ISSN 2007-9737

Explainable and Evaluative Artificial Intelligence: Alternatives to Ensure Equity in Decision-Making

Yenny Villuendas-Rey[1], Oscar Camacho Nieto[1,*], Claudia C. Tusell-Rey[2], Viridiana Salinas-García[1], Joel Pino Gómez[3]

[1] Instituto Politécnico Nacional, CIDETEC-IPN, México

[2] Instituto Politécnico Nacional, CIC-IPN, Mexico

[3] Mondragon Unibertsitatea, Faculty of Engineering, Spain

{yvilluendasr, ocamacho, vsalinasg}@ipn.mx, {joelpinogomez, clautusellrey2014}@gmail.com

Abstract. Explainable Artificial Intelligence (XAI) and Evaluative Artificial Intelligence (EAI) are crucial approaches to ensuring fairness in decision-making. XAI refers to the ability of Artificial Intelligence (AI) systems to be understood and explained by humans, while EAI is a new paradigm that focuses on identifying possible decisions for intelligent algorithms by formulating hypotheses for and against them. These paradigms are crucial in a world where AI increasingly influencing key areas, such as employment, healthcare, and justice, among others. XAI allows developers, regulators, and users to understand how AI systems make decisions. This is essential for identifying biases and preventing unfair outcomes. Transparency in how algorithms work can help correct problems before they cause significant harm. This is crucial for the widespread adoption of AI across different sectors. Furthermore, it is crucial to assess the fairness and impartiality of AI algorithms. Since algorithms can reflect and amplify biases existing in the data they are trained on, it is essential to assess and mitigate these biases to ensure that AI decisions do not perpetuate discrimination or injustice. This chapter focuses on exploring existing paradigms and techniques to ensure that AI is used fairly and equitably in different contexts. Promoting transparency, understanding, and evaluation of AI systems will help build a future where technology benefits everyone fairly and equitably.

Keywords. Explainable artificial intelligence, evaluative artificial intelligence, decision-making, equity and justice.

1 Introduction

Artificial Intelligence (AI) has proven to be extremely useful in various aspects of modern life. For example, Insilico Medicine uses AI to identify potential compounds in a fraction of the time it would take using traditional methods, facilitating drug design, among many other benefits. Also in healthcare, there are AI systems, such as IBM Watson Health, that are used to analyze large volumes of clinical data and medical literature, providing support in complex medical decision making.

However, AI, like any advanced technology, has the potential to be used for both good and bad purposes. For example, AI algorithms can be trained to identify vulnerabilities in computer systems or to launch more sophisticated phishing attacks [1;2]. AI can be used to automate large-scale hacking attacks, carry out financial fraud more efficiently, or coordinate criminal activities such as drug and human trafficking [3-5]. Text and image generation algorithms can also be used to create convincing fake news or even to manipulate social media. Furthermore, image and video generation algorithms, powered by artificial intelligence, have the ability to create highly realistic visual content that can be used to impersonate people in various ways. For example, Generative Adversarial Networks (GANs) can generate photographs of people who

do not exist in reality, but which are indistinguishable from real images. Also, AI algorithms, such as Deepfakes, can create fake videos that superimpose one person's face onto another's body in real time, making it appear that they are performing actions or saying things that never happened [6-8].

Beyond purely malicious use, AI is not without risks inherent to its very nature. Thus, even its well-intentioned use can have negative consequences. Artificial intelligence algorithms are designed by humans and can therefore reflect the conscious or unconscious biases of their creators. For example, if developers do not adequately consider the diversity of perspectives and cultural contexts in the design of the algorithm, they could introduce biases that affect the automatic decisions made by the system. A highly notable example internationally was the discovery that facial recognition systems performed significantly worse with people from certain ethnic groups, which made headlines [9] and has continued to be a source of scientific interest [10;11]. It should also be noted that AI models learn from historical data, and if this data contains inherent biases (such as racial, economic, or gender discrimination), the model can learn and perpetuate those biases in its predictions and recommendations. For example, AI-based recruitment systems have shown age and gender biases toward certain ethnic groups, attracting international attention from both the media [12, 13] and the scientific community [14, 15].

