplinary technology that also spans many areas. Intelligent systems produce a significant percentage of company reports and communications; the customer service channels of many companies are managed by chatbots; we all use digital personal assistants, unlock our mobile devices using our fingerprint, find directions with our mobile apps, use recommendation systems on content, fashion, etc.; and there are intelligent applications in industries, agriculture, tourism, medicine, and education, among other fields. In other words, virtually every sector today is a technological sector. If we look for a definition of Artificial Intelligence in the Oxford Dictionary we will find that it is an area of Computer Science that deals with the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. Some of its most well-known areas are:

- Machine Learning (ML), which trains algorithms and models to learn from data and make decisions on classifying or predicting situations.
- Natural Language Processing (NLP), which focuses on enabling machines to understand, interpret and generate human language.
- Computer Vision (CV), which involves enabling machines to understand and interpret visual information from images or videos, etc.

The current state of AI differs significantly from the initial aspirations of AI pioneers. Initially, the goal was to create a form of AI that possessed general intelligence akin to that of humans, referred to as general AI. However, this objective remains far from realization even today. Instead, we have achieved specific AI, which involves the development of algorithms and machines capable of performing tasks associated with human intelligence within specific and specialized domains. These AI systems can learn, understand, and reason within their designated areas of expertise. For instance, we have seen AI programs excel in games like chess or Go, defeating world champions and grandmasters. Nevertheless, these AI systems lack the versatility of human general intelligence. In contrast, a human chess player can readily apply their knowledge to learn a different game like checkers. Achieving similar adaptability in AI would need the development of distinct algorithms, although there is steady ongoing progress in the field of transfer learning, which aims to bridge this gap.

Since his birth in the late 50s, the discipline has gone through various historical phases. In that first meeting, scientists aimed to replicate brain capabilities using algorithms, and there were two different approaches to building these algorithms. One group of scientists believed that symbolic representation was primarily based on logic and syntax, opting for mathematical solutions. The other group thought it was primarily based on semantics, and since intelligence is too complicated and probably computationally intractable, it couldn't be solved with the type of homogeneous system that requires precise requirements. And

precisely, each of these two approaches has led to the two brightest stages (called springs) of AI, interspersed with the so-called "winters", in which the discipline mostly survives within academic research environments. The first spring arises from the first of these approaches and constitutes what is called symbolic AI, which gave birth to the development of expert systems based on expert knowledge in various fields, such as DENDRAL (which interprets molecular spectra), MYCIN (diagnosis and treatment of infectious diseases), or PROSPECTOR (mineral ore location). These systems make knowledge explicit and provide explanations for their reasoning but often have costly and complex maintenance. The second spring, in the 2010s, emerged with new machine learning algorithms that take advantage of the massive explosion of data and new computing platforms, with highly disruptive algorithms, mainly those based on deep learning, reinforcement learning, etc. These are highly precise systems but lack transparency in that they do not provide a transparent explanation of their reasoning (see Figure 1).

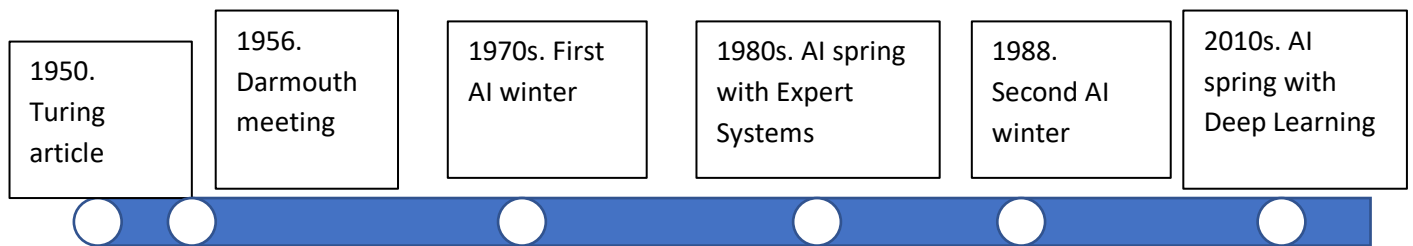


Figure 1. Timeline for Artificial Intelligence springs and winters.

