





Gender and Artificial Intelligence: Challenges and Imminent Possibilities

Mar Souto-RomeroUniversidad Rey Juan Carlos de Madrid (España) ✉ **Mario Arias-Oliva**Universidad Complutense de Madrid, España ✉ **Kiyoshi Murata**Centre for Business Information Ethics, Meiji University (Japan) ✉ **Orlando Lima Rua**Center for Organisational and Social Studies (CEOS.PP), Porto School of Accounting and Business (ISCAP), Polytechnic of Porto (P.PORTO); and Research Center of Business Sciences (NECE), University of Beira Interior (UBI) ✉ <https://dx.doi.org/10.5209/infe.98150>

Recibido: Octubre 2024 • Evaluado: Noviembre 2024 • Aceptado: Diciembre 2024

Abstract: Introduction. The present work addresses gender-related aspects in the development of Artificial Intelligence (AI). It begins by conceptualising AI and situating its context and historical development, as a preliminary step to describing the current and future gender issues related to AI. Objective. To define the main areas of study in gender and AI, including the unequal participation of women in the sector, as well as data and algorithmic biases that may lead to gender biases in outcomes. Methodology. This study has employed qualitative content analysis, given that the topic is emerging and there are no established theoretical or paradigmatic frameworks on which to base other types of research. Results. The study outlines key areas to consider for achieving responsible and ethical AI development, incorporating gender aspects into the future development of this technology. Findings. The study describes the main existing problems and highlights future challenges to address gender biases in AI. These challenges include explainability, the responsible and ethical development of AI systems, and accountability. Additionally, other challenges are presented, such as enhancing the education of women in STEM fields and their incorporation into the AI industry to avoid gender biases in algorithm development teams.

Keywords: gender; information and communication technologies; artificial intelligence.

[es] Género e Inteligencia Artificial. Retos y posibilidades inminentes

Resumen: Introducción: El presente trabajo presenta los aspectos relativos al género en el desarrollo de la Inteligencia Artificial. Comienza conceptualizando la IA, y situando su contexto y desarrollo histórico, paso previo para poder describir los actuales y futuros retos de género en la IA. Objetivo. Definir las principales áreas de estudio en género e IA, entre las que se encuentran la desigual participación de la mujer en el sector, así como los sesgos de datos y algoritmos, que hacen que los resultados puedan presentar sesgos de género. Metodología. El presente estudio ha realizado un análisis de contenido cualitativo, ya que la temática es emergente, y no existen todavía marcos teóricos y paradigmáticos asentados sobre los que realizar otro tipo de investigaciones. Resultados. Se presentan las principales áreas a tener en cuenta para lograr un desarrollo responsable y ético de la IA, integrando los aspectos presentados de género en el desarrollo futuro de esta tecnología. Aportación. Se describen los principales problemas existentes, presentado los aspectos a tener en cuenta en el desarrollo de la IA para evitar los sesgos de género, entre los que destacamos al explicabilidad, el desarrollo responsable y ético de los sistemas de IA, y la rendición de cuentas. Se presentan además otros retos, como la potenciación de la educación de la mujer en ámbitos STEM o su incorporación a la industria de la IA para evitar los sesgos de género en los equipos de desarrollo de algoritmos, así como el desarrollo de técnicas que permiten evitar los sesgos en el entrenamiento de conjuntos de datos y algoritmos.

Palabras clave: género, tecnologías de la información y la comunicación, inteligencia artificial.

Table of contents: 1. Introduction: conceptualisation, context and historical development of Artificial Intelligence. 2. Gender perspective in AI. 2.1. Gender in the development of AI. 2.2. Gender biases in data used by AI. 2.3. Biases in the algorithms used by AI. 3. Challenges of gender and AI. 3.1. Enhancing gender diversity in the development of AI at all levels. 3.2. Enhancing technological systems for the correction of data and algorithmic biases. 3.3. Enhance the explainability of AI systems. 3.4. Enhance responsible development and accountability in AI development. 4. Conclusions. 5. Limitations and future lines of research.

How to cite: Souto-Romero, M.; Arias-Oliva, M.; Murata, K.; Lima Rua, O. (2025). Gender and Artificial Intelligence: Challenges and Imminent Possibilities. *Investigaciones Feministas* 15(2), 385-394. <https://dx.doi.org/10.5209/infe.98150>

1. Introduction: conceptualisation, context and historical development of Artificial Intelligence

In order to address the gender aspects of Artificial Intelligence (AI), it is first necessary to conceptualise it, contextualise it, and know the basic milestones that have marked its contemporary historical development. Reviewing these aspects in the introduction is fundamental to subsequently understanding some imminent challenges and possibilities.

Following Ekmekci and Arda (2020), the word “artificial” implies that it does not exist naturally and requires the intervention of human beings for its creation. Therefore, the term “artificial” alludes to that which is synthetic, imitation or not real. It is used to describe manufactured things, such as artificial flowers, that are similar to the real thing, but lack its innate characteristics (Lucci and Kopeck, 2016). On the other hand, “intelligence” is defined as the ability to acquire and apply knowledge. A broader definition refers to the skilful use of reason, the act of understanding, and the ability to think abstractly, as measured by objective criteria. Therefore, “Artificial Intelligence” (AI) refers to an entity created by humans that possesses the ability to understand and assimilate knowledge, reason using that knowledge and even act accordingly (Ekmekci and Arda, 2020: 17).