Furthermore, but even more seriously, predictive policing has shown clear biases [16] in considering certain minorities as “more criminal,” generating corresponding social concern [17].

As can be seen, even with well-intentioned use, biases in AI systems are a significant problem that requires diligent attention and action. Biases in AI algorithms can perpetuate and amplify existing inequalities in society, making it more difficult for certain groups to access fair and equitable opportunities. Even if AI systems are used to make critical decisions in areas such as healthcare, employment, credit allocation, and criminal justice, among others, biases can lead to unfair and discriminatory decisions [18].

To avoid these unwanted AI behaviors, and to ensure that the technology is used ethically and fairly, without perpetuating bias and discrimination, transparency must be promoted regarding how data is collected, selected, and used in models.

Explainable artificial intelligence (XAI) and, more recently, evaluative artificial intelligence (EAI) play a fundamental role in reducing inequity, discrimination, and bias in AI systems by providing transparency and understanding of how models work and why they make certain decisions. This paper addresses some of the advances in explainable artificial intelligence.

2 Advances in Explainable Artificial Intelligence and Evaluative Artificial Intelligence

For a decade now, there has been growing interest from the scientific community and society at large regarding the importance of ensuring fairness, transparency, and accountability in Artificial Intelligence systems. Initially, in 2014, the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) was created as an international forum for the study of this topic, and since 2018, the Association for Computing Machinery (ACM) has followed up on it through the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), as well as other similar forums.

These forums seek to bring together a diverse community of academics from different disciplines such as computer science, law, and social sciences to research and address issues related to fairness, transparency, and accountability in intelligent models. Specifically, it aims to "evaluate technical solutions to existing problems, reflecting on their benefits and risks; address fundamental questions about economic incentive structures, perverse implications, power distribution, and welfare redistribution; and ground research on equity, accountability, and transparency in existing legal requirements" [19].

Among the advances achieved in these areas, the awareness of intelligent algorithm development and application environments [20] stands out, particularly the use of performance

metrics, as well as the reflection on how to deal with changing data, how to better understand model outputs, and how to improve the transmission of social practices and the communication of aspects related to algorithmic performance.

Zhang et al. created a development framework to consider equity in the design of AI products. Their framework includes fairness aspects related to statistical parity, demographic parity, equal opportunity, equal probabilities, fairness of evidence, equal treatment, counterfactual fairness [21], fairness in relational domains, fairness through awareness, and fairness through ignorance [22]. This development framework has an online tool, as well as reflection cards, to assist AI development teams and stakeholders in the AI solution design process. Plečko and Bareinboim also developed a toolkit for studying fairness in AI systems, using a causal approach [23]. Other researchers have focused on the theoretical study of the trade-off between fairness and accuracy in AI models [24], as well as on the search for software defects related to the fairness of AI systems [25]. Likewise, strategies have been developed to verify the absence of bias even when there are regulations that prohibit, in some countries, the use of sensitive features that can lead to algorithmic biases. It is known that, even when sensitive features are omitted, they could be related in some way to other features, called proxy features. The proposal of [26] shows how to reveal whether a black-box model, which complies with the regulations, is still biased or not. To do so, an end-to-end bias detection approach is presented that exploits a counterfactual reasoning module and an external classifier for sensitive features. In detail, the counterfactual analysis finds the minimum cost variations that give a positive result, while the classifier detects nonlinear patterns of proxy features that represent sensitive characteristics. In this way, it is possible to detect classifiers that learn from proxy features. Other approaches include manual annotation of protected features [27] and analysis of user perception when using protected and/or proxy features [28].

On the other hand, several studies have been conducted to understand users' perceptions of fairness and bias in AI systems [29-31]. These studies can help us understand how to involve humans in the development and evaluation process of algorithmic decision-making systems, how to create personalized explanations based on the system's characteristics, as well as the users' personality and demographic characteristics, and, consequently, how to improve users' perceptions of fairness regarding these systems. However, there is still much progress to be made in this regard, since the explanations provided by AI systems do not always contribute to users' trust [18].