This new AI has shaped itself as a success due to several reasons, which have formed the primordial soup that has made the discipline the most influential in the transition to the so-called Society 4.0. One of the most important factors is the tremendous amount of data we have today, which serves as the primary fuel for AI algorithms. We are immersed in a progressive and intense process of digitization that is reaching all personal and business areas. Personal experiences are becoming increasingly digital, and we have a large number of sensors continually recording data about how our environment behaves, as well as sensorizing almost every industrial process we can think of. Also, connectivity is an essential value, increasingly shaping a hybrid world in which the physical and digital coexist. This presents a significant economic opportunity, as seen in the importance of connectivity for individuals and businesses and the shift to the teleworking model. However, it also poses other regulatory and social challenges, some of which are related to the role of women taking on teleworking at home and caring for their families. Currently, Spain has launched the 5G network, which represents a leap towards hyper-connectivity, allowing for higher speed, greater bandwidth, and very low latency, with the capacity to connect millions of devices. The consulting firm Gartner estimates

that by 2020 we will have around 2.8 billion devices connected to the Internet. This new protocol will make interactions possible in a matter of milliseconds, facilitating tasks as simple as downloading movies on platforms or viewing 360° sports events, up to autonomous cars, intelligent city management, or real-time remote surgeries. The goal is to achieve complete digitization of companies (especially SMEs), public administrations, and citizens.

Another influential factor in AI progress is having the necessary computing power, thanks to the decreasing cost of cloud computing, the availability of new parallel and distributed computing platforms, and the significant advances in high-performance computing technologies. This allows for fast and economically viable processing of huge volumes of heterogeneous data (text, images, videos, etc.) that are generated at high speed (the so-called Big Data). Finally, there are significant software advancements, including new types of databases that can store structured heterogeneous data (those with defined length, format, and size) and unstructured data (without a specific format, stored in various formats like multimedia files, PDF, email, Word, etc.), beyond classic scientific data. We also have algorithms that have been highly disruptive, achieving very high precision results in many complex fields, often superior or comparable to human performance. The result of all this is a successful AI with a significant impact on the economy.

But AI also entails social changes, which should be confronted in a context of rapid pace and breadth. Because despite all the successes of AI, there are still many important open research directions and some identified problems that need solutions. Some of them are related to algorithm biases. Bias is an inherent aspect of human decision-making, shaped by cultural backgrounds and societal norms, which vary among individuals, leading to the formation of preferences and prejudices. This bias is deeply ingrained in our cognitive processes, making it challenging to detect and necessitating educational reforms for change. While AI offers the promise of quicker, ostensibly unbiased decision-making driven by data, it has proven not to be entirely impartial. This might happen for different reasons, for example because certain use cases have not been considered, or because when using real-world data, they end up incorporating gender, race, or religious biases implicit in them.

At this time when technology is a dominant force shaping our world, digitalization is an unstoppable trend, and the recent pandemic has underscored the importance of medium to high-level digital skills in most jobs, there's a substantial risk that women may be left behind. This risk is not only due to the gender biases mentioned before that might be present on AI tools, but also due to the significant gender gap in access to technology education. As nearly every sector undergoes a technological transformation, and with the

European Union (EU) projecting that approximately 90% of jobs will require advanced digital skills in the near future, addressing this issue is imperative. Additionally, the underrepresentation of women in tech, especially within fields like AI and ML (one of its most successful subareas), means that most innovations are currently developed by predominantly male teams, often with higher economic status and of white ethnicity.

Consequently, it is imperative to detect and eliminate biases that are present in the AI algorithms and tools, some of which can perpetuate gender stereotypes and inequalities, but also to work on education and employment, where gender bias is also present. To foster a more inclusive and equitable technological landscape, it is essential to work towards solutions for these challenges.

3. SOME BARRIERS: ACCESS TO TECHNOLOGY, EDUCATION AND EMPLOYMENT

Digitalization is one of the processes leading to technological change in the world. In the EU the European Commission has called to strengthen our position to be leaders in technology in the near future, with an investment of 250 billion euros to digitalization in Next Generation EU funds, and a goal of having 80% of our population have acquired digital skills by 2030². However, in a study funded by UNESCO [10], women have a 25% lower likelihood than men to possess basic digital technology skills, are 4 times less likely to be proficient in computer programming, and are thirteen times less likely to submit an ICT patent application. Thus, one important aspect to deal with is the need to foster access and education in technology for girls.

The limited participation of women in careers and studies related to science, technology, engineering, and mathematics, commonly referred to as STEM, threatens to leave us women behind. In the workplace, the number of women employed in the sector also reflects a significant inequality. Some significant figures are that 6% of professional software developers are women, and if centering on the AI field, 12% of researchers in AI and 22% of employees in technical roles in major ML companies are women globally, with the percentage descending as going higher in the responsibility of the role. In the UE, there is still a persistent gender gap: only one in five ICT specialists and ICT graduates are women, which may affect the way digital solutions are devised and deployed [9]. This is compounded by the demographic decline across the EU, and a lack of specialized education offered in key digital areas. No wonder that with these reduced percentages, in the EU and other developed areas of the world, around

² https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age_en.

14% of Computer Science students are women. Specifically in Spain, it is the lowest percentage among all STEM disciplines, although it has increased very slightly during the last 5 years.