Another definition sees AI as the development of computational systems capable of autonomously performing tasks that, before the advent of AI, required the intervention of human intelligence for their execution. This includes visual perception, speech recognition, decision making, or translation between languages (Russell and Norvig, 2020). These systems attempt to emulate or surpass human cognitive capabilities through the use of algorithms, computational models and large amounts of data (Nilson, 2009). Kurzweil conceptualises AI as the art of creating machines that are capable of performing tasks that require intelligence when carried out by humans (Kurzweil, 1990). Raphael (1976), in the same vein, defines AI as the science of making machines do things that would require intelligence if performed by humans.

Since the Stone Age, some 2.6 million years ago, humans have been making a range of man-made artefacts, from simple tools for hunting or working the land to complex industrial machines. But thus far, in all the artefacts invented, there has always been human control over these artefacts, whose skills and capabilities, designed by humans, have always been predictable and controllable (Ekmekci and Arda, 2020). AI represents a paradigm shift, as control over artefacts has already changed, and they have the possibility of self-improvement, randomness and creativity, autonomously and without human intervention. These three characteristics (self-improvement, randomness, and creativity) are the ones that make us overcome the existing concept of computers, which until now were considered as systems capable of solving unsolvable codes for humans. With the birth of AI, we are moving on to a more complex concept, since they are capable of imitating human behaviour, behaving in a way that would qualify as intelligent if a human were to act in that way (McCarthy et al., 2006).

The evolution of AI has allowed the development of different types of Artificial Intelligence: weak, and strong or general Artificial Intelligence (AGI). Weak AI refers to artificial intelligence systems designed and trained for a specific task. These systems are capable of performing specific tasks efficiently but lack the capacity for general understanding or awareness (Russell and Norvig, 2020). In contrast to weak AI, strong or general AI refers to systems with the ability to perform any intellectual task that a human being can do, but without necessarily involving consciousness or subjective experiences (Goertzel and Pennachin, 2007). Thus, weak AI focuses on specific tasks, which can be very complex, such as autonomous driving of a vehicle; whereas strong or general AI seeks a broad intellectual capacity comparable to that of a human and may even possess the presence of consciousness and subjective experiences, going beyond a mere intellectual capacity (Wang and Goertzel, 2012).

Although we will later analyse the contemporary evolution of AI, we consider it necessary in these first paragraphs devoted to conceptualisation to define *Large Language Models* (LLMs), the generative AI systems that have represented the mass dissemination of AI. Generative AI refers to machine learning algorithms designed to generate new content, such as text, images, audio or video, that are similar, but not identical to the data on which the model has been trained (Goodfellow et al., 2020). LLMs are a subcategory of generative AI that has been noted for its ability to generate high-quality, coherent text.

These models, trained on huge amounts of textual data, have revolutionised fields such as natural language processing and content generation (Jordan & Mitchell, 2015). During training, LLMs learn to predict the next word in a sentence based on the context provided by previous words. This is achieved by assigning probabilities to the recurrence of words that have been tokenised (split into smaller sequences of characters). LLMs demonstrate a deep understanding of human language, enabling them to perform tasks such as translating, summarising text, or answering complex questions (IBM, n.d.).

In the early days of AI, Alan Turing posed whether machines were capable of thinking autonomously (Turning, 1950). For him, thinking was conceptualised as an act of reasoning and providing appropriate answers to questions related to various topics. From this conceptualisation, it was affirmed that machines currently have the capacity to think, as they are capable of providing increasingly adequate, detailed and precise answers to all kinds of questions.

Despite the current popularisation of the term, the origins of AI can be traced back to the mid-20th century. Although there is no consensus on its precise beginning, some authors consider the genesis of the current revolution to be in 1943, the year in which the first mathematical model of a neural network was developed. This laid the foundations for the development of artificial neural networks (McCulloch and Pitts, 1943). In 1950, Alan Turing published his article “Computing Machinery and Intelligence”, where he introduced the famous Turing Test to evaluate the ability of a machine to exhibit intelligent behaviour equivalent to that of a human (Turing, 1950).

Subsequently, Arthur Samuel developed a program with the ability to autonomously learn to play chess, being one of the first examples of automatic learning (Samuel, 1952). However, it was John McCarthy who in 1956 used the term “Artificial Intelligence” for the first time at the Dartmouth Conference (Moor, 2006), and this milestone is considered one of the initial starting points of AI. Since that time, the technology has continued to improve unstoppably with countless advances, such as the development of symbol manipulation languages (such as LISP, POP and IPL), as well as advances in hardware such as processors and memory (Buchanan, 2005). During the following decades, expert systems, computer programs capable of solving complex problems by simulating human knowledge work in a specific domain, were created (Russell and Norving, 2016). However, during the same period, great expectations were generated that were not fulfilled, and the 1970s and 1980s were considered the “winter” of AI, with investments in its development being cut back (Lighthill, 1973).

