International Journal of Finance, Economics, and Management Studies
ISSN: 3108-0030 | Volume 2 Issue 3, 84-93, July-September 2026
DOI: https://doi.org/10.64137/31080030/IJFEMS-V2I3P110

Received: 15/07/2026 Revised: 08/08/2026 Accepted: 20/08/2026 Published: 24/08/2026

Original Article

Wage Disparities and Gender Gap in STEM Fields – Investigating the Long-term Impact with AI

AVANI MALHOTRA
Delhi Public School, Mathura Road, New Delhi, India.

ABSTRACT: *This study explores how artificial intelligence (AI) and other technological advances can align with the gender-based wage gap and their implications for job opportunities over time. Irrespective of changes in technology and social norms, the study highlights that there are significant concerns related to the potential of AI to either increase or mitigate current gender inequalities. Though there are opportunities available in AI for fairness and decision-making, there are significant risks that poor representation in STEM and biased algorithms may worsen the gender gap. This study is based on a critical analysis of the ethical and structural challenges in deploying AI and the overlooking of women in the governance and design of AI programs. There is also an uneven effect of automation on industries led by women. With a systematic review approach, this study has identified various sources of AI bias, their implications, persistent gaps in the promotion and participation of women in STEM fields, and both policy-based and technical ways to arrive at more equitable outcomes as AI becomes increasingly integrated into society.*

KEYWORDS: *AI Bias, STEM Fields, Gender Gap, Artificial Intelligence, Job Opportunities, Gender Inequalities*.

1. INTRODUCTION

Over the past decades, there has been a dramatic shift in norms, values and language. Uncertainty related to automation has always been relevant. Nobel laureate Herbert Simon expressed the opinion in 1956 that machines would be able to do anything a man can do within 20 years. Hence, a lot of new technologies would emerge, and many jobs would become obsolete, especially blue-collar jobs in the manufacturing sector. Computers have been around for those born around the 2000s, especially in the office, at home, and at the bank. Today, computers are part of restaurants and are used for simply ordering food. These days, one cannot think of jobs that could not be affected by these machines and computers (Ernst et al., 2019).

Unlike Prof. Simon, no one fears computers anymore. These days, there are fears around the capabilities of machines to predict using big data and act in unstructured or complex environments, such as artificial intelligence (Agrawal et al., 2018). Under uncertainty, complex decision-making is at the core of economies. Whether an employee is choosing the right career or job, a manager is running day-to-day operations, or a consumer is deciding which services or products to use, there are complex and constantly interrelated issues that require natural intelligence to be well prepared.

No machine has recently been capable of matching human intelligence potential, even though the concept of smart machines emerged with the invention of the computer in the 1930s. Long before the invention of silicon computers, Alan Turing and Alonzo Church found in 1936 that any kind of reasoning can be simulated by digital devices, including problems in management and economics. There might come a time when humans cannot differentiate between talking to a chatbot and a human (Turing, 1950). Given the recent experiences of top AI firms, that time no longer appears far away.

Apart from wage disparities, gender inequality raises another important question in the world of artificial intelligence (AI), which is redefining the very fabric of society. AI is now itself a game changer that will alter the educational framework as we know it, along with work and our interactions in society significantly. However, this revolution in technology can also exacerbate and entrench those existing gender inequalities. With the onset of the AI era, we have to analyse potential effects positive or negative on gender bias and predict future direction as well (Yu, 2024).

AI poses a different range of implications with regard to the structure and dynamics of employment; indeed, AI has been seen as either deepening or closing gender gaps. AI programs are no longer confined to a single sector; they span three different sectors. Moreover, what they do alters both gender expressive outcome paths how they are portrayed in these roles and sectors. Women are usually underrepresented in "science, technology, engineering, and mathematics (STEM)" fields, which are the core of managing and developing AI approaches (Yu, 2024).

There are concerns about the participation of women in shaping and creating AI programs. By relying solely on data analysis, with the potential for AI models to reinforce and reflect gender biases that are already embedded in the training set due to a

This work is an open access and licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License
(https://creativecommons.org/licenses/by-nc-nd/4.0/)

lack of diversity within those perspectives, discrimination can occur in project assignment decisions as well as hiring practices and promotions. Furthermore, areas of activity with larger proportions of those working in customer service, health and education are most at risk from AI-based efficiencies and automation. The displacement arising from automation could affect women disproportionately, leading to widening gaps in terms of income and jobs. Work in New Job Roles Created or Promoted by AI: If women are not successful at transitioning into new job roles created or promoted through the use of artificial intelligence, they may still be relegated to lower-paying jobs and even pushed out of work altogether (Yu, 2024).

Working in most gig jobs has very few of the protections or certainty offered to full-time employees, which AI platforms have heavily mediated. Although gig work is often pursued by women due to its added flexibility, it can lack stability and opportunities for progression. Such precarity can intensify the gender pay gap and restrict women's economic empowerment (Yu, 2024).

2. LITERATURE REVIEW

Many emerging technologies bring with them a host of economic challenges that deserve scrutiny, and AI is no different. While AI has a bright future, in that its potential is undisputed, many big questions have emerged regarding the social structure and wealth distribution. Challoumis (2024) makes the argument that as processes once performed by humans are being taken over more and more by machines, they should be critically examined against macroeconomic inequality factors, and addresses points on what to do about such imbalances in order for everyone to have a better future.