Let us draw our attention to the UNESCO-EQUALS report [10] on their analysis of what they refer to as the gender equality paradox (see Figure 2). Despite the awareness in many countries, the gender digital gap seems to be widening, even though we have seen nearly two decades of interventions in many nations. This divide deepens as technology becomes increasingly sophisticated and expensive, enabling more transformational and impactful uses.

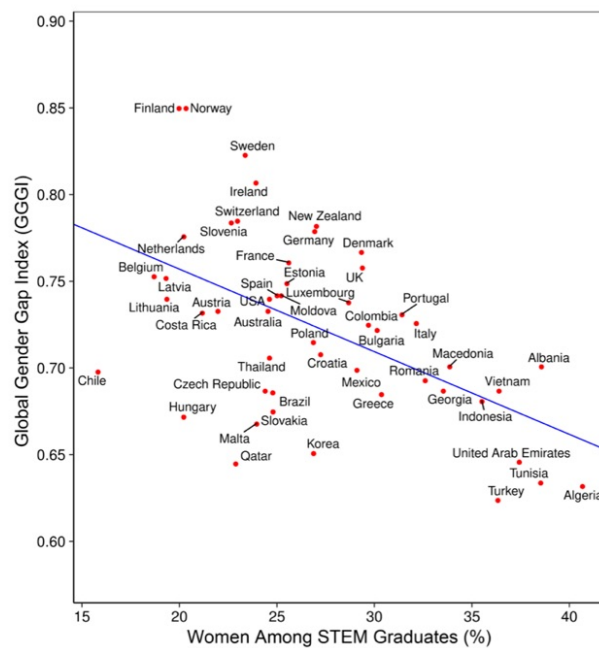


Figure 2. The Gender Equality Paradox. The efforts for gender equality in Western countries do not increase the STEM vocations in girls

As depicted in Figure 2, countries that are closer to achieving gender equality in general (the x-axis is the percentage of women acquiring the advanced skills required for careers in the technology sector, and the y-axis represents the percentage of women graduating in Information and Communications Technology - ICT), have lower percentages of women STEM graduates. Conversely, countries with low levels of gender equality, such as the Arab countries, have the highest percentage of women pursuing advanced technology degrees. To provide a more concrete example, in Belgium, only 6% of ICT graduates are women, whereas in the United Arab Emirates, this figure stands at 58%. This paradox highlights that despite increased awareness in many countries, the gender digital gap appears to either remain stable or widen. This trend persists, despite interventions spanning at least two decades across

numerous nations. This divide deepens as technology becomes increasingly sophisticated and expensive, enabling more transformational and impactful uses.

The reasons for this situation are various and complex, from historical to educational ones. At the end of World War II, which marked the beginning of modern computing, programming was considered a relatively low-importance task, and it was predominantly women who filled these roles, which were crucial for achieving the progress we see today. As computers became integrated into all aspects of daily life, programmers began to accumulate influence and prestige. Strangely and almost imperceptibly, women gradually reduced their presence in the field in favor of men. Before the era of personal computers, computer science students were almost evenly split between men and women because the industry was very new. This situation changed with the introduction of computers into households. A statistic from the 1990s in the United States revealed that boys were twice as likely to receive a computer as a gift compared to girls. Parents tended to place computers in boys' rooms rather than girls. Moreover, it's also evident that parents are more inclined to encourage boys rather than girls to learn computational skills. All these factors have solidified the shift toward a stereotype of a male professional in just one generation.

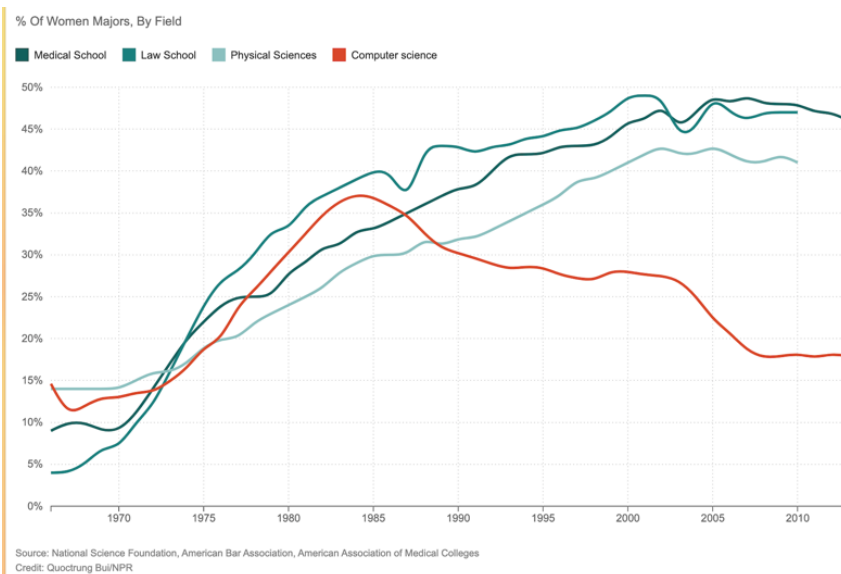


Figure 3. Percentage of women students across time and several specialties in the U.S.A.