In the 1990s, supervised and unsupervised learning systems were developed through the use of machine learning algorithms that allowed machines to learn from data without the need to be explicitly programmed, as well as deep neural networks. In addition, improvements in computational processing power continued (Lundstrom, 2003), allowing the complexity of the tasks that systems are able to undertake, such as image recognition or natural language processing, to continue to increase (LeCun et al., 2015). The development of the Internet has also played a key role in the rise and development of AI. It has provided an immense and growing amount of data from a variety of sources, such as social networks, multimedia content, academic publications, and many other types of digital sources on the web.

Such data, which are constantly generated and updated on a global scale, have been, and continue to be, used to train and refine increasingly complex and sophisticated algorithms (Jordan and Mitchell, 2015). The use of information from the Internet to train AI systems poses significant intellectual property challenges, as these models use information created by third parties, without their prior authorisation, to create their content, leaving an unresolved legal loophole (Hartmann et al., 2020). In addition, there is the legal attribution of the very content generated by the AI itself (Cubert and Bone, 2018).

All of these advances allowed for the development of more sophisticated and capable systems. Deep Blue's victory over Garry Kasparov in 1997 represented a turning point in the development of AI. It demonstrated that machines could outperform humans in complex tasks such as chess. This event marked the beginning of a new era in which AI began to be perceived as a tool capable of tackling problems of any kind (Campbell et al., 2002).

The first decade of the 21st century saw the development and refinement of deep learning systems, which can recognise complex patterns in data such as images, text, audio and video, representing another important advance in the development of AI (Krizhevsk et al., 2012).

Table 1. Time required for technology platforms and services to reach 100 million users

| Technology platform / service | Number of months needed to reach 100 million users |
|-------------------------------|--|
| Threads | 0.07 |
| ChatGPT | 2 |
| TikTok | 9 |
| YouTube | 18 |
| Instagram | 30 |
| Facebook | 54 |
| X/Twitter | 60 |
| Spotify | 132 |
| Netflix | 216 |

Source: Statista (2024)

However, it was not until 2018, with the creation of the *Generative Pre-trained Transformer* (GPT) model, that AI through LLM systems became popular and were broadly used (Radford et al., 2018). According to ChatGPT itself, its name combines “Chat”, which refers to its ability to interact in real time with users by answering questions and assisting with various tasks, and “GPT”, an artificial intelligence architecture developed by OpenAI. Using a pre-trained model with large amounts of data, it generates text in a coherent and natural way, based on the efficient Transformer architecture (OpenAI, 2024). There are a growing number

of alternative LLM models, such as Anthropic's Claude, Google's Gemini, and Meta's Llama (open model to download and use locally), among others. All of them are developments in generative artificial intelligence, which allow content to be generated in various formats (text, image, audio, or video) from the content with which the system has been trained.

As can be seen in Table 1, ChatGPT has been one of the technological applications that has taken the least time – two months – to reach 100 million users (Statista, 2024). It is currently only surpassed by Instagram's Threads, as a microblogging network that aims to position itself as an alternative to X (formerly Twitter) (Bonet, 2023).

These data show that AI has become a sophisticated, as well as every day, tool. It is only a matter of time before it is used by practically the entire world population in all kinds of daily tasks, including at work. With the near world-wide use of AI coming upon us quickly and the fact that it has been trained on data created by people, it would be natural to conclude that AI has inherited and incorporated human biases. This leads to whether this technology presents problems with respect to gender. The following sections analyse the gender issues currently presented by AI.

2. Gender perspectives in AI

2.1. Gender in AI development

Women's participation in certain scientific fields is significantly lower than that of men. Despite advances in women's presence in STEM (*Science, Technology, Engineering, and Mathematics*) fields, there is still a significant gap. According to UNESCO (2021), globally, only 28% of science researchers are women. In the United States, women represent approximately 27% of the STEM workforce, although they constitute 47% of the total workforce (National Science Foundation, 2021). In the European Union, 41% of scientists and engineers are women. However, the situation changes when we look at women as self-employed professionals in science, engineering, and information and communication technologies, representing only 7% (European Commission, 2021).

The development of AI falls within the STEM area, so the participation of women in the development of this technology, which will radically transform the society we live in, represents a potential problem. The lower participation of women in many work activities has had consequences on the development and design of many products and services. There are numerous examples of this issue. For example, smartphones, which have mostly been designed by men, are sized to be more suitable for male hands, almost half a centimetre larger than female hands (Garber, 2014).

Likewise, car safety systems were historically designed with dummies used in crash tests that had been modelled after the average male body, resulting in automotive safety systems that do not equally protect women (Criado, 2019). Because of this design and development bias, largely caused by developing products with mostly male teams, women are 73% more likely to be seriously injured in car crashes than men (Samarrai, 2019). As a final example, only 19% of practising industrial product designers are women, and only 11% of leadership roles in industrial design are held by women. This shortage of female designers results in many of the new products launched on the market perpetuating stereotypes about female consumers, failing to meet the real needs of future female users (Samarri, 2023).

