https://doi.org/10.25115/riem.v15i1.10970 ISSN: 2173-1950

Racial/Ethnic Bias in AI Systems: A PRISMA Systematic Review of Magnitude, Domains, and Mitigation Strategies (2014–2025)

Juan Sebastián Fernández-Prados[1], Yolanda Cara-Fernández[2], María José Torres-Haro[3].

Abstract: Objective. To examine the extent, direction, and domains in which ethnic/racial bias appears in AI/ML systems and what mitigation strategies show evidence of effectiveness.

Methods. We conducted a search in Scopus (2014–present) using the search strings bias AND "artificial intelligence" AND (racial OR ethni *), following PRISMA 2020 guidelines. After deduplication, 526 records were screened using double review (concordance and discrepancy resolution). To ensure a robust synthesis, we focused the analysis on 10 recent systematic reviews addressing racial/ethnic disparities in health, safety/justice, credit/finance, employment/selection, education, and digital platforms. We extracted results by subgroups (e.g., accuracy/error differences between groups), reported equity metrics (e.g., comparison of sensitivities and false positives between groups, overall performance differences, and "disparate impact"), and mitigation measures (before, during, and after modeling, and governance measures).

[1] Center for the Study of Migration and Intercultural Relations (CEMyRI), University of Almería, Spain, jsprados@ual.es

[2] Center for the Study of Migration and Intercultural Relations (CEMyRI), University of Almería, Spain, yolandacara@gmail.com

[3] University of Almería, Spain, mth748@ual.es

Results. All 10 reviews agree that many artificial intelligence (AI) systems reproduce or amplify racial/ethnic inequalities across various domains. In health, poorer accuracy or more errors are common for Black, Latino, and Indigenous people; in security/justice, more false positives are documented for minorities. Strategies with supporting evidence include data improvements (representativeness and quality), reweighting/ rescaling of losses, adjustment of thresholds for subgroups, and external audits with transparent results; however, heterogeneity and the measurement of race/ethnicity by inference or *proxy* limit certainty.

Conclusions. The evidence consistently points to racial/ethnic biases in AI. Mitigation measures can reduce disparities, but their effect is partial and context-dependent. We recommend reporting results by subgroups, external validation, reproducible code and data, and governance (audits and accountability). Limitation: single source (Scopus).

Keywords: algorithmic justice; racial bias; artificial intelligence; systematic reviews; PRISMA.

1. Introduction

Artificial intelligence (AI) is being rapidly incorporated into critical areas of society—from healthcare and criminal justice to recruitment and public services—with the promise of improving efficiency and objectivity in decision-making (Ferrara, 2025; Murphy, Bowen, El Naqa, Yogarajah, & Green, 2024; Kekez, Lauwaert, & Begicevic, 2025). However, concerns are growing about algorithmic biases, particularly those related to race or ethnicity, which can lead to discrimination and inadvertent inequalities (Willem, Fritzsche, Zimmermann, & Sierawska, 2024). In fact, racial bias in AI is now recognized as one of the most pressing issues in contemporary technological development, given that machine learning systems tend to reflect and even amplify historical inequalities present in their training data (Kekez et al., 2025). In other words, algorithms "inherit" social biases: if they learn from biased data, their predictions and decisions can perpetuate those same injustices (Cau, Pisu, Suri, & Saba, 2025; Omar et al., 2025). This raises serious ethical and human rights implications, especially regarding the principle of non-discrimination (Willem et al., 2024; Ferrara, 2025).

Numerous case studies have raised concerns about the real impact of these biases. For example, commercial facial recognition systems have shown higher error rates when identifying non-white faces compared to Caucasian faces, due to training data disproportionately focused on white faces. In the legal field, predictive models of criminal risk (used in bail or sentencing decisions) have been criticized for systematically scoring African American defendants as more likely to reoffend than white defendants in equivalent situations, thus reproducing racial disparities in judicial decisions (Ferrara, 2025). Even within the public sector, a scandal in the Netherlands revealed that a biased tax algorithm wrongly flagged approximately 26,000 families for welfare fraud, disproportionately affecting immigrant families and causing serious economic and emotional harm (Kekez et al., 2025; Murphy et al., 2024). These instances demonstrate that AI can cause "disparate impact": vastly different outcomes for different groups despite its purported neutrality (Kekez et al., 2025). In all examples, the consequence was to harm historically disadvantaged populations— whether ethnic or racial minorities—widening equity gaps that the technology was supposed to help close.