Some of the ways in which XAI contributes to reducing inequity, discrimination, and bias in AI systems are [32-36]:

- Transparency and accountability: XAI allows developers and users to understand the decision-making process of AI models. It provides a clear explanation of which characteristics or variables influence the model's predictions or decisions, helping to identify and address potential biases or discrimination.
- Identification of biases and discrimination: By explaining how each decision or prediction is reached, XAI facilitates the detection of biases inherent in the training data or algorithm design. This allows models to be adjusted and improved to reduce the impact of discriminatory factors, such as race, gender, or ethnicity.
- Clarity in the interpretation of results: XAI provides understandable explanations of how specific conclusions are reached, helping practitioners and stakeholders assess the validity and fairness of the decisions made by the system.
- Increased trust and acceptance: Transparency and explainability improve public and professional trust in AI systems [37]. When users understand how and why a model makes decisions, they are more willing to adopt it and use it ethically and responsibly.
- Improving Policy and Regulatory Design: XAI provides critical evidence to inform the design of policies and regulations that mitigate the risks of discrimination and bias in AI systems.

This includes establishing ethical standards and governance practices that promote fairness and justice in the application of the technology.

- User Training and Empowerment: By better understanding how AI works and how to interpret its outputs, end-users and practitioners can make more informed and ethical decisions. This includes adjusting parameters, selecting appropriate training data, and ensuring that AI systems are used responsibly in different contexts.

In short, explainable AI plays a crucial role by providing transparency and clarity about the decisions of AI models, which significantly helps mitigate the inequity, discrimination, and bias inherent in these technologies [38]. This not only improves public trust and acceptance but also promotes more ethical and responsible use of artificial intelligence for the benefit of society as a whole.

However, several dilemmas remain to be resolved in this area. In particular, we can refer to the five dilemmas of fairness in AI systems [35]:

- Fairness versus performance. Trade-offs are sometimes necessary to obtain an impartial, well-performing model.
- (Dis)agreement and incompatibility regarding “Fairness.” There is no consensus in the literature on whether individual or group fairness should be prioritized.
- Tensions with context and policy. There is an argument that, rather than striving to “minimize” unfairness, greater awareness of context-based aspects of discrimination is needed. In this regard, there is an additional challenge in the datasets used to train AI models because the data represent past decisions, and as such, the biases inherent in these decisions can be amplified.
- Democratizing machine learning to address the fairness skills gap. Lowering the barrier to using machine learning through democratization may increase (un)intentional and socially insensitive uses of machine learning technologies. Furthermore, the justice field has relatively few open-source tools available and there are few adaptations for different levels of technical proficiency.
- Scientific progress versus data reality. As data preparation is central to machine learning, the justice literature is slowly showing research that considers forms of data preparation that are not justice interventions per se. One example is the treatment of missing data. Much of the existing research assumes (often implicitly) that data are complete and clean, but realistically, this will never be the case.

Moreover, the "harmful" use of AI raises serious ethical and legal concerns about who is responsible for actions carried out by autonomous or automated systems [39]. Therefore, it is crucial to develop robust regulatory frameworks that mitigate the risks associated with the misuse of AI, ensuring that adequate safeguards are in place to protect individuals and organizations from potential abuse. To protect citizens from these risks, some countries have developed regulations governing the use of Artificial Intelligence. Below is an overview of the challenges and opportunities that exist in Mexico in this area.

3 AI for Human Resources Selection and Promotion – Impact on the Gender Gap

Gender bias in artificial intelligence (AI)-based decision-making systems is considered a sociotechnical problem [40]. This bias has been widely documented in talent selection and promotion systems and has garnered media attention, such as the case of Amazon's algorithm favoring male candidates [41] or the dismissal of British makeup artist Anthea Mairoudhiou following an automated HireVue body language analysis [12].

Recent research confirms these trends. Rao and Zhao [42] analyzed more than 70,000 applicants in the information technology sector, showing that women are less likely to be hired than men. In addition, Yarger et al. [43] warn that even algorithms designed for equity purposes are not neutral and require audits to identify morally

or legally problematic decisions. Fabris et al. [44] argue that historical patterns in training data can generate systematic disadvantages for certain groups due to sensitive attributes, such as gender, age, disability, religion, ethnicity, or sexual orientation, thus perpetuating historical discrimination and widening the gender gap [45].