Like all technologies before it, AI will reflect the values of its creators (Crawford, 2016). There is a diverse crisis in the AI sector in terms of gender. For example, the participation of women is substantially lower than that of men. A recent study by Perrault and Clark (2024) points to more male than female graduates in AI-related programmes, such as computer science, computer engineering and information technology studies, taking into account undergraduate, master's and doctoral levels. In the case of Spain, in the year 2022, men represented 85.68% of students in the aforementioned areas, compared to 14.32% of women (Perrault and Clark, 2024). In industry, we find a similar situation. According to Simonite (2018), for example, at Google only 21% of technical roles are occupied by women. In the area of "machine intelligence", there were 641 people working, of which only 10% were women. Facebook reported that 22% of its technical workers are women. The company's AI research group pages listed 115 people (Simonite, 2018).

Therefore, and taking into account that AI is categorised within the STEM area, there is a clear deficit of women's participation in the development of this technology. Although there is a positive trend to correct this diversity deficit in the AI sector, male employment in the sector is still preponderant. According to Zeki (2024), women's labour market participation has grown by an annual average of 16% in the period between 1990 and 2023. Despite the positive data, there is still a need for measures to increase women's participation in the AI market.

Considering the enormous transformative impact of AI, and the problems that can arise from not incorporating women in the design of this technology, the potential concerns must be addressed as one of the priority challenges for the ethical and responsible development of AI.

2.2. Gender biases in data used by AI

As noted above, AI requires an enormous amount of data for training, and the quality and nature of those data will determine the quality of the output. If the data used presents gender-stereotyped information, the technological applications that emerge from this learning will not only reproduce these biases, but also contribute to perpetuating them (Leavy, 2018). This raises important ethical and social questions, as more and more AI systems that provide scoring by potentially biased algorithms are increasingly used to make decisions that affect people's lives in areas such as finance, employment, insurance, and many others (Keats and Pasquale, 2014).

For example, a system used by judges to set probation found that the assessment of the likelihood of recidivism was biased against black defendants (Ali et al., 2010). Or the fact that facial recognition systems on mobile phones tend to work better for white men than for other races and/or genders (Buolamwini and Gebru, 2018).

AI systems, despite being developed at a time when gender aspects are already imperative in the development of any activity, may present a bias that not only does not correct, but reinforces certain structural inequalities. This negatively impacts various groups and consolidates discriminatory narratives in automated decisions.

2.3. The biases of the algorithms used by AI

In addition to biases in the data, there are biases that can be induced by the algorithm itself. Algorithmic bias refers to systematically erroneous or unfair results produced, which may reflect and amplify human biases present in the design of the algorithm (Buolamwini and Gebru, 2018). As an example, a recent study revealed that ChatGPT shows a notable and systematic bias in favour of the Democratic Party in the United States and Labour in the United Kingdom (Motoki et al., 2024).

These biases can arise from the selection of certain characteristics or variables that inadvertently influence the decision-making process. By choosing which data to include or exclude, and how they are weighted in the analysis, it is possible to introduce biases that negatively affect the results, even when they were not intended to do so (MELT Group, 2023). For example, an algorithm trained on thousands of images available on the Internet learned to associate women with kitchens, due to the higher representation of women in that environment in online photographs. This fact would imply an initial data bias for training, which later, during the learning process, the algorithm not only replicated, but amplified.

This led to a reinforcement of the association between women and kitchens (Zhao et al., 2017). Another example of this type of gender bias was in a 2016 US presidential prediction. The algorithm was trained with images of former US presidents, which predicted Donald Trump's victory in the 2016 election. The algorithm, based on its information, knew that there had never been a female president in the country, so its prediction favoured the male candidate over the female (Polonski, 2016).

These biases are present in a multitude of AI systems that make more and more decisions every day. A recent study by Lambrecht and Tucker (2019) found that even though a STEM job ad is designed to be gender-neutral, the advertising platform's optimisation algorithm showed the ad to more men than women. This occurred because women are considered a more valuable demographic for advertisers, resulting in a higher cost to show them other types of ads. As a result, the algorithm, in trying to optimise cost and reach, ended up showing the ad disproportionately to men. In the area of targeting, there have also been cases of discrimination. Recent research, in which eighteen companies using algorithmic recruitment services were analysed, found that several recruitment algorithms showed preferences for male candidates, especially in traditionally male-dominated fields (Raghavan, Barocas, Kleinberg & Levy, 2019).

Common practices, such as removing explicit demographic information, may be insufficient to prevent algorithmic discrimination (Raghavan et al., 2020). Credit scoring algorithms, by employing proxy variables¹, can cause indirect discrimination against women, who often have lower credit scores due to historical and socio-economic factors, such as the wage gap and disrupted employment patterns. AI algorithms can amplify pre-existing biases by learning from historical data that reflect past discriminatory practices. As a result, it has been observed that women, on average, receive lower loan sizes and higher interest rates compared to men with similar risk profiles (Aggarwal et al., 2021).

Therefore, the assessment of the development of ethical and responsible AI is becoming increasingly relevant. Among the criteria for assessing the degree of responsible AI development is fairness, understood as the need to create algorithms that are fair and equitable, avoiding bias or discrimination, and considering the diverse needs and circumstances of all stakeholders. This aligns with broader societal standards of equity. Designing fair algorithms implies that they have a proven ability to avoid bias in recommendations, ensuring that users from all demographic groups, including women, receive equitable and non-discriminatory care (Perrauk and Clark, 2024).