The global dimension is examined by Tzannatos 1999), describing changes in men's and women's participation, occupational segregation and the male/female wage gap. The study provides strong evidence that gender gaps in employment are closing faster in developing than globalised nations, and estimates of the inefficiencies arising from ongoing labour market disparities associated with sex are large. According to the estimates, it is mainly women who bear most of the deadweight losses while men benefit proportionally more from ongoing arrangements. It finds that market-only growth is a poor instrument for tackling gender inequality and suggests improving access to education and training, equal job opportunities and pay for both genders, together with a benefits-and-taxation system that reflects reproduction as a legitimate contribution to the economy.

AI and its implications on sustainable development have also been evaluated in the context of other industries. Based on a consensus-based assessment, AI is found to help achieve 134 targets and have adverse effects across all goals for 59 targets. Therefore, regulatory insight and oversight should keep pace with the rapid development of AI to ensure algorithms develop at a sustainable rate; neglecting to engage in proactive regulation will further widen unrivalled gaps in safety, transparency and ethical standards (Vinuesa et al., 2020).

Emerging economies also face particular risks to the achievement of the UN SDGs due to their vulnerability here, as well as from unregulated roll-out of experimental AI by Big Tech. Unregulated and poorly designed decision-making processes pose a risk to financial inclusion, influencing the important finances of users. Automated decision-making models are not only biased, opaque and poorly governed creating uneven access to finance but have also produced discrimination. Emerging economies are at risk of being exploited by both à la carte harvesting of profits and data through AI, threatening the SDGs and poverty reduction progress on avoiding corruption/financial crime is compromised as a result. Truby (2020) argues that the promotion of mega corporations creates a highly questionable rationale to operate as unregulated entities and proposes global standards for anticipatory AI governance in which coders/developers would be liable, while addressing SDGs-aligned parameters and processes via ideological aspects connected with any features related to outer-space threats. The study also warns that overregulating the technology, and subsequent backlash against it in a climate of global economic competition, could harm society's system-wide capture of important benefits from AI.

Advances in technology, particularly AI, have also been connected with increasing inequality and unemployment. Goyal and Aneja (2020) study the increasing income inequality along with technological change using Gini coefficients and automation data to find a direct as well as indirect effect of AI technology on work processes, place and worker position. Our new study shows the robust synergistic result of unemployment, rising income inequality between high-skill and middle-skill labour sectors, as well as a consistent negative relationship regarding distributional effects on active AI proliferation among jobs operating at different levels. Gini coefficients are higher for developing countries compared to developed countries, indicating that the income gap in AI-applying countries is more serious in lower economic stages.

2.1. RESEARCH OBJECTIVES

- To investigate the biases of AI responsible for inequalities in the job market.
- To identify the widening risks to gender parity in STEM fields.
- To discuss potential solutions for AI to reduce the gender gap in the job market.

3. RESEARCH METHODOLOGY

A systematic review approach was adopted to fulfil the objectives of this study. This perspective aids in systematising prior literature related to the topics examined here, and helps construct relations among the different approaches used by researchers examining gender equity within a common wage gap framework. Additionally, it highlights consistent patterns in AI research (Lim et al., 2022), and identifies fertile areas for future investigation [29].

A systematic review approach can help to summarise areas where additional research is required (Snyder, 2019). The approach applied here is based on the methodology suggested by Tranfield et al. (2003), which is a three-stage model: data gathering, analysis and reporting. The literature search for AI is based on Web of Science, Scopus and other databases. The search, mostly applied to the abstract, title and keywords of each source, employed such key terms as wealth generation; wage gap; income inequality; gender gap; sustainable development goals; responsible innovation; gender inequalities; decarbonisation, etc. Only reviews, conference proceedings, book chapters, and articles published in peer-reviewed journals were selected for analysis.

4. DATA ANALYSIS

4.1. BIASES OF AI RESPONSIBLE FOR INEQUALITIES IN THE JOB MARKET

AI can redefine many industries and improve the overall lives of people in different ways. However, the presence of bias is one of the most significant challenges facing the deployment and development of AI programs. Bias refers to systematic errors occurring in the process of decision-making, resulting in unfair outcomes. Bias in AI can come from algorithm design, data collection and human interpretations. AI system ML models can replicate patterns and learn biases that are in the data on which they are trained, ultimately producing biased or unfair results.

Different sources of bias have been studied as algorithmic bias, data bias, and user or human-related biases, with examples from the real world. Bias is the systematic error in a decision process that leads to unfair outcomes, and it can come from various levels of an ML pipeline (Crawford & Calo 2016; Selbst et al.).

The notion that using incomplete or misleading data to train models contributes to biased outputs refers to the concept of data bias. This may occur when the data is insufficient, collected from biased sources, contains inaccuracies or omits crucial information. On the other hand, algorithmic bias can arise when machine learning algorithms are based on prejudices built in—for example, models that rely on biased assumptions or decision-making criteria(1). User bias occurs when those who use AI systems impose their own prejudices into the system, either intentionally or unintentionally: through using biased training data and/or via their interactions with the system.

Various strategies have been proposed to alleviate these types of bias, such as introducing bias-aware models, data augmentation and user-feedback systems. Data Augmentation: This process helps in adding more varied data into training datasets, including additional measures for helping bias and providing better representation over time. In bias→aware models, algorithms are created that anticipate biases and lessen their impact on the outputs of larger systems. Such systems ask their users for feedback in order to detect and repair biases. This research remains ongoing, and dimensions are growing to build black-box AI programs that are fairer.