These findings underscore the need to employ diverse training data, implement continuous monitoring for bias detection, and establish ethical guidelines throughout the development of AI systems [46]. They also highlight the importance of interdisciplinary teams that include ethicists, as well as conducting external audits and promoting transparency to ensure fairness and maintain public trust [47;48].

Recently, Tusell-Rey et al. proposed a novel artificial intelligence algorithm for machine learning, the Evaluative Customized Naïve Associative Classifier (ECNAC), and showcased its capabilities for automatic assessment of human resource promotion [49]. The research showed interesting gender differences, and provided insights into how to design artificial intelligence models according to the EAI paradigm.

4 Challenges and Opportunities of AI in Mexico

Regulations are essential to promote the safe and ethical use of Artificial Intelligence in modern society. Establishing clear and rigorous regulatory frameworks helps mitigate potential risks of negative impacts on employment and public safety. Furthermore, regulations promote transparency and accountability surrounding the development, implementation, and operation of AI systems, ensuring that human rights are respected and risks to users and society at large are minimized.

Likewise, regulations encourage responsible innovation by establishing ethical standards that guide the research and deployment of AI technologies, thus facilitating the development of an environment where the technology can equitably benefit all sectors of society without compromising the safety or integrity of individuals.

In the case of Mexico, the National Alliance for Artificial Intelligence (ANIA) has monitored legislative initiatives that in some way consider regulating AI within the country [50]. The alliance has 15 initiatives in the Senate of the Republic and 25 in the Chamber of Deputies. It should be noted that the first registered initiative dates back to 2021.

Below, some of the elements of interest in these initiatives and/or legislation related to AI are presented as examples. Initiative LXV/3SPO-87-3353/140653 [51] adds various provisions to the General Law on Women's Access to a Life Free of Violence and consequently introduces reforms to the Federal Penal Code. This initiative proposes to include in current legislation the concept of "digital falsification media," understood as the technology based on AI and digital processing used to create and alter audiovisual content, that is, images, videos, sounds, texts, and virtual news that appear to be true but whose content is false.

Furthermore, this initiative proposes to criminalize "digital sexual violence," considering that it is committed by "any person who, through the use of information and communications technologies, including the use of digital falsification means, generates, exposes, distributes, disseminates, exhibits, transmits, markets, offers, exchanges, or shares real or simulated images, audios, or videos of intimate sexual content of a person, without their explicit consent or express authorization."

The initiative itself clearly demonstrates that, currently, in Mexico, it is possible for a malicious person to generate, using artificial intelligence, for example, a pornographic video of a colleague, ex-partner, or another person, and distribute it, without this being classified as a crime. Thus, the victim is currently unprotected and does not even have the possibility of reporting their aggressor, because this type of crime does not exist in the penal code. Linked to the above, and even more serious, is the fact that Mexican citizens do not have the right to their own digital image. The initiative "Reforms to Guarantee the Right of Citizens to Their Own Image" [52] seeks to resolve this situation by amending Article 87 of the Federal Copyright Law.

The current law only considers a "portrait" as an image of a person, thus excluding digital photographs, avatars, and images generated by AI or "any graphic representation of a person's image made by other means." According to the initiative, the legal precedent is given in "the ruling by the First Chamber in direct amparo number 7/2022 on February 8, 2023, where the word "portrait," provided for in Article 87 of the Federal Copyright Law, should not be interpreted restrictively to consider only portraits but also photographs or drawings that evoke the person, regardless of the technique used to reproduce them."

Another element of interest in Mexico is that there is currently no legal framework regulating the use of AI and protecting citizens from discrimination, algorithmic bias, false representation, digital identity theft, or any other use that undermines their rights as human beings.

To our knowledge, it was not until May 2023 that the first proposal to generally regulate the use of AI in Mexico was presented in the Chamber of Deputies through the initiative to create the Law on the Ethical Regulation of Artificial Intelligence and Robotics [53]. This bill sought to regulate the use of these technologies in government, economic, commercial, administrative, communication, and financial sectors to ensure that they are used ethically and legally.