3. Gender challenges and AI

3.1. Enhancing gender diversity in the development of AI

The AI sector is in need of a profound change in its approach to the current diversity crisis. It is recommended that both industry and society become aware of the potential problems that a lack of diversity can cause for the ethical and responsible development of a technology. This technology will be the backbone of society in the near future. Furthermore, the impact of diversity in AI teams will enable the development of less biased and more inclusive systems. The incorporation of women in these teams not only contributes to equity, but also to an increase in innovation and the consideration of diverse perspectives during algorithm design (UN Women, 2024).

¹ Proxy variables are characteristics or data used in place of a variable of interest that is unavailable or difficult to measure directly. In the context of statistical models and algorithms, proxy variables serve as surrogates to represent a desired measure that cannot be captured accurately. For example, if you want to assess a person's creditworthiness, but do not have access to their full income history, you could use a proxy variable such as education level or employment type, which may be correlated with income. Although these variables are not direct measures of income, they can provide relevant indirect information. In the case of credit scoring algorithms, proxy variables may include data such as employment type, utility payment history, or even postcode. However, the use of proxy variables can lead to problems if these characteristics are correlated with factors that introduce bias, such as gender or race, which could be considered indirect discrimination.

The promotion of women's studies in STEM fields, although it has shown positive impacts in recent years, needs to be further enhanced. Early encouragement programmes significantly increase women's interest and confidence in these fields, as well as visibility campaigns of successful women in these disciplines, as they serve as role models, inspiring more girls to opt for careers in these areas (Cheyan et al., 2011).

Similarly, with regard to the incorporation of women in the industry, inclusive policies have improved the percentage of women working in the technology industry. Companies that implement diversity and inclusion policies show an increase in the recruitment and retention of women in STEM positions. The development of mentoring programmes and professional support networks also favours the incorporation and retention of women in technology companies. Fostering changes in organisational culture, focused on promoting gender equality, will lead to more women in leadership roles in these sectors (Wang and Degol, 2017).

To reduce the gender gap in STEM, it is essential to address the cognitive, motivational and socio-cultural factors that influence it by maximising the career options that women perceive as attainable and compatible with their skills and goals. This is achieved through collaboration between researchers, practitioners, and policymakers. They must work together to eliminate gender stereotypes, cultural barriers, and misinformation, and thus increase female interest and participation in these areas (Cheyan et al., 2011).

3.2. Enhancing technological systems for the correction of data and algorithmic biases

To mitigate this problem, technical measures, such as developing and employing unbiased data frameworks, enhancing algorithmic transparency, as well as management measures, such as internal corporate ethical governance and external oversight, are recommended (Chen, 2023).

Biased datasets result in logically biased results. Therefore, the way to address algorithmic data bias is to reconfigure datasets that are gender-biased. The development of techniques that allow for identification and correction is one solution, although it presents the ethical dilemma of the change that bias modification represents, modifying the actual data to avoid bias. Moreover, rigorous elimination of biased data, given the enormous amount of existing data, involves very high costs (Bornstein, 2018). Another challenge is not to rely solely on large data collections, but also to enhance accuracy by combining *big data* with *small data*, the latter being comprised of more accurate, smaller and more manageable sources. Combining the enormity of 'big data' with the accuracy of 'small data' can help mitigate, to some extent, algorithmic errors and biases (Kitchin and Lauriault, 2015).

The other way to avoid bias is in the algorithm itself. To address gender bias in algorithms, several strategies can be applied at different stages. In pre-processing, re-sampling techniques can be used to balance the gender representation, and sensitive features that are not relevant to the task can be removed. During algorithm processing, fairness constraints can be incorporated into the objective function and regularisation can be employed to penalise biased decisions. In post-processing, decision thresholds can be adjusted to ensure similar error rates across gender groups, and calibration techniques can be applied to balance predictions. A holistic approach considers the social and ethical context at all stages of development, including the involvement of diverse teams in design and evaluation (Mehrabi et al., 2021).

3.3. Enhancing the explainability of AI systems

AI acts as a "black box" model, in which we ask it to perform a process, and it returns a result, without having the slightest idea of how it arrived at that result. If we consider that we increasingly trust the results proposed by algorithms, we are facing a major problem. On many occasions, they present us with erroneous or biased results, which are difficult to detect since we lack the necessary information to know how they arrived at the results.

Therefore, enhancing the transparency and explainability of results will be critical to both improving AI systems and correcting potential gender biases. Algorithmic explainability, also known as algorithmic interpretability or explainable/explicable AI (XAI), refers to the degree to which decisions or predictions made by an AI or machine learning system can be understood and interpreted by humans (Adadi and Berrada, 2018). It involves methods and techniques that make the operation of complex algorithms transparent and understandable, allowing users, developers, and stakeholders to understand how and why a particular result was generated (Arrieta et al., 2020). According to Lipton (2018), explainability allows users to understand the behaviour of a model, identify possible errors and build trust in the system. Algorithmic explainability becomes particularly relevant in high-risk domains such as healthcare, finance and criminal justice, where algorithmic decisions can have a significant impact on people's lives (Rudin, 2019).