Misleading Models: A number of instances have been documented throughout the health care industry as well, and across various sectors of business. In a similar study conducted in Wisconsin, the COMPAS system was shown to overestimate the risk posed by African-American offenders who had not previously been convicted of any offence; they were more likely than white offenders to be labelled as high-risk (Angwin et al., 2022). Another example, and indeed already an AI program in use for predicting patient mortality rates at the hospital level, was found to be biased against African-American patients, who were more frequently given higher risk scores compared with their non-Black counterparts even when health status/age co-factors are statistically accounted for (Obermeyer et al., 2019); such bias directly impacts treatment provision and access. Facial recognition: Another well-documented example of bias relates to AI programs used by law enforcement for facial recognition, which are often less accurate for African-American people and people with darker skin types, resulting in a higher rate of false positives and an increased risk that innocent individuals will be wrongly arrested or convicted as criminals (Schwartz et al., 2022).

As the capabilities of generative AI (GenAI) systems have improved, so has the danger posed by bias. For example, large Generative Adversarial Networks (GANs) based text-to-image GenAI systems such as DALL-E (OpenAI), Stable Diffusion and Midjourney portray stereotypical/racial biases were reported (Mittelstadt et al., 2016). According to the gender bias model, when they prompted it to produce images of CEOs, it exclusively generated males; similarly, with its criminals or terrorists prompts, it disproportionately generated people of colour.

This shows the danger of GenAI systems, which are trained on pictures out in the world, just to suck up and recreate all these disparities present. This highlights the importance of balanced and varied datasets in AI development to achieve fair representation. Table 1 summarises examples of AI bias types and their impacts, which require mitigation and evaluation strategies.

TABLE 1 Types of AI biases and examples

Types of AI biases	Description	Examples
Algorithm	Caused by the implementation and design of a model that prefers certain attributes, resulting in unfair outcomes.	A model that prefers a particular gender or age, resulting in unfair hiring practices.
Representation	Occurs when the population is not accurately represented in the model, resulting in incorrect predictions.	A medical dataset that under-represents women, causing inaccurate diagnoses for women.
Sampling	Training data does not represent the target population, resulting in biased predictions for some groups.	A facial recognition program that performs poorly on people of other races because it was trained mainly on white faces.
Confirmation	Occurs when an AI model confirms existing beliefs or biases held by its users or creators.	An AI system that predicts the success of job candidates on the basis of hiring managers' existing biases.
Generative	Common in GenAI models used to create synthetic text, data, or images; emerges when model outputs disproportionately reflect certain attributes, patterns, or perspectives present in the training data, causing unbalanced or skewed generated content.	A text-generation model trained mainly on Western literature that over-represents Western idioms and norms while misrepresenting or under-representing other cultures; similarly, an image-generation model trained with limited human diversity may struggle to represent a range of ethnicities accurately.
Measurement	Occurs when data measurement or collection systematically under- or over-represents certain groups.	A survey-collection bot that gathers more responses from urban populations and under-represents rural groups.
Interaction	AI systems become biased through their interactions with humans, causing unfair treatment.	A chatbot that gives different responses to women than to men, resulting in biased communication.

(Source: Self, compiled from Crawford and Calo, 2016; Mittelstadt et al., 2016; Selbst et al., 2019)

4.1.1. ETHICAL AND SOCIETAL IMPLICATIONS OF AI BIAS

The fast-paced development of AI has provided numerous advantages as well, but it also represents possible hurdles and dangers. One of the biggest concerns about AI is bias and its negative impact on society and upon individuals. Bias can reinforce and even amplify existing inequities, leading to discrimination against underserved populations, which harms their access to essential services. In addition to this, it can also reproduce current discrimination and stereotypes and generate new discriminatory practices based on economic status, physical appearance, ethnicity or skin colour. Thus, the identification and mitigation of AI bias is crucial in order to create equitable systems that treat all users fairly.

To begin with, biased AI has different ethical implications, including the risk of discrimination, accountability by policymakers and developers in implementing algorithms substitute, impact on human autonomy/agency, as well as erosion of public trust in technology overall. This implies an all-stakeholder effort to do something about these implications, and it is worthwhile exploring regulatory frameworks as well as ethical guidelines that foster transparency, fairness and accountability around AI usage/development. This is important due to the prevalence of discrimination, as AI bias can reinforce or amplify pre-existing inequalities (Sweeney 2013), but also because in areas such as justice, biased models may result in partial treatment against specific groups and harsher punishment of individuals with particular characteristics, whether they be sex-based aspects or ethnicity-based attributes (Buolamwini and Gebru 2018).

AI bias also affects people's access to essential services, such as finance and healthcare. In credit-scoring models, for instance, biased models can result in under-representation of groups like people of colour from poorer backgrounds (Dwork et al., 2012), which puts financial services or loans out of reach for these groups. AI bias can further stimulate discrimination and stereotyping for instance, gender-biased facial recognition, where models trained on male faces only are more likely to misidentify female faces (Buolamwini & Gebru, 2018), or GenAI asked to generate a CEO image would very often provide images of men only (Mittelstadt et al.).

Apart from widening prejudice, AI bias can create new stereotyping by skin colour, ethnicity or even broader physical appearance. Gender-biased models also disproportionately present people of colour as terrorists or other criminals. The direct consequence of these systems being used at scale is severe: they can lead to refusal of jobs or provision of services, as well as wrongful arrests or convictions. This impacts how people see themselves and view others, as well as influence the way they interact with others or what opportunities lie before them. Continuously using biased systems could solidify discrimination and get in the way of being more inclusive or equal. With the widespread use of AI in everyday life, this technology may increasingly control many aspects of social structures and cultural norms; hence, it is crucial to consider these issues at every step along the way when developing an AI system (Ferrara 2022 4).