Furthermore, as previously mentioned, the lack of diversity in the teams designing technology, where gender is just one aspect to consider, means that the tools being developed often fail to consider the perspective of women regarding the future world. In some cases, this perpetuates traditional female roles characterized by inequality and submissiveness. In even worse cases, we are witnessing the development of potentially very harmful tools, as is the case of the scandals that have occurred with applications that create false images,

and that on some occasions have been used against famous women by disseminating false images of their nudes on the Internet.

But gender equality in technology is not only a matter of justice but also of the economy. An analysis carried out by McKinsey reveals a projected tech talent shortage of 1.4 million to 3.9 million individuals by 2027 in the EU-27 countries. If Europe were to double the representation of women in the technology workforce to approximately an in principle achievable 45% (around 3.9 million women more by 2027), it could effectively bridge this talent gap and potentially experience a boost in GDP ranging from €260 billion to €600 billion [11]. In this report, the authors have identified that there are two critical points in which there is a notable decline in the proportion of women pursuing STEM disciplines: the first, during the transition from primary and secondary education to university, resulting in an 18-percentage-point decrease; and the second, during the shift from university to the professional workforce, leading to an additional 15-point drop. Regarding gender balance in tech companies, including social networks, is relatively equal, but the representation of women in technical positions like developers and data engineers is significantly lower, as said above. Finally, and a worrying fact, and is that this is expected to worsen, as the graduation rate of women in STEM fields during higher education is on a downward trend according to Eurostat data. If this trend persists, the percentage of women in tech roles in Europe is projected to decrease to 21 percent by 2027, instead of increasing. And not only this rate is decreasing, but also half the women in the technology field leave the industry at the midpoints of their careers (more than doubling the men rate) for lack of support and good opportunities. Thus, several interventions need to be done, such as creating a less isolated environment for women in tech, improving flexibility at work, improving retention rates and trying to encourage girls into technological disciplines earlier in the educational process.

It is also quite striking that the decline in female enrollment coincides with the rise in job opportunities and salaries in the profession, not to mention that both in Spain and across Europe, there aren't enough graduates to meet the demands of the market. But why are women not interested in entering a rapidly growing sector with virtually no unemployment, above-average salaries, and a significantly lower gender pay gap? Why is this drop in female enrollment in Technology and Computer-related careers, fields that will shape our future world? The reasons are manifold and can be found in various areas, including social, family, and educational contexts, as mentioned above.

We need to reconsider how we educate our sons and daughters from an early age because several studies indicate that 90% of girls aged 6 to 8 believe that engineering requires skills typically associated with males. There exists a certain technophobic image associated with women in society, which perpetuates gender roles that should have been eradicated by now and influence what girls aspire to become in the future. We must promote the study of

technology-related subjects starting from primary education levels. Society still harbors numerous stereotypes. Both education and media reinforce distinct aesthetic and behavioral standards for males and females. Attitudes toward life that are seen as positive and desirable for males, such as professional ambition, are often viewed negatively and criticized when exhibited by females. Another influential factor is the lack of role models and the limited visibility of women in technology, despite their significant contributions to computing, telecommunications, and STEM fields in general. While 5% of 15-year-old boys consider pursuing a career in ICT, only 0.5% of girls entertain this possibility. However, as previously mentioned, women have made substantial contributions to ICT, starting with its inception. Computing wouldn't be what it is today without the contributions of women, many of whom remain unfamiliar to the public. For instance, we can name a few, such as Ada Lovelace, who was the first person to define a general-purpose programming language, the current programming language "Ada" is named in her honor; Grace Hopper created the first compiler for a programming language; the famous Hollywood actress Hedy Lamarr played a pivotal role in developing the precursor technology for Wi-Fi, Bluetooth, and GPS; Evelyn Berezin invented the office computer and is considered the mother of word processors, while Margaret Hamilton developed the code critical to the first moon landing. In Spain, Ángela Ruiz Robles, a teacher in a village near Ferrol, pioneered the e-book with her mechanical encyclopedia patented in December 1949, even though most of the world credits American Michael Hart's Gutenberg Project in the 1970s. Many other women have been and continue to be significant in the field, although, regrettably, they remain relatively unknown to society. We must highlight the work of women in scientific and technological disciplines, revising educational texts, which feature very few female figures, with an average representation of only 7% across all fields, according to several studies in Spain and other countries [12, 13]. It's an urgent necessity that we reform the educational system, introducing technology from an early age so that our boys and girls feel vocally drawn to the discipline and perceive that they possess the same capacity to pursue a career in the sector. To increase girls' access to technological disciplines, we may need to change our approach to education, while also conducting parallel efforts in family education and addressing the messages perpetuating sexist roles in the media.