In the context of machine learning, Lipton (2018) argues that explainability encompasses several aspects, including *transparency*, i.e. the ability to understand the structure and learning process of the model; *interpretability*, which involves the ability to provide human-understandable explanations for individual predictions; and *post-hoc interpretability*, which refers to the methods used to explain the behaviour of a model once it has been trained.

Therefore, the lack of transparency in algorithmic models makes it difficult to effectively assess and mitigate bias (Raghavan et al., 2020), making this a priority area both for the improvement of AI systems in general and the avoidance of gender bias in particular.

3.4. Enhancing responsible development and accountability in the development of AI

Responsibility and accountability in the field of AI have become increasingly relevant issues, as AI systems continue to play ever significant roles in decision-making processes in various domains. They will therefore be key factors in ensuring the ethical development and implementation of AI technologies.

Accountability in AI refers to the obligation of AI developers, implementers, and users to ensure that AI systems are designed and used in ways that are beneficial and not harmful to individuals or society (Dignum, 2019). *Accountability* implies the ability to determine and address the consequences of the actions and decisions of AI systems (Wieringa, 2020), including gender biases.

However, implementing responsible and accountable AI is not easy, as many actors are involved in development and implementation, making it difficult to attribute responsibility to a single entity (Floridi, 2016). Moreover, the inherently opaque nature of algorithms complicates the understanding and explanation of their decision-making processes, further hindering efforts to ensure accountability (Mittelstadt et al., 2016). To this end, different frameworks have been developed, such as the ART (*Accountability, Responsibility, and Transparency*) proposed by Dignum (2019), which emphasises the need to enhance these three key elements in the design and implementation of AI systems. The IEEE has also recommended *Ethically Aligned Design* (IEEE, 2019), which offers a comprehensive approach to integrating ethical considerations into the design of AI systems.

Regarding legal and regulatory aspects, governments and international organisations are increasingly developing regulations to ensure responsibility and accountability in AI. For example, the proposed European Union AI Law seeks to establish a legal framework for the development and use of AI systems, including provisions for accountability and transparency (European Commission, 2021).

In the area of technical solutions, researchers are developing methods to improve accountability in AI. These include XAI techniques, which aim to make AI decision-making processes more interpretable (Adadi and Berrada, 2018), as well as algorithmic auditing tools to detect and mitigate biases in AI systems (Raji et al., 2020).

It will be essential to encourage the ethical development of this technology. More and more companies in the AI industry are publishing AI principles that address potential problems with their algorithms, initiating self-regulatory processes (Chen, 2023). Microsoft has formed an AI ethical standards committee to enforce these principles, subjecting all future AI products to ethical scrutiny (Smith and Shum, 2018). Google has introduced the concept of the '*Model Card*', a sort of algorithm manual, which explains the algorithm used, highlights strengths and weaknesses, and even shares operational results from various datasets (Mitchell et al., 2019).

4. Conclusions

AI represents a new technological revolution with, if possible, greater economic and social impacts than previous ones. Therefore, promoting responsible and ethical development is a priority. Within the challenges to achieve an adequate development and use of AI, gender aspects must be included. The main problems are to be found in the unbalanced participation of women in the technology sector, resulting in a lower proportion of women studying in STEM areas. In addition to the gender inequality of the sector's workforce, this fact leads to potential development failures due to the lack of a female perspective in development. It is also advisable to explore mechanisms that develop and foster a culture of diversity inclusion in organisations, beyond merely promoting numerical balance or quotas (Collett and Dillon, 2023).

In addition, many of the gender biases are inherited from previous data, as these systems are trained with previous information in which there was biased content. Finally, the algorithms themselves, in their design, may incorporate biases that do not take into account some gender-relevant dimensions.

To solve these problems, women need to be encouraged to enter both STEM education and industry. Technology must develop mechanisms to avoid bias in the datasets used, and care must be taken in algorithmic development to avoid gender biases as well. Enhancing the explainability of AI systems will enable ethical and responsible AI development, making it easier to detect and correct gender biases.

5. Limitations and future lines of research

The methodology used has many limitations, as it is based on a content analysis to explore the areas in which AI presents or may present gender problems in the future. In the theoretical review, the most relevant topics have been selected, leaving other aspects to be studied, such as the gender bias in interfaces, voice assistants, and the concerns derived from language biases.

Voice assistants, such as Siri, Alexa or Google Assistant, have been criticised for reproducing gender stereotypes in their design and functionality. One of the main aspects pointed out is that these systems often use default female voices, which reinforces the stereotype of women as assistants or subordinate figures (West et al., 2019). However, some studies justify this, as people prefer female bots over male ones because they perceive the former to possess more warmth and humanity, qualities that are fundamental to the perception of humanity but absent in machines (Borau et al., 2021).

With respect to problems stemming from language biases, they directly affect the performance of AI systems, particularly in areas such as speech recognition, machine translation and text generation. For example, AI algorithms may associate words such as "doctor" with men and "nurse" with women, perpetuating gender stereotypes (Caliskan et al., 2017). Such associations not only reinforce biases, but also have serious consequences when AI systems are applied in sensitive environments such as criminal justice or recruitment, where automated decisions can have a negative impact on vulnerable populations.