4.2. WIDENING RISKS TO GENDER PARITY IN STEM FIELDS

AI is transforming the job market, promising new avenues to address significant concerns such as generating new jobs and improving productivity. With rising demand for AI expertise, employers might be expected to draw on the widest possible talent pool to support this transition. This section focuses on the STEM field because of the limited participation of women within it. According to the World Economic Forum, just 28 per cent of women work in STEM compared with 47 per cent in non-STEM workforces. Over a third of the graduates in STEM are women, but already just 12% (World Economic Forum, 2025) have gotten to executive positions in those fields. The dangers posed by AI seem to widen the gap between men and women in STEM fields rather than shrink it. Cohen et al. 2018) provide insights into the lower presence of women in STEM as compared to non-STEM fields, with evidence from between 2016 and 2024 (Figure1).

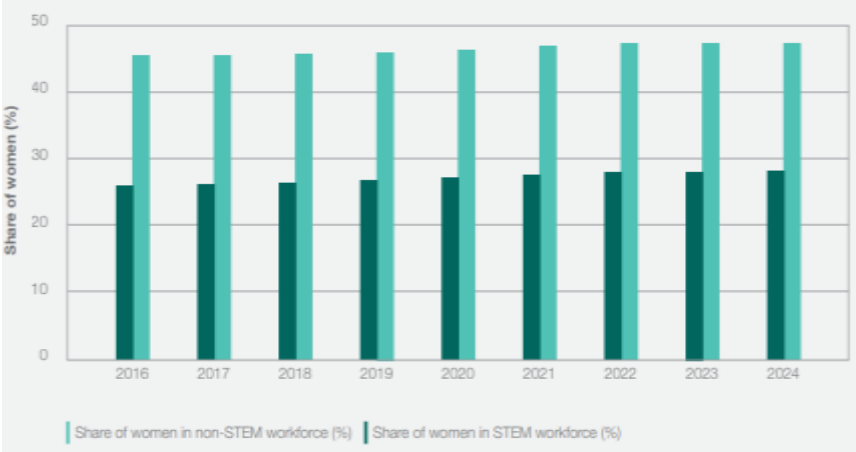

FIGURE 1 Participation of Women in the Workforce: STEM Vs Non-STEM Fields

(Source: World Economic Forum (2025))

There has been a narrowing gap between men and women in technology-related roles over time, but considerably more progress is needed to avoid losing talented women from the workforce. Women still make up only 28.2% of the STEM workforce less than a third as reported by the Global Gender Gap Report (Pal et al., 2024). While more women are graduating in STEM subjects than before, there remains a significant dropout rate from STEM careers: of the 35.5% of graduates who were women in 2017, only 29.6% of them were still working in STEM fields a year later, and this dropout rate has shown little change over time (World Economic Forum, 2025).

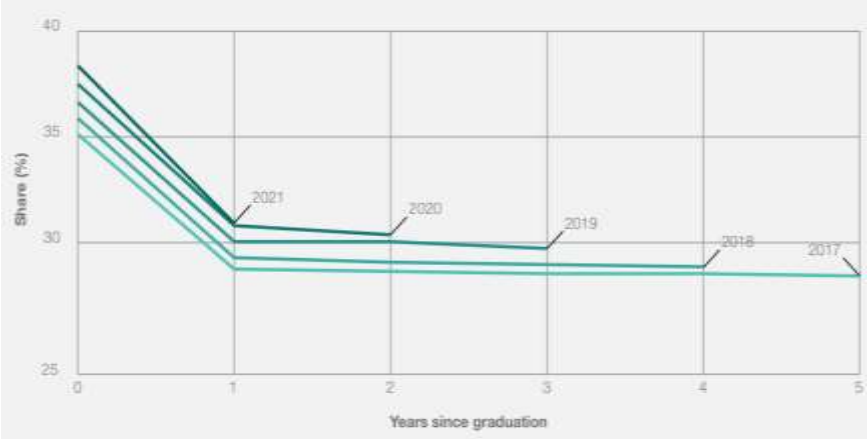

FIGURE 2 Share (%) Of Women among STEM Graduates by Year

(Source: World Economic Forum (2025))

Added to this is a "drop to the top" phenomenon (Figure 3), whereby the representation of women declines further as they rise through organisational ranks, and this decline is particularly pronounced in STEM fields. Only 24.4% of women reach managerial positions, and just 12.2% reach the C-suite. This points to an opportunity for those working in generative AI to promote gender equity in seniority and career progression (World Economic Forum, 2025).

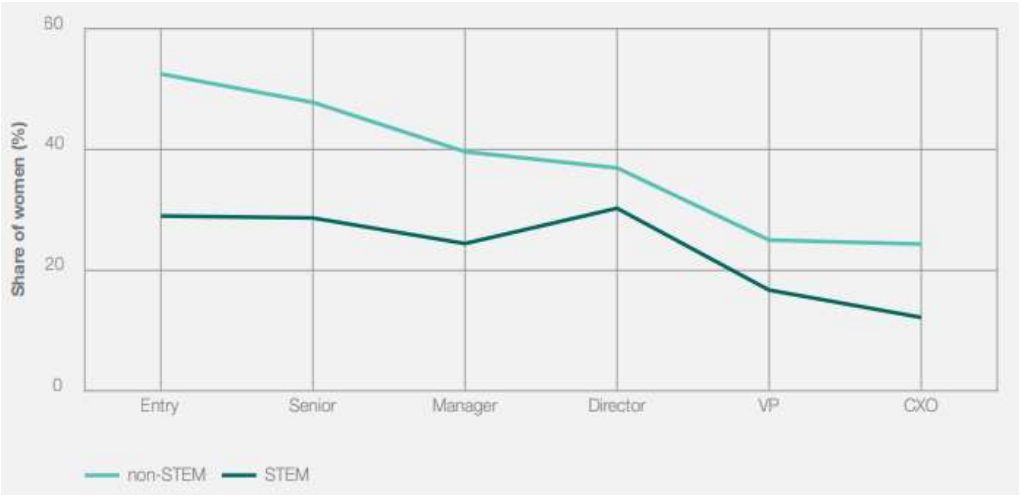

FIGURE 3 the "Drop To the Top" Phenomenon in STEM Fields versus Other Fields
(Source: World Economic Forum (2025))