The intention of this proposal was to establish public policies in Mexico to ethically regulate the use of AI and robotics within the country. Furthermore, this initiative sought to promote the creation of Mexican Official Standards based on ethical principles to ensure the beneficial use of AI and robotics in Mexican society, respecting human rights, gender equality, and prohibiting any form of discrimination.

This bill proposed the creation of the Mexican Ethics Council for AI and Robotics as a decentralized public body. It also proposed establishing the National Network of Statistics on the Use and Monitoring of Artificial Intelligence and Robotics, as well as involving autonomous organizations in regulating its use in Mexico. The National Institute of Statistics and Geography would be designated as the primary entity generating information on the use of AI in Mexico.

The text of this proposal recognizes the international concern to establish appropriate regulations and legal frameworks for AI, considering its potential ethical and legal implications. The importance of addressing the challenges related to the potential misuse of these technologies is highlighted, emphasizing that, although AI is constantly evolving, the human being must always be considered the center of the law. However, this initiative was not approved.

On the other hand, in September 2023, the first initiative was presented that would empower the Congress of the Union to legislate on Artificial Intelligence [54]. This initiative proposed the reform of Section XVII of Article 73 of the Political Constitution of the United Mexican States, in order to grant the Congress of the Union the power to issue the necessary regulations to regulate the research, development, and applications of AI. In particular, the proposed text states: “XVII. To enact laws on general means of communication, artificial intelligence, information and communication technologies, broadcasting, telecommunications, including broadband and the Internet, postal services and mail, and on the use and exploitation of waters under federal jurisdiction.”

Subsequently, in February 2024, the second proposal for AI regulation was introduced in the Mexican Senate, this time in the form of the proposed Federal Law Regulating Artificial Intelligence [55]. Notably, this initiative contemplates extraterritorial application, meaning that compliance would be mandatory even for providers of Artificial Intelligence Systems (AIS) established abroad that offer services in Mexico or whose generated data is used in the country.

This initiative proposes that the Federal Telecommunications Institute (IFT) would be the authority responsible for regulating and authorizing AIS providers. Furthermore, it proposes the creation of a National Commission on Artificial Intelligence, which would act as an advisory body to the IFT and be composed of scientific experts in the field. The proposed regulation seeks to classify AIS according to the risks they may pose, adopting an approach similar to that used by the European Union (EU) regulation. Three risk levels are established:

"Unacceptable Risk," "High Risk," and "Low Risk," each with specific characteristics and requirements. Finally, prior authorization from the IFT is mandatory for commercializing AI in Mexico, even for those offered free of charge, and fines are imposed on violators of these provisions.

In May 2024, the "National Artificial Intelligence Agenda for Mexico 2024-2030" [56] was presented to the Senate. The document seeks to establish a set of principles that promote the integration of AI as a catalyst for the country's inclusion and social, economic, and educational development. It also seeks to encourage the ethical and equitable use of this tool as a means of driving scientific research, technological development, innovation, and entrepreneurship.

This proposal includes recommendations on public policies and rights, education and labor markets, cybersecurity and risk management, gender, inclusion and social responsibility, infrastructure and data, as well as innovation, research, and industry. It also establishes a comprehensive framework that incorporates public policies, specific regulations, and governance strategies. This framework is designed to be collaborative, involving multiple actors, disciplines, and sectors of society.

These elements reflect the current concern of some Mexican legislators regarding the dangers of the use and adoption of AI without adequate legal support and an attempt on their part to comprehensively and thoroughly regulate the use and commercialization of AI in the country, ensuring the existence of a legal framework that protects both users and national interests in this area.

5 Conclusions

This article explores the impacts of AI development on modern societies and how this technology itself offers possibilities to address the main challenges posed by its use. In this regard, the XAI and EAI paradigms were addressed, specifically for their benefits in mitigating discrimination and biases in the most sensitive areas where these technologies have an impact. These issues could become even more unfair if ignored during the design and use of AI tools. The main challenges in regulating the use of AI to enhance good practices and exploit its potential without great harm were addressed in the context of Mexico. This demonstrates that this developing country has taken some steps, but that much remains to be done in this regard. Advances in AI are becoming more known and enjoyed every day; however, as its acceptance and adoption in different spheres grow, the challenges associated with its use also increase.