Regarding future lines of research, in addition to the areas included in this paper, the intersectional analysis of AI biases should be strengthened. Beyond studying only gender differences, it is crucial to understand how other factors, such as race, class, or disability, interact with gender in AI systems. Biases in algorithms not only affect women in general, but can have differential effects according to race, which requires an intersectional analysis to adequately address inequities in AI (Noble, 2018).

Another area for future research is the study of the role of regulation and public policy in ensuring that AI is inclusive and equitable, exploring how regulatory frameworks can influence the design, implementation, and oversight of gendered AI technologies (Whittaker et al., 2018). Similarly, from a methodological perspective, empirical development and testing is needed, with both qualitative and quantitative studies. Also, interdisciplinary studies, bringing together technical and social aspects, are needed to have a proper understanding of gender aspects in the future development of AI.

AI will continue to evolve, bringing about disruptive changes, and there is a growing need for interdisciplinary approaches to address responsibility and accountability in AI. This includes collaboration between technology scientists and social scientists in areas such as ethics and public policymaking to identify and develop the regulatory and social frameworks needed to correct existing deficiencies (Coeckelbergh, 2020).

Funding

This research was supported by the JSPS Grant-in-Aid for Scientific Research (C) 23K01545; and the Telefonica Chair on Smart Cities of the Universitat Rovira i Virgili and Universitat de Barcelona (project number 42.DB.00.18.00).

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Aggarwal, N., Gupta, A., & Rani, S. (2021). Fairness in AI-based credit scoring: A systematic review. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 712–724). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445931>
- Ali, O., Flaounas, I., De Bie, T., Mosdell, N., Lewis, J., & Cristianini, N. (2010). Automating news content analysis: An application to gender bias and readability. In *Proceedings of the First Workshop on Applications of Pattern Analysis* (pp. 36–43).
- 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. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bonet, R. (2023, July 6). Threads en Instagram: Qué es, cómo funciona y qué promete esta red social. *Xataka*. <https://www.xataka.com/basics/threads-instagram-que-como-funciona-que-promete-esta-red-social>
- Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing*, 38(7), 1052–1068. <https://doi.org/10.1002/mar.21490>
- Bornstein, S. (2018). Antidiscriminatory algorithms. *Alabama Law Review*, 70(2), 519–572.
- Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AI Magazine*, 26(4), 53–60. <https://doi.org/10.1609/aimag.v26i4.1848>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91).
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10.1126/science.aal4230>
- Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep blue. *Artificial Intelligence*, 134(1–2), 57–83. [https://doi.org/10.1016/S0004-3702\(01\)00129-1](https://doi.org/10.1016/S0004-3702(01)00129-1)
- Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities & Social Sciences Communications*, 10(1), 1–12. <https://doi.org/10.1057/s41599-023-02079-x>
- Cheryan, S., Siy, J. O., Vichayapai, M., Drury, B. J., & Kim, S. (2011). Do female and male role models who embody STEM stereotypes hinder women's anticipated success in STEM? *Social Psychological and Personality Science*, 2, 656–664. <https://doi.org/10.1177/1948550611405218>
- Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.
- Collett, C., & Dillon, S. (2023). AI and gender: Four proposals for future research. University of Cambridge. <https://doi.org/10.17863/CAM.41459>
- Comisión Europea. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence. <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence>
- Crawford, K. (2016, June 25). Artificial intelligence's white guy problem. *The New York Times*. <https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html>
- Criado Perez, C. (2019). *Invisible Women: Data Bias in a World Designed for Men*. Abrams Press.
- Cubert, J. A., & Bone, R. G. (2018). The law of intellectual property created by artificial intelligence. In *Research Handbook on the Law of Artificial Intelligence* (pp. 411–427). Edward Elgar Publishing.
- Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer.
- Ekmekci, P. E., & Arda, B. (2020). History of artificial intelligence. In *Artificial Intelligence and Bioethics* (pp. 1–27). Springer.