Concerns are also increasing around the AI-based economy. AI can augment some existing job roles while displacing or disrupting others; some roles requiring skills that cannot be readily replicated by AI have so far remained largely unaffected. Men are more likely to work in AI-related jobs and to retain their positions, while women are more likely to occupy positions vulnerable to AI-driven change; both men and women hold a broadly similar share of niche jobs unlikely to be affected. The two genders also appear to hold somewhat different attitudes toward technology and AI according to the Global Gender Gap Report, only 54% of women believe the skills required for their jobs will change substantially over the next five years, compared with 61% of men, and women report being slightly more reluctant than men to use AI (Pal et al., 2024) (Figure 4).

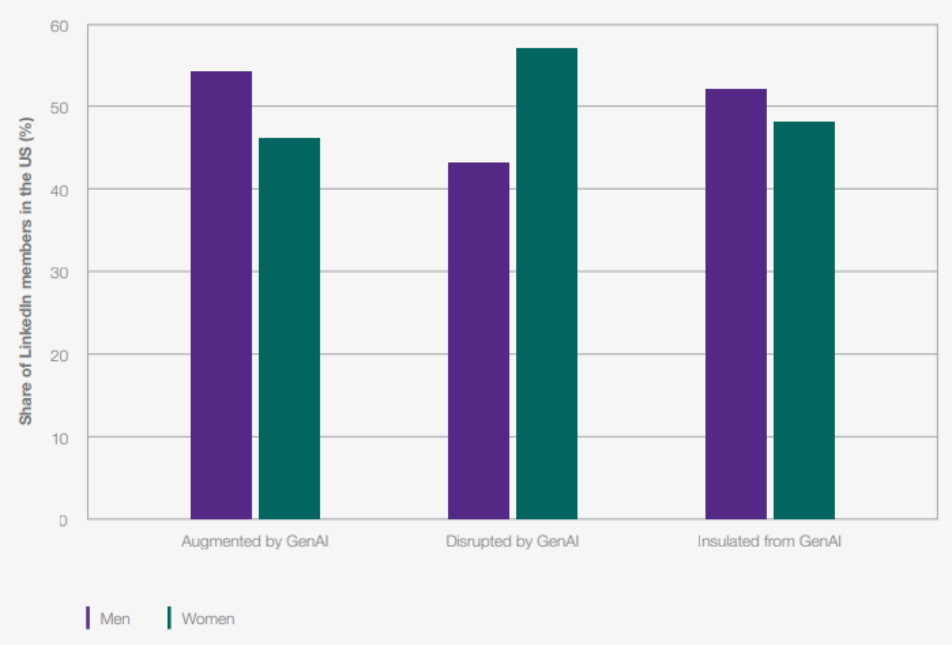

FIGURE 4 Share of Men and Women Augmented, Disrupted, and Insulated By Genai
(Source: Pal et al. (2024))

AI ranks among the top three skilling priorities for 40% of men and women worldwide, a 29% increase since 2024, according to the Randstad Workmonitor report (2025). While a gap exists between men and women in this respect (44% versus 36%), it is less pronounced when it comes to confidence in AI skills specifically. This trend is echoed in LinkedIn data: whereas only 23.5% of workers with AI-engineering skills listed on their profiles were women in 2018, that share had risen to 29.4% by

2024 (Pal et al., 2024). It will take continued momentum to close the final gap. Recent recruitment challenges mean organisations could extend their talent pool, as they embrace new technologies that integrate GenAI into workflows more aggressively this moment in time may be a turning point for groups previously discounted, with the eye towards equity and helping women gain an advantage through roles linked to AI. With these sectors advancing, it is important that policymakers work to ensure women are not left behind either (World Economic Forum 2025).

4.3. POTENTIAL SOLUTIONS FOR AI TO REDUCE THE GENDER GAP IN THE JOB MARKET

Several practitioners and researchers have proposed methods to reduce AI bias, broadly grouped into approaches based on model selection, preprocessing of information, and post-processing of decisions. Each approach carries its own challenges and limitations, including a shortage of sufficiently diverse training data, difficulty in measuring and detecting different kinds of bias, and potential trade-offs between accuracy and fairness. There are also ethical questions about which forms of bias to prioritise and which groups to focus on when reducing bias. Despite these challenges, mitigating bias in AI remains essential to building equitable systems that benefit society as a whole, and research into mitigation techniques continues.

Addressing bias in AI is fundamentally difficult, with a variety of methods suggested by researchers. One common approach is preprocessing training data so that it reflects the whole population, including groups which have been historically marginalised this covers under-sampling, over-sampling and synthetic data generation (Ferrara 2023). On the technology side, for example, Buolamwini and Gebru (2018) saw that oversampling people of colour fixed some disproportionality in facial verification systems. Preprocessing may also detect and eliminate bias in data prior to training the model using data-augmentation techniques that create synthetic positive or negative examples of under-represented groups, or biases are reduced during this preprocessing step as part of a complete debiasing method (Yan et al., 2020). Such processes and the biases they attempt to address must be documented (Corbett-Davies et al., 2017; Corbett-Davies and Goel, 2018; Gebru et al., ItemType).