Acknowledgments

The authors would like to thank Instituto Politécnico Nacional (Secretaría Académica, Secretaría de Investigación y Posgrado, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Centro de Investigación en Computación, Red de Computación, Red de Inteligencia Artificial y Ciencia de Datos), the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SCHITI), and the Sistema Nacional de Investigadoras e Investigadores (SNII), for their support in developing this work.

References

1. **Schmitt, M., Flechais, I. (2024)**. Digital Deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review, 57(12), 324.
2. **Ai, L., Kumarage, T., Bhattacharjee, A., Liu, Z., Hui, Z., Davinroy, M., . . . Kirchner, M. (2024)**. Defending Against Social Engineering Attacks in the Age of LLMs. arXiv preprint arXiv:2406.12263.
3. **Peters, K. (2019)**. 21st century crime: How malicious artificial intelligence will impact homeland security. (Master of Arts in Security Studies (Homeland Security and Defense)), Naval Postgraduate School,
4. **Blauth, T. F., Gstrein, O. J., Zwitter, A. (2022)**. Artificial intelligence crime: An overview of malicious use and abuse of AI. Ieee Access, 10, 77110-77122.
5. **Di Nicola, A. (2022)**. Towards digital organized crime and digital sociology of

organized crime. Trends in organized crime, 1-20.

6. **Karnouskos, S. (2020)**. Artificial intelligence in digital media: The era of deepfakes. IEEE Transactions on Technology and Society, 1(3), 138-147.
7. **Botha, J., Pieterse, H. (2020)**. Fake news and deepfakes: A dangerous threat for 21st century information security. Paper presented at the ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited.
8. **Ali, A., Ghouri, K. F. K., Naseem, H., Soomro, T. R., Mansoor, W., Momani, A. M. (2022)**. Battle of deep fakes: Artificial intelligence set to become a major threat to the individual and national security. Paper presented at the 2022 International Conference on Cyber Resilience (ICCR).
9. **Hardesty, L. (2018)**. Study finds gender and skin-type bias in commercial artificial-intelligence systems. MIT News.
10. **Cavazos, J. G., Phillips, P. J., Castillo, C. D., O'Toole, A. J. (2020)**. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE transactions on biometrics, behavior, and identity science, 3(1), 101-111.
11. **Lohr, S. (2022)**. Facial recognition is accurate, if you're a white guy. In Ethics of Data and Analytics (pp. 143-147): Auerbach Publications.
12. **Lytton, C. (2024)**. AI hiring tools may be filtering out the best job applicants. BBC. Retrieved from https://www.bbc.com/worklife/article/20240214-ai-recruiting-hiring-software-bias-discrimination
13. **13. Curry, R. (2023)**. In job hiring process, most workers say they already sense AI, but the bias issue is far from solved. CNBC Technology Executive Council. Retrieved from https://www.cnbc.com/2023/12/28/in-the-job-hiring-process-most-workers-say-they-already-sense-ai.html
14. **Dastin, J. (2022)**. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of data and analytics (pp. 296-299): Auerbach Publications.
15. **Sipior, J. C., Ward, B. T., Rusinko, C. A., Lombardi, D. R. (2023)**. Bias in Using AI for Recruiting: Legal Considerations. Information Systems Management, 1-14.
16. **Heaven, W. D. (2020)**. Predictive policing algorithms are racist. They need to be dismantled. MIT Technology Review, 17, 2020.
17. **National Association for the Advancement of Colored People (NAACP). (2024)**. Artificial Intelligence in Predictive Policing Issue Brief [Press release]. Retrieved from https://naacp.org/resources/artificial-intelligence-predictive-policing-issue-brief#:~:text=Jurisdictions%20who%20use%20this%20tool,public%20trust%20in%20law%20enforcement.
18. **Schoeffer, J., De-Arteaga, M., Kuehl, N. (2024)**. Explanations, fairness, and appropriate reliance in human-AI decision-making. Paper presented at the CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, HI, USA.
19. **ACM (2024)**. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT). Retrieved from https://facctconference.org/
20. **Veale, M., Van Kleek, M., Binns, R. (2018)**. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. Paper presented at the Proceedings of the 2018 chi conference on human factors in computing systems.
21. **De Schutter, L., De Cremer, D. (2024)**. How counterfactual fairness modelling in algorithms can promote ethical decision-making. International Journal of Human–Computer Interaction, 40(1), 33-44.
22. **Zhang, J., Shu, Y., Yu, H. (2023)**. Fairness in design: A framework for facilitating ethical artificial intelligence designs. International Journal of Crowd Science, 7(1), 32-39.
23. **Plečko, D., Bareinboim, E. (2024)**. Causal fairness analysis: a causal toolkit for fair machine learning. Foundations and Trends® in Machine Learning, 17(3), 304-589.