4. ALGORITHMIC DISCRIMINATION: SOME USE CASES IN AI TOOLS

As said, AI is deeply integrated into numerous applications and tools that are an integral part of our daily lives. It assists in determining who gets selected for a job, who receives a loan from the bank, or which is the diagnosis for an illness, for instance – all matters and decisions that were traditionally made by human experts. However, many of these tasks that were once performed by individuals can now be automated, with the advantage that machines can analyze vast

amounts of data more quickly and accurately than humans, and with fewer errors. However, ML bias might arise when an algorithm delivers systematically biased results because of erroneous assumptions, although unconscious, of the ML process. This is mainly because ML algorithms rely on real-world data, data that inevitably contain the biases present in the world we live in. Thus, the AI algorithms should be consciously reviewed, and these biases should be addressed in our algorithms, so to avoid these situations.

Gender bias is one of these possible unfair situations, but others as race, religion, etc. are also possible. One example is the algorithm used in USA hospitals that predicted which patients required additional medical care in 2019, and which favored by a considerable margin white patients. It was found that this was so because it was using the expense of healthcare as a parameter reflecting healthcare needs, but black patients with similar diseases spent less than white ones [14]. Once the bias was detected, researchers reduced it by 80%. Another system used in the USA for parole decisions was biased against black males. Gender and race bias are also present in most facial recognition systems, which perform better for male and white users [15]. A well-known example is Amazon's hiring system, which failed to select women for tech positions. The software was to assign ratings to job applicants, spotting similarities in the applications. As most candidates were male, the algorithm learned to prefer them. Although the company changed the program to be neutral on this matter, other biases occurred, and thus the recruiters started to use the program rating as just a suggestion [16], until eventually the project was closed. Google's online job advertising system predominantly displayed high-paying jobs to men [17], and as a final example, if one were to search Google images using the word "CEO", only 11% of the images were women, which in fact account for around 27% in the USA. All these issues appeared because the learning systems considered historical data, thereby perpetuating biases in the data used. These are real-world examples that are often selected by the engineering teams developing the system, who, in these instances, used more examples of white males with high-tech mobile devices, more black males with criminal histories, or fewer women with technology-related curricula. This led the system to build a model of the world based on these data, a model that perpetuates a history of bias and discrimination. The lack of transparency and the complexity of the learning models used make it even more challenging to detect biases. Because of that, transparent and explainable AI are key aspects of an ethical AI.

Diversity in development teams could be a relevant aspect to help identify such situations. A compelling demonstration of this is that many leading scientists in the field of detecting and addressing biases in intelligent algorithms are women. Researchers like Susan Leavy [4] are focusing on identifying specific challenges to ensure that the ML process does not become an amplifier of the gender biases present in language and written texts. Specifically, Leavy raises concerns about the inadequate incorporation of feminist linguistic theory into the ML process, especially in systems that rely on data extracted from texts, such as

some recommendation systems or social media data collection systems. This linguistic theory has centered on studying how gender ideology, conveyed through language and writing, influences both the perception of men and women and the associated expectations of behavior for each group. Such incorporation is necessary to prevent gender biases in the ML process. Among the gender biases in language that could affect the machine learning process relying on linguistic data and texts, Leavy identifies these five in particular:

- Naming practices, which sometimes use distinct terms to refer to men (e.g., "family man") and women (e.g., "single mother") without equivalent expressions for the other gender.
- Preference in word order, often favoring the male gender (e.g., "sons and daughters," "fathers and mothers," "husband and wife," etc.).
- Biased descriptions, such as generic references to a profession based on whether it is predominantly male or female (e.g., chairman, fireman, ombudsman, etc.).
- Use and types of metaphors, with descriptions of males often focusing more on their behavior than their appearance or sexuality, in contrast to descriptions of females, where metaphors are more frequent and derogatory.
- The degree of presence/absence of women in written texts is significantly lower than that of men, especially in certain fields such as business or engineering.