Another method to reduce AI bias is careful model selection. There have also been suggested selection methods that maximise fairness for either individuals or groups (Zafar et al., 2017; Yan et al., 2020); one such method is Kamiran and Calders' classifier-selection algorithm, which specifically seeks to achieve demographic parity by balancing positive and negative outcomes across all the different groups. Some other model-selection techniques reduce bias by use of regularisation, which penalises models for biased predictions, and others through ensemble approaches combining several models to induce a lower overall level of bias (Dwork et al.).

Another method is attempting to remedy bias after the model has already produced its output, by modifying a model's outputs for greater fairness or reduced bias. In other approaches, researchers have formulated post-processing algorithms that alter decisions so that equalised odds are satisfied where false-positive and false-negative rates are equal between groups (Yan et al., 2020). Although promising, these methods also face challenges: information preprocessing is often prohibitively expensive, and may fail to address systematic label bias if the training data are already strongly biased.

An incomplete theory of fairness in a context could also limit model-selection approaches; post-processing methods can require significant additional data and may reflect complexity (Barocas & Selbst 2016). We therefore need to keep developing new bias-mitigation methods. The task of eliminating bias in generative AI is particularly arduous (Ferrara, 2022; 2024) and involves a holistic approach that needs to start from the preprocessing stage from an equitable perspective, with consideration for diversity as well as representativeness by including data sources reflecting diverse human experience while taking into account the ratio between different demographic groups represented within training data.

Then the selection of models must favour transparent ones that can identify their biased outputs as they are produced: techniques like adversarial training, in which a model is constantly tested against difficult scenarios, help. After processing, the AI-generated content can be carefully examined and modified to address biased results using approaches such as transfer learning or additional filters that fine-tune model outputs. Ensuring that equity and fairness are adequately sustained during the life cycle of a GenAI system will require establishing feedback loops along with mechanisms for ongoing monitoring, guided by ethical principles (i.e., justified design decisions) combined with active engagement from diverse AI teams, supplemented by interdisciplinary collaboration.

These approaches must also be used thoughtfully, taking into consideration their social and ethical implications. For instance, amending the model's predictions for better fairness can involve trade-offs between different notions of bias and effects on distributions of outcomes (Ferrara 2024). Table 2 summarises the main approaches to mitigating AI bias and their associated challenges.

TABLE 2 Approaches to Mitigate AI Bias and Associated Challenges

AI Approach	Description	Examples	Challenges	Implications
Preprocessing of information	Addressing and identifying biases in data before training. Under-sampling, over-sampling, or generating synthetic data ensures that data represents the whole population.	Oversampling people of colour in a dataset for facial recognition; augmentation to represent under-represented groups; models trained with adversarial debiasing.	Time-consuming; may not be effective if the underlying data is already biased.	Risk of under- or over-representing specific groups; privacy concerns relating to data use and collection.
Model Selection	Methods that prioritise fairness. Models may be selected on the basis of individual or group fairness, or penalised for biased predictions.	Classifiers selected for demographic parity; selection methods based on group fairness; multiple models combined to reduce bias.	Constrained by an incomplete understanding of what constitutes fairness.	Fairness needs to be balanced against efficiency and accuracy.
Post-processing	Model outputs are adjusted for fairness and bias removal, for example to achieve equalised odds in decision-making.	Equalised odds achieved through post-processing approaches so that false positives and false negatives are evenly distributed across groups.	Requires substantial additional data and can become complex.	Trade-offs among different forms of bias when optimising for fairness; possible unintended effects on the distribution of outcomes.

(Source: Self, compiled from Yan et al., 2020; Kamiran and Calders, 2012; Dwork et al., 2012)

4.3.1. FAIRNESS IN AI

Fairness is a significant and hot topic in AI, attracting attention from both academic research as well as industry. Essentially, fairness in AI is the absence of discrimination or bias from an AI program a challenging goal to actualise given that many sources of bias can arise within such systems. There are many definitions of fairness, such as individual-based fairness [58], group fair outcomes and counterfactual characterisation. These two concepts have a strong symmetry bias and fairness are related, yet they are not the same while whatever is biased is perhaps fairer than it might be against some collective target population.

Fairness has become a hotly debated, multidimensional and complex concept in both industry and academia. At its most fundamental, it refers to the idea of no discrimination or bias in AI programs (Ananny and Crawford 2018); achieving this is difficult because significant attention must be paid towards all forms of bias that may arise within these systems and how they can possibly act. Various definitions of fairness have been suggested, such as counterfactual individual and group-based fairness (Zafar et al., 2017).

Whereas group fairness aims for equal treatment among groups or proportional representation of different groups in AI systems. It can be split into equal opportunity tuning, which consists of true- and false-positive rates across groups being equally distributed; demographic parity values, where positive outcomes are spread uniformly across the demographic groups; or general concepts for unfairness characterised by misclassification patterns (Yan et al., 2020; Zafar et al., 2017; Kamiran & Calders, 2012).

By contrast, individual fairness relates to treating similar individuals who belong to one or more sensitive groups and can be approached with distance- or similarity-based measures that guarantee that people similar in relevant attributes (or traits) are treated similarly by such a system (Ferguson, 2012). A newer concept, counterfactual fairness (Wachter et al., 2017), aims to guarantee that an AI will still make the same fair decision for an individual under hypothetically changed group membership. Hence, procedural fairness targets a fair and transparent decision process, while causal fairness concerns bias or unfairness within the system itself (Kleinberg et al., 2018).