24. Tang, H., Cheng, L., Liu, N., Du, M. (2023). A Theoretical Approach to Characterize the Accuracy-Fairness Trade-off Pareto Frontier. arXiv preprint arXiv:2310.12785.

25. Monjezi, V., Trivedi, A., Tan, G., Tizpaz-Niari, S. (2023). Information-theoretic testing and debugging of fairness defects in deep neural networks. Paper presented at the 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE).

26. Cornacchia, G., Anelli, V. W., Biancofiore, G. M., Narducci, F., Pomo, C., Ragone, A., Di Sciascio, E. (2023). Auditing fairness under unawareness through counterfactual reasoning. Information Processing Management, 60(2), 103224.

27. Consuegra-Ayala, J. P., Gutiérrez, Y., Almeida-Cruz, Y., Palomar, M. (2024). Automatic Annotation of Protected Attributes to Support Fairness Optimization. Information Sciences, 120188.

28. Goyal, N., Baumler, C., Nguyen, T., Daumé III, H. J. a. p. a. (2023). The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features. Paper presented at the 29th International Conference on Intelligent User Interfaces (IUI '24), Greenville, SC, USA.

29. Shulner-Tal, A., Kuflik, T., Kliger, D. (2023). Enhancing fairness perception–Towards human-centred AI and personalized explanations understanding the factors influencing laypeople's fairness perceptions of algorithmic decisions. International Journal of Human–Computer Interaction, 39(7), 1455-1482.

30. Ochmann, J., Michels, L., Tiefenbeck, V., Maier, C., Laumer, S. (2024). Perceived algorithmic fairness: An empirical study of transparency and anthropomorphism in algorithmic recruiting. Information Systems Journal.

31. Fahnenstich, H., Rieger, T., Roesler, E. (2024). Trusting under risk–comparing human to AI decision support agents. Computers in Human Behavior, 153, 108107.

32. Akinrinola, O., Okoye, C. C., Ofodile, O. C., Ugochukwu, C. E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. GSC Advanced Research Reviews, 18(3), 050-058.

33. Barocas, S., Hardt, M., Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities: MIT Press.

34. Miller, T. (2023). Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. Paper presented at the Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.

35. Caton, S., Haas, C. (2024). Fairness in machine learning: A survey. ACM Computing Surveys, 56(7), Article 166.

36. Le, T., Miller, T., Singh, R., Sonenberg, L. (2024). Towards the new XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence. arXiv preprint arXiv:.01292.

37. Dlugatch, R., Georgieva, A., Kerasidou, A. (2023). Trustworthy artificial intelligence and ethical design: public perceptions of trustworthiness of an AI-based decision-support tool in the context of intrapartum care. BMC Medical Ethics, 24(1), 42.

38. Barrera Ferro, D., Brailsford, S., Chapman, A. (2024). Improving fairness in machine learning-enabled affirmative actions: a case study in outreach activities in healthcare. Journal of the Operational Research Society, 1-12.

39. Farayola, O. A., Olorunfemi, O. L. (2024). Ethical decision-making in IT governance: A review of models and frameworks. International Journal of Science Research Archive, 11(2), 130-138.

40. Nadeem, A., Marjanovic, O., Abedin, B. (2022). Gender bias in AI-based decision-making systems: a systematic literature review. Australasian Journal of Information Systems, 26.