The risks are not merely theoretical, but tangible. In the healthcare sector, an analysis revealed that a certain commercial patient prioritization algorithm (used to allocate admission to intensive care management programs) exhibited a hidden bias in favor of white patients: given equal medical need, it recommended less care and fewer resources for Black patients, a fact only discovered when applied on a large scale (Cau et al., 2025;

Murphy et al., 2024). Similarly, studies of online advertising have shown that ads for higher-paying jobs tend to be shown more to men than to women, and that criminal background checks appear more frequently when searching for names associated with Black people (Hafner, Hafner, & Corizzo, 2025). These types of findings have generated increasing public and regulatory awareness. International bodies and regulators are taking action: the European Union, for example, through legislative proposals such as the AI Act, has emphasized the need to ensure that AI systems respect fundamental rights and do not perpetuate discrimination (Willem et al., 2024; Ferrara, 2025). In short, there is an emerging consensus that achieving fair and unbiased AI is both an ethical imperative and a prerequisite for the social acceptance of these technologies.

2. Theoretical framework

Academic interest in algorithmic fairness and bias in AI has grown exponentially in the last five years. During this time, the specialized literature has developed a common language of metrics and principles of "fairness" (algorithmic equity) to measure and compare bias. Both legal concepts (e.g., disparate impact) and statistical definitions of fairness are adopted (Ferrara, 2025). For example, the notion of "demographic parity" (related to avoiding *disparate impact)* requires that the rate of positive outcomes of a model — such as the proportion of people selected for a benefit or approved for a loan— be similar between a protected group (minority) and the majority group (Kekez et al., 2025). Another family of metrics, inspired by Hardt et al. (2016) refers to equality of opportunity or *odds,* which requires balancing error rates between groups: specifically, ensuring that the *true positive rate* (TPR) and/or false positive rate (*FPR)* of the algorithm do not differ significantly based on race or ethnicity. This is equivalent to requiring, for example, that a diagnostic model has comparable sensitivity and specificity in patients from different backgrounds, reducing any performance gaps between groups. Such metrics formally quantify disparity and have become standard for auditing AI systems. In fact, what does it mean for an algorithm to be biased? From a US legal perspective, an algorithm is biased (by disparate impact) if its selection process produces widely different results for different protected groups even when the decision rule appears neutral (Kekez et al., 2025).

In parallel, data science researchers have developed multiple technical approaches to detect and mitigate bias in AI systems. A common distinction is between preprocessing, processing, and postprocessing strategies. Preprocessing techniques act on the data before it is fed into the algorithm, for example, by resampling or reweighting the data to

ensure a balanced representation of demographic groups (there are even methods that "re - label" or modify sensitive attributes in the training data to neutralize discriminatory effects). Processing strategies are incorporated during model training, introducing additional constraints or objectives that promote fairness; for example, regularizers that penalize differences in error rates by group or algorithms that learn "unbiased" latent representations with respect to race. Finally, post-processing techniques adjust the outputs or decisions of the already trained model —for example, by calibrating risk scores by group or establishing different thresholds— to correct detected unfair deviations. Each approach has advantages and limitations, and they are often combined in practice (Sasseville et al., 2025; Cau et al., 2025).

Empirical studies have tested these bias mitigation interventions with mixed results. For example, the *equalized metric has been applied Odds* matching has been applied to clinical prediction models to recalibrate their probabilities until they achieve balanced error between white and black patients (Sasseville et al., 2025). While this can reduce disparity in certain measures, it has also been observed that forcing equal errors between groups sometimes increases the overall error rate or introduces calibration problems into the model (Sasseville et al., 2025; Ferro Desideri et al., 2025). A study by Sasseville et al. reported that approaches such as group recalibration or strict application of fairness metrics achieved partial improvements, but in some cases exacerbated prediction errors for specific subgroups or degraded the overall reliability of the algorithm. This underscores that there is no single, trivial solution: achieving fairness may involve *trade-offs* with other performance metrics and therefore finding methods to mitigate bias while minimizing the loss of accuracy is the subject of intense research. Even with these complexities, the trend is toward integrating *fairness considerations* from the very design of systems. For example, Celis et al. (2018) demonstrated how it is possible to generate fairer data summaries by incorporating demographic diversity constraints into sample selection algorithms, obtaining subsets that do not under- or over - represent a certain gender or ethnicity (Hafner et al., 2025). Similarly, “embedded” ethics approaches propose integrating social - ethical analysis throughout the technology development cycle, involving interdisciplinary teams that continuously evaluate aspects of bias, transparency, and accountability *in* AI projects (Willem et al., 2024). In short, the state of the art is progressing from a discrimination discovery phase (identifying where biases occur) to an active prevention phase, with increasingly sophisticated *fairness - aware* machine *learning* methodologies.

From an empirical standpoint, a considerable body of evidence already exists documenting racial/ethnic biases in various AI application domains. In the healthcare field, for example, multiple studies have highlighted disparities in the performance of clinical algorithms. Gichoya et al. (2022) found that a computer vision model for diagnosing pathologies in chest X-rays performed worse in certain groups: it tended to misclassify Black (and female) patients who had the disease as "healthy" compared to white male patients (Cau et al., 2025). This implied a lower true positive rate