These different flavours of fairness are not orthogonal and in practice tend to have a high degree of overlap; they may also conflict with one another so that trade-offs must be made when aiming for multiple dimensions of fairness (Ananny & Crawford, 2018). As a result, fairness cannot be treated as one-size-fits-all and more generally requires this nuanced, context-sensitive approach drawing on the various forms of fairness to erect balanced AI systems. Although bias and fairness are interrelated, bias is a systematic difference in the output of a model from what would be expected if there were no influencing factor on its value (Žiobaitė, 2017), whereas AI Fairness or algorithmic non-discrimination implies that an individual/ group should not receive favouritism or disfavour based upon its gender/race/skin colour /age/religion, etc. The most important

difference is that fairness can be a deliberate and conscious goal, whereas bias may not necessarily be. Thus, bias is largely a technical consideration, while fairness we argue to be an ethical and social one (Ananny and Crawford, 2018).

Bias can be either positive or negative, whereas fairness is concerned specifically with avoiding discrimination or negative bias (Žiobaitė, 2017). Negative bias occurs when a model consistently favours or discriminates against one group/individual over the other. Bias and fairness tend to be interrelated topics; mitigating bias is an important advocate/ingredient of achieving fair outcomes, i.e. for a specific prediction decision, which especially occurs when the training data contains inherently biased information (Corbett-Davies et al. 2017).

5. DISCUSSION AND CONCLUSION

The use of biased AI raises several ethical implications that merit careful consideration. Chief among these is the potential for discrimination against individuals or groups on the basis of age, race, gender, or disability: biased AI systems can preserve existing biases and promote discrimination against minority groups, a concern that is particularly acute in healthcare, where discriminatory AI programs can result in uneven access to treatment. The responsibility of companies, governments, and developers to design and use AI systems transparently and fairly is a further major concern; where an AI system produces biased outcomes, responsibility lies with both the system and those who created it, underlining the importance of regulatory frameworks and ethical guidelines that hold developers accountable for biased outcomes. The use of biased AI programs can also erode public trust, leading to reduced adoption or outright rejection of the technology, with potentially significant social and economic consequences if the benefits of AI go unrealised as a result of this lost trust. Lastly, think about how biased AI impacts human autonomy and agency: a tool with bias may infringe on individual freedom or help preserve the status quo power relationship (e.g. in hiring, their data-challenged continue to wrongly screen out candidates). This article explored the varied types of bias in AI and ML models and their potential implications for society, while also highlighting particular concerns regarding generative AI.

Where such computational tools are not adequately audited and designed, they risk amplifying existing biases related to gender, race, and other social characteristics. A number of examples of biased AI systems have been considered, with particular focus on generative AI, underlining the need for in-depth strategies to identify and mitigate bias across the full AI development pipeline. To help address bias, this study has also discussed strategies such as applying counterfactual fairness, augmenting training data, and pursuing unbiased and diverse methods of data collection. Taken together, the findings suggest that while AI carries real risks of widening the gender gap in STEM and the broader labour market, deliberate, well-designed interventions across data, model design, and governance can help ensure that AI narrows, rather than deepens, existing gender disparities in wealth generation and career progression.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest concerning the publication of this paper.

REFERENCES

[1] A. Agrawal, J. S. Gans, and A. Goldfarb, “Prediction, Judgment and Complexity: A Theory of Decision Making and Artificial Intelligence,” SSRN Electronic Journal, 2018, doi: 10.2139/ssrn.3103156.

[2] M. Ananny and K. Crawford, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,” New media & Society, vol. 20, no. 3, 2018, doi: 10.1177/1461444816676645.

[3] J. Angwin, J. A. Larson, S. Mattu, and L. Kirchner, “Machine Bias *,” pp. 254–264, Mar. 2022, doi: 10.1201/9781003278290-37.

[4] E. Arrigo, A. Di Vaio, R. Hassan, and R. Palladino, “Followership behavior and corporate social responsibility disclosure: Analysis and implications for sustainability research,” Journal of Cleaner Production, vol. 360, p. 132151, Aug. 2022, doi: 10.1016/j.jclepro.2022.132151.

[5] “Barocas, S., & Selbst, A. D. (2016). Big Data’s Disparate Impact. California Law Review, 104, 671-732. - References - Scientific Research Publishing,” Scirp.org, 2016. https://www.scirp.org/reference/referencespapers?referenceid=3614745 (accessed Aug. 27, 2026).

[6] J. Buolamwini and T. Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” proceedings.mlr.press, Jan. 21, 2018. https://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline&ref=akusion-ci-shi-dai-bizinesumedeia (accessed Aug. 27, 2026).

[7] Constantinos Challoumis - Κωνσταντίνος Χαλλουμής, “AI AND ECONOMIC INEQUALITY - ADDRESSING THE CHALLENGES AHEAD,” Dec. 31, 2024. https://www.researchgate.net/publication/387577305_AI_AND_ECONOMIC_INEQUALITY_-_ADDRESSING_THE_CHALLENGES_AHEAD (accessed Aug. 27, 2026).

[8] S. Corbett-Davies and S. Goel, “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning,” arXiv:1808.00023 [cs], Aug. 2018, Accessed: Aug. 27, 2026. [Online]. Available: https://arxiv.org/abs/1808.00023

[9] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, “Algorithmic Decision Making and the Cost of Fairness,” Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2017, doi: 10.1145/3097983.3098095.

[10] K. Crawford and R. Calo, “There is a blind spot in AI research,” Nature, vol. 538, no. 7625, pp. 311–313, Oct. 2016, doi: 10.1038/538311a.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12, 2012, doi: 10.1145/2090236.2090255.

[12] E. Ernst, R. Merola, and D. Samaan, "Economics of Artificial Intelligence: Implications for the Future of Work," IZA Journal of Labor Policy, vol. 9, no. 1, Aug. 2019, doi: 10.2478/izajolp-2019-0004.