41. Dastin, J. (2018). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved from https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-

showed-bias-against-women-idUSKCN1MK0AG/

42. **Rao, S., Zhao, T. (2025)**. Ethical AI in HR: A Case Study of Tech Hiring. Journal of Computer Information Systems, 1-18.
43. **Yarger, L., Cobb Payton, F., Neupane, B. (2020)**. Algorithmic equity in the hiring of underrepresented IT job candidates. Online information review, 44(2), 383-395.
44. **Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., . . . Biega, A. J. (2025)**. Fairness and bias in algorithmic hiring: A multidisciplinary survey. ACM Transactions on Intelligent Systems and Technology, 16(1), 1-54.
45. **Jackson, M. C. (2021)**. Artificial intelligence algorithmic bias: the issues with technology reflecting history & humans. J. Bus. & Tech. L., 16, 299.
46. **Du, J. (2024)**. Exploring gender bias and algorithm transparency: Ethical considerations of AI in HRM. Journal of Theory and Practice of Management Science, 4(03), 36-43.
47. **Yam, J., Skorburg, J. A. (2021)**. From human resources to human rights: Impact assessments for hiring algorithms. Ethics and Information Technology, 23(4), 611-623.
48. **Köchling, A., Wehner, M. C. (2020)**. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795-848.
49. **Tusell-Rey, C. C., Pino-Gómez, J., Villuendas-Rey, Y. (2024)**. Evaluative Customized Naïve Associative Classifier: Promoting Equity in AI for the Selection and Promotion of Human Resources. Paper presented at the International Conference on Intelligent Data Engineering and Automated Learning.
50. **Alizanza Nacional de Inteligencia Artificial (2024)**. Seguimiento Legislativo. Retrieved from https://www.ania.org.mx/seguimiento-legislativo
51. **López-Hernández, R. A. (2024)**. Proyecto de decreto por el que se reforman y adicionan diversas disposiciones de la Ley General de Acceso de las Mujeres a una Vida Libre de Violencia y del Código Penal Federal. LXV/3SPO-87-3353/140653. Retrieved from https://www.senado.gob.mx/65/gaceta_del_senado/documento/140653
52. **Ascencio-Ortega, R. C. (2024)**. Reformas para garantizar el derecho de los ciudadanos a su propia imagen. Retrieved from http://sil.gobernacion.gob.mx/Archivos/Documentos/2023/03/asun_4513056_20230314_1677182151.pdf
53. **Senado de la República (2023)**. Proyecto de decreto por el que se expide la Ley de Regulación Ética de la Inteligencia Artificial y la Robótica. LXV/2SPR-5-3226/135000. Retrieved from https://www.senado.gob.mx/65/gaceta_comision_permanente/documento/135000
54. **Bañuelos de la Torre, G., Pinedo-Alonsoa, C. C., Márquez Alvarado, M. C., Padilla-Peña, J. (2023)**. Proyecto de decreto por el que se reforma la fracción XVII al artículo 73 de la Constitución Política de los Estados Unidos Mexicanos, para facultar al Congreso de la Unión para emitir las normas necesarias para regular la investigación, desarrollo y aplicaciones de la Inteligencia Artificial. Retrieved from http://sil.gobernacion.gob.mx/Archivos/Documentos/2023/09/asun_4595174_20230906_1693930883.pdf
55. **Monreal-Avila, R. (2024)**. Proyecto de decreto por el que se expide ley federal que regula la Inteligencia Artificial. Retrieved from https://ricardomonrealavila.com/wp-content/uploads/2024/02/Inic_Morena_inteligencia_artificial.pdf
56. **Senado de la República (2024).** Presentan en el Senado Agenda Nacional de la Inteligencia Artificial para México 2024-2030. Retrieved from https://comunicacionsocial.senado.gob.mx/informacion/comunicados/8982-presentan-en-el-senado-agenda-nacional-de-la-inteligencia-artificial-para-mexico-2024-2030

Article received on 11/06/2025; accepted on 20/10/2025.
**Corresponding author is Oscar Camacho Nieto.*