These biases persist in tools and applications that use these datasets. Let's examine the case of "word embeddings," one of the most commonly used structures in ML and NLP tasks (for example, for automatic translators). Word embeddings are employed to represent text data as vectors. In a study conducted by Tolga Bolukbasi and colleagues [18] it was demonstrated that embeddings trained on Google News articles exhibited a significant number of stereotypes. In their research group, they have employed methodologies to modify these embeddings to eliminate stereotypes, such as removing the association between words like "receptionist" and "woman," while maintaining desired associations like "queen" and "woman." This approach reduces gender bias while preserving the advantages of the formulation. Bolukbasi demonstrated that one of the most widely used word embedding spaces, Word2Vec, could encode gender social biases. To illustrate this, he used it to train an analogy generator that would complete missing words in phrases. For example, the analogy "man is to king as woman is to x" yields $x = \text{queen}$. However, it also reveals implicit sexism in texts. For instance, in the case of "man is to programmer as woman is to x," it filled in $x = \text{homemaker}$, or "father is to doctor as mother is to x" yields $x = \text{nurse}$.

This issue represents a growing area of research in ML, with specific characteristics depending on the language being used. Languages derived from Latin, for instance, lack neutral terms, whereas English does have them ("You are very tall" can refer to either gender, while in Spanish, we would need to specify

"Eres muy alto (masculine)" or "Eres muy alta (feminine)"; "The doctor" in Spanish could be "El doctor (m.)" or "La doctora (f.)". In the case of automatic translators, like Google Translate, a few years ago, only the masculine translation was provided, but now both alternatives are displayed. In the world's most important Computational Linguistics conferences, sections and workshops dedicated to bias in language processing are already organized, and some companies, like Google, allocate specific funds to this topic.

In other work [19], the researchers conducted a similar investigation in the field of facial recognition systems evaluating three commercial classification systems (Microsoft, IBM, and Face++). In all cases, the performance was better for men than for women (with differences in error ranging from 8.1% to 20.6%), and better for light-skinned individuals than for dark-skinned individuals (with differences in error ranging from 11.8% to 19.2%). As a contribution, the researchers have provided a new dataset suitable for this area, balanced in terms of both gender and skin color. Since their article, consequent progress has been made for these systems.

It might seem that these algorithms are not particularly harmful, but let's imagine the consequences if they are part of an intelligent system that examines our skin to diagnose whether we have a disease or are the voice recognition system (these also perform better for men) in an autonomous car, for example. This is why we need algorithms that are more transparent, explainable, and auditable. Furthermore, these algorithms should provide information about the demographic and phenotypic characteristics of the data used for training to address potential biases.

Another example that highlights inadvertently introduced biases is that of intelligent personal assistants. This area is relevant as an example because this technology is rapidly entering the consumer market, and it does have a gender perspective. The most common assistants are voice-based assistants like Cortana, Siri, Alexa, and Google Assistant, which represent 90% of the market in terms of user volume and frequency of use. Voice searches began in 2008, increasing on the internet by a factor of 35 in 10 years. Currently, in 2023 more than 1 billion voice searches take place every month, with more than 50% of the adults reporting that they use voice search daily. To put this into context, it took mobile phones, which are now ubiquitous, around 30 years to reach this level of prevalence. Increasingly, human-computer interaction will be hands-free and voice-driven. Most of these assistants were initially designed as synthetic young women's voices or are set as such by default, although most now offer a male/female option. Companies justify this female selection based on consumer preferences. As an example, Siri was exclusively a female voice at its launch in 2011. In 2013, it became the default female voice (with the option of a male voice), but it defaults to a male voice if the user selects Arabic, British English, Dutch, or French. Several studies suggest that this preference is rooted in traditional social norms since automatic personal assistants are associated with adjectives like

"helpful" and "humble," aiming to make us feel like "bosses," which are stereotyped attributes for women. This association aligns with the roles of women in video games (assistants to a male central character) or the typical roles they often have in TV series and movies.

Alternatively, there are situations where male voices are used by default to provide instructions and directions in a navigation system. This is because in some countries, like Germany, there were complaints about female navigation assistants, with users stating that they didn't want to take orders from women. In other cases, such as in Japan, female voices in stock market assistants provide data, but male voices facilitate and confirm chosen transactions. When IBM Watson won the game Jeopardy in 2011, its voice was unmistakably male. From their launch to nowadays, many things have changed. Siri, Google Assistant, Cortana or Alexa do not acknowledge gender when asked, and some of the responses that they give to sexual questions from the users (see UNESCO report [10]) have been changed. This reinforces the fact that if we do not change the surveillance of AI applications, we may inadvertently perpetuate a pattern of perceiving women as caregivers and men as decision-makers, reinforcing sexist roles.

In Table 3.1, we can see some of the characteristics of the most used commercial assistants in the world. As you can observe, all of them initially featured a female voice.