- European Commission. (2021). We still need more women in science. https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_24_732
- Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160112. <https://doi.org/10.1098/rsta.2016.0112>
- Garber, M. (2014, September 19). Are the new iPhones too big for women's hands? *The Atlantic*. <https://www.theatlantic.com/technology/archive/2014/09/are-the-new-iphones-too-big-for-womens-hands/379911/>
- Goertzel, B., & Pennachin, C. (2007). *Artificial General Intelligence* (Vol. 2). Springer.
- Goodfellow, I., Bengio, Y., & Courville, A. (2020). *Deep Learning*. MIT Press.
- Gutiérrez, M. (2021, February 8). Los sesgos de género en los algoritmos: Un círculo perverso de discriminación en línea y en la vida real. *eldiario.es*. https://www.eldiario.es/tecnologia/sesgos-genero-algoritmos-circulo-perverso-discriminacion-linea-vida-real_129_7198975.html
- Hartmann, C., Allan, J. E., Hugenholtz, P. B., Quintais, J. P., & Gervais, D. (2020). Trends and developments in artificial intelligence: Challenges to the intellectual property rights framework. <https://data.europa.eu/doi/10.2759/683128>
- IBM. (n.d.). ¿Qué son los grandes modelos de lenguaje (LLM)? <https://www.ibm.com/es-es/topics/large-language-models>
- IEEE. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems. <https://standards.ieee.org/industry-connections/ec/autonomous-systems.html>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6249), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Keats, D., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89(1), 1–33.
- Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, 80, 463–475. <https://doi.org/10.1007/s10708-014-9601-7>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
- Kurzweil, R. (1990). *The Age of Intelligent Machines*. MIT Press.
- Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7), 2966–2981. <https://doi.org/10.1287/mnsc.2018.3093>
- Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering* (pp. 14–16).
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lighthill, J. (1973). Artificial intelligence: A general survey. In *Artificial Intelligence: A Paper Symposium*. Science Research Council.
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Lucci, S., & Kopec, D. (2016). *Artificial Intelligence in the 21st Century: A Living Introduction*. Mercury Learning and Information.
- Lundstrom, M. (2003). Moore's law forever? *Science*, 299(5604), 210–211. <https://doi.org/10.1126/science.1079567>
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, 27(4), 12–14. <https://doi.org/10.1609/aimag.v27i4.1904>
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133. <https://doi.org/10.1007/BF02478259>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
- MELT Group. (2023, November 24). Sesgos de la IA: ¿Cuáles son y qué consecuencias tienen? *MELT Group*. <https://meltgroup.com/sesgos-de-la-ia/>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., & Gebru, T. (2019). Model cards for model reporting. In V. Conitzer, G. Hadfield, & S. Vallor (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency*. The Association for Computing Machinery.
- Moor, J. (2006). The Dartmouth College artificial intelligence conference: The next fifty years. *AI Magazine*, 27(4), 87–91. <https://doi.org/10.1609/aimag.v27i4.1911>
- Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice*, 198, 3–23. <https://doi.org/10.1007/s11127-023-01097-2>
- Nass, C., & Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press.
- National Science Foundation. (2021). *Women, minorities, and persons with disabilities in science and engineering: 2021*. National Center for Science and Engineering Statistics. <https://nces.nsf.gov/pubs/nsf21321>

- Nilsson, N. J. (2009). *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge University Press.
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- OpenAI. (2024). ChatGPT (versión del 15 de mayo) [Modelo de lenguaje de gran tamaño]. <https://chat.openai.com/chat>
- Perrault, R., & Clarck, J. (2024). *Artificial Intelligence Index Report 2024*. Human-Centered Artificial Intelligence, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf
- Polonski, V. (2016, November 6). Would you let an algorithm choose the next U.S. president? *TechCrunch*. <https://techcrunch.com/2016/11/06/would-you-let-an-algorithm-choose-the-next-u-s-president/>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20)** (pp. 469–481). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372828>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). <https://doi.org/10.1145/3351095.3372873>
- Raphael, B. (1976). *The Thinking Computer: Mind Inside Matter*. W.H. Freeman.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Russell, S. J., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson Education.
- Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- Samarrai, F. (2019, July 10). Study: New cars are safer, but women are still more likely to suffer injury. *University of Virginia*. <https://news.virginia.edu/content/study-new-cars-are-safer-women-most-likely-suffer-injury>
- Samarri, K. (2023, January 18). Shrink it and pink it: Gender bias in product design. *Harvard Advanced Leadership Initiative*. <https://www.sir.advancedleadership.harvard.edu/articles/shrink-it-and-pink-it-gender-bias-product-design>
- Samuel, A. L. (1952). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>
- Simonite, T. (2018). AI is the future – but where are the women? *WIRED*. <https://www.wired.com/story/artificial-intelligenceresearchers-gender-imbalance/>
- Smith, B., & Shum, H. (2018). *The Future Computed: Artificial Intelligence and Its Role in Society*. Microsoft.
- Statista. (2024, September 1). Time taken for selected online platforms and services to reach 100 million followers as of August 2024 (in months) [Graph]. *Statista*. <https://www.statista.com/statistics/1489983/selected-platforms-services-reach-one-hundred-million-followers/>
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- UN Women. (2024). Placing gender equality at the heart of the Global Digital Compact. UN Women. <https://www.unwomen.org/sites/default/files/2024-03/placing-gender-equality-at-the-heart-of-the-global-digital-compact-en.pdf>
- UNESCO. (2021). *UNESCO Science Report: The Race against Time for Smarter Development*. UNESCO Publishing. <https://unesdoc.unesco.org/ark:/48223/pf0000377250>
- Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, 29(1), 119–140. <https://doi.org/10.1007/s10648-015-9355-x>
- Wang, P., & Goertzel, B. (2012). *Theoretical Foundations of Artificial General Intelligence*. Atlantis Press.
- West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. *AI Now Institute*. <https://ainowinstitute.org/discriminatingystems.html>
- Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., ... & Schwartz, O. (2018). *AI Now Report 2018*. AI Now Institute.
- Zeki. (2024). Women in AI 2024. *Horizon Report*. <https://zekidata.com/report/women-in-ai-2024/>
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.