[13] A. Ferguson, "Predictive Policing and Reasonable Suspicion," Emory Law Journal, vol. 62, no. 2, p. 259, Jan. 2012, Accessed: Aug. 27, 2026. [Online]. Available: https://scholarlycommons.law.emory.edu/elj/vol62/iss2/1/

[14] E. Ferrara, "GenAI against humanity: nefarious applications of generative artificial intelligence and large language models," Journal of Computational Social Science, vol. 7, Feb. 2024, doi: 10.1007/s42001-024-00250-1.

[15] E. Ferrara, "Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models," arXiv:2304.03738 [cs], Apr. 2023, Accessed: Aug. 27, 2026. [Online]. Available: https://arxiv.org/abs/2304.03738

[16] E. Ferrara, "The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness," Machine Learning with Applications, vol. 15, no. 15, p. 100525, Mar. 2024, doi: 10.1016/j.mlwa.2024.100525.

[17] T. Gebru et al., "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723.

[18] A. Goyal and R. Aneja, "Artificial intelligence and income inequality: Do technological changes and worker's position matter?," Journal of Public Affairs, vol. 20, no. 4, Sept. 2020, doi: 10.1002/pa.2326.

[19] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," Knowledge and Information Systems, vol. 33, no. 1, pp. 1–33, Dec. 2011, doi: 10.1007/s10115-011-0463-8.

[20] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human Decisions and Machine Predictions," The Quarterly Journal of Economics, vol. 133, no. 1, Aug. 2017, doi: 10.1093/qje/qjx032.

[21] W. M. Lim, S. Kumar, and F. Ali, "Advancing Knowledge through Literature reviews: 'what,' 'why,' and 'how to Contribute,'" The Service Industries Journal, vol. 42, no. 7–8, pp. 481–513, Mar. 2022, doi: https://doi.org/10.1080/02642069.2022.2047941.

[22] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The Ethics of algorithms: Mapping the Debate," Big Data & Society, vol. 3, no. 2, pp. 1–21, Dec. 2016, doi: 10.1177/2053951716679679.

[23] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[24] K. K. Pal, K. Piaget, Saadia Zahidi, and S. Baller, "Global Gender Gap Report 2024," World Economic Forum, June 11, 2024. https://www.weforum.org/publications/global-gender-gap-report-2024/?gad_source=1&gad_campaignid=22228224717&gbraid=0AAAAAoVy5F4x6KcDDQCh4ldE2ZECxmJGX&gclid=Cj0KCQjwnbrUBhDOARIsAKKhPpeO-fLhsfYL4prGqNsYy5MQOpDnWU5BYLWoK-tsqh3PUiIgtd7Y1s0aAm9kEALw_wcB (accessed Aug. 27, 2026).

[25] "Workmonitor | Randstad," www.randstad.com. https://www.randstad.com/workmonitor/

[26] R. Schwartz, A. Vassilev, K. K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," NIST, Mar. 2022, Accessed: Aug. 27, 2026. [Online]. Available: https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence

[27] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, pp. 59–68, Jan. 2019, doi: 10.1145/3287560.3287598.

[28] H. Snyder, "Literature review as a research methodology: An overview and guidelines," Journal of Business Research, vol. 104, no. 1, pp. 333–339, Nov. 2019, doi: https://doi.org/10.1016/j.jbusres.2019.07.039.

[29] L. Sweeney, "Discrimination in online ad delivery," Communications of the ACM, vol. 56, no. 5, pp. 44–54, May 2013, doi: 10.1145/2447976.2447990.

[30] J. Truby, "Governing Artificial Intelligence to benefit the UN Sustainable Development Goals," Sustainable Development, vol. 28, no. 4, Feb. 2020, doi: 10.1002/sd.2048.

[31] A. Turing, "Computing Machinery and Intelligence," Mind, vol. 59, no. 236, pp. 433–460, Oct. 1950, doi: https://doi.org/10.1093/mind/LIX.236.433.

[32] Z. Tzannatos, "Women and Labor Market Changes in the Global Economy: Growth Helps, Inequalities Hurt and Public Policy Matters," World Development, vol. 27, no. 3, pp. 551–569, Mar. 1999, doi: 10.1016/S0305-750X(98)00156-9.

[33] R. Vinuesa et al., "The role of artificial intelligence in achieving the Sustainable Development Goals," Nature Communications, vol. 11, no. 1, p. 233, Jan. 2020, doi: 10.1038/s41467-019-14108-y.

[34] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," SSRN Electronic Journal, vol. 31, no. 2, 2017, doi: 10.2139/ssrn.3063289.

[35] World Economic Forum, Gender parity in the intelligent age, White Paper, pp. 1-20, World Economic Forum, 2025. [Online]. Available: https://reports.weforum.org/docs/WEF_Gender_Parity_in_the_Intelligent_Age_2025.pdf

[36] S. Yan, H. Kao, and E. Ferrara, "Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes," Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Oct. 2020, doi: 10.1145/3340531.3411980.

[37] C. Yu, "Gender Inequality in the Age of AI: Predictions, Perspectives, and Policy Recommendations," Jan. 2024, doi: 10.31219/osf.io/5zrh9.

[38] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact," Proceedings of the 26th International Conference on World Wide Web - WWW '17, 2017, doi: 10.1145/3038912.3052660.

[39] I. Žliobaitė, "Measuring discrimination in algorithmic decision making," Data Mining and Knowledge Discovery, vol. 31, no. 4, pp. 1060–1089, Mar. 2017, doi: 10.1007/s10618-017-0506-1.