	SIRI	CORTANA	ALEXA	GOOGLE ASSISTANT
Release date	October 2011	April 2014	November 2014	November 2016
Release voice	Feminine	Feminine	Feminine	Feminine
Masculine voice date	June 2013	November 2019	July 2021	October 2017
Feminine by default in most countries	Yes	Yes	Yes	Yes
Masculine by default	Only if operating system language is Arabic, German, Dutch or British english	No	No	No
Personality	A sense of assistance and camaraderie, brave without being sharp, joyful without being caricatural.	Supportive, helpful, kind, empathetic	Intelligent, humble, occasionally humorous	Humble, helpful, occasionally a bit playful

Table 1. Some characteristics of most common voice assistants

Obviously, the companies and in general the AI community should address these problems of bias from a technological and ethical perspective. However, due to the significant implications these systems may have on our individual rights, governments need to implement regulations in their development and usage. In this regard, as said above, the UE began addressing these issues as early as 2019 with the publication of the Ethical Guidelines for an Ethical and Trustworthy AI, which were further solidified in 2021 and 2022 with the AI and the Data Acts, being the first discussed during these last months by the European Parliament [3,5,6,20]. The EC has also released a white paper on AI [21], acknowledging that the growing reliance on algorithms in Europe brings about specific risks concerning the protection of fundamental rights, particularly in relation to equality and non-discrimination. These risks are also recognized in the Commission's recent Gender Equality Strategy for 2020-2025, which acknowledges that AI has the potential to exacerbate gender inequalities [22]. In response, the EU has advocated for the establishment of a "trust ecosystem" that insists on European AI being firmly rooted in EU values and fundamental rights, with the right to equality and non-discrimination being of central importance. This commitment is necessary to ensure fair, explainable, and auditable technology that, to the greatest extent possible, guarantees a more equitable future for everyone.

5. CONCLUSIONS

AI is not at all immune to algorithm discrimination and gender bias, which potentially can even widen the already existing gender gap. Thus, the algorithms must be developed in a way that allows the identification and filtering of biases in the datasets and the models employed. Some of these biases, as we have seen above, are computationally identifiable and solvable using strategies that consider the context and avoid the perpetuation of biases. As a recommendation, it would also be advisable for AI systems to be developed transparently, to be able to detect and arrange these complex problems more easily.

To summarize, six major challenges AI algorithms pose to gender equality [23], and that we have described in the examples seen in the previous section:

- The human factor and the challenge of stereotypes and cognitive bias. This refers to how the biases, stereotypes, and prejudices that people naturally have could impact the algorithms they create, which may end up making biased or unfair decisions also because they have learned from biased human data.
- The data challenge is about recognizing that the information used for training algorithms often reflects historical patterns of discrimination deeply embedded in our society. When algorithms are trained with this biased data or with data that are incorrect, not representative, or

unbalanced, they might end up perpetuating the existing inequalities and biases that are built into the data.

- The correlation and proxies challenge. Correlation is about how algorithms can mistakenly treat connections or patterns as if they are causes. For example, if women historically received more negative evaluations for their work (not because they performed worse, but due to bias), an algorithm might think that being a woman causes poor performance. It can then make decisions based on this incorrect idea. The proxies challenge is based on the fact that removing certain characteristics (such as gender) from the input variables is not enough to remove the bias, as the algorithm can find other clues (proxies) that can indirectly reveal these characteristics. So, even if one hides certain traits, the algorithm might still figure them out and make biased decisions based on these hidden clues. As an example, imagine an algorithm designed to predict creditworthiness for loans. To avoid bias, it's programmed not to consider gender. However, it still finds a proxy: it notices that applicants who belong to certain women-dominated professions (like nursing or teaching) tend to have lower credit scores. As a result, it could unfairly deny loans to individuals in those professions, most of whom happen to be women, even though their creditworthiness is just fine. This demonstrates how algorithms can detect and use proxies for protected characteristics, leading to unintended biases.
- Transparency and explainability. The ML algorithms are more accurate each day, but most times at the expense of being complex black boxes, in the sense that it is difficult to know in detail how they work internally, especially if code and data are proprietary. Thus, it is complicated to prove them for gender discrimination. These two factors are among the ones recommended for a trustworthy and ethical AI.
- The scale and speed challenge describes how algorithms can spread their biased decisions at a larger and faster rate than human discriminations. The good news in this challenge is that bias can be corrected also at a much faster rate in machines than in humans.
- The responsibility, liability and accountability challenge, which revolves around the problem of identifying who should be held responsible, accountable, or liable when discrimination occurs due to complex relationships between humans and machines, and the many different parties involved in creating, selling, and using algorithms, making it hard to pinpoint who's ultimately to blame for any discriminatory outcomes.

It is essential to try to reduce the gap in access to technology and STEM education for women and increase diversity and the presence of women among those who develop algorithms and conduct research in the field of AI. While this may not automatically guarantee the solution to gender bias, diversity within teams can help identify biases that might otherwise go unnoticed. Additionally, it would be recommended to establish minimum ethical standards that

organizations and companies must adhere to, ensuring the protection of human rights, including gender equality, in the face of various potential situations and risks posed using AI.

Finally, governments must take responsibility for implementing regulations and policies that protect human rights from the undesirable effects of AI tools. In this regard, the EU has proposed several regulations that impose restrictions for certain algorithms that might pose risks for humans, among which gender inequalities are. But in this age of rapid technological advancement, we need to keep pace with the implementation of regulations and policies. The choices we make today will shape the future of AI, and by extension, our society. The true power of innovation lies not in what we create, but in the impact it has on the lives of every individual, regardless of their gender.

BIBLIOGRAFÍA

- [1] Chat GPT. Open AI, <https://chat.openai.com>
- [2] González de la Garza, Luis M. Teoría de sesgos en el sistema educativo de la democracia del siglo XXI Nuevas garantías para la libertad de pensamiento, el “Derecho a no ser engañados”. *Revista de Educación y Derecho*. Nº 22, Abril - septiembre 2020.
- [3] Regulation 2016/679, General Data Protection Regulation, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.
- [4] S. Leavy, "Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning," 2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE), Gothenburg, Sweden, pp. 14-16, 2018.
- [5] High-Level Expert Group on Artificial Intelligence. Ethical Guidelines for Trustworthy AI. April, 2019.
- [6] Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts [https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI\(2021\)698792_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf)
- [7] European Commission, Directorate-General for Justice and Consumers, Gerards, J., Xenidis, R., Algorithmic discrimination in Europe – Challenges and opportunities for gender equality and non-discrimination law, Publications Office, 2021, <https://data.europa.eu/doi/10.2838/544956>
- [8] Turing, Alan M. (1950) Computing Machinery and Intelligence. *Mind* 49: 433-460. <https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf>
- [9] <https://digital-strategy.ec.europa.eu/en/policies/desi>
- [10] West, Mark, Kraut, Rebecca, Ei Chew Han. I'd blush if I could: Closing gender divides in digital skills through education. UNESCO and EQUALS Skill Coalition, 2019. <https://unesdoc.unesco.org/ark:/48223/pf0000367416>
- [11] Blumberg, Sven, Krawina, Melanie, Mäkelä, Elina and Soller, Henning. Women in tech: The best bet to solve Europe's talent shortage. McKinsey Digital, January, 2023.
- [12] Torre de la Sierra, Ana M., and Guichot-Reina, Virginia. The influence of school textbooks on the configuration of gender identity: A study on the unequal representation of women and men in the school discourse during the Spanish Democracy, *Teaching and Teacher Education*, Vol 117, 103810, 2022.
- [13] Sunderland, J. New understandings of gender and language classroom research: Texts, teacher talk and student talk *Language Teaching Research*, 4 (2) (2000), pp. 140-173, 10.1177/13621688000400204.
- [14] Obermeyer, Ziad; Powers, Brian; Vogeli, Christine and Mullainathan, Sendhil. Dissecting racial bias in an algorithm used to manage the health of populations. *Nature*, Vol. 36, Nº 6464, pp 447-453, 2019.

- [15] Turner Lee, Nicol; Resnik, Paul and Barton, Genie. <https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>, 2019.
- [16] Martin, Kirsten. Ethics of Data and Analytics, Auerbach Pub, 2022.
- [17]Merill, Jeremy B. <https://themarkup.org/google-the-giant/2021/02/11/google-has-been-allowing-advertisers-to-exclude-nonbinary-people-from-seeing-job-ads>, 2021.
- [18] Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh and Kalai, Adam . Man is to Computer programmer as Woman is to Homemaker? Debiasing Word embeddings. 30th Conference on Neural Informations Processing Systems (NIPS 2016).
- [19] Raji, Inioluwa Deborah; Gebru, Timnit; Mitchell, Margaret; Buolamwini, Joy; Lee, Joonseok, and Denton, Emily. Saving face: Investigating the Ethical concerns of facial recognition auditing. AAI/ACM AI Ethics and Society Conference, pp 145-151, 2020.
- [20] Document 52022PC0068. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonized rules on fair access to and use of data (Data Act). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN>, 2022.
- [21] European Commission, 'White Paper on Artificial Intelligence: A European approach to excellence and trust' COM (2020) 65 final, 2020.
- [22] European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions. "A Union of Equality: Gender Equality Strategy 2020-2025" COM (2020) 152 final, 2020.
- [23] Gerards, Janneke and Xenidis, Raphaëlle. Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law. Special report. European Commission, 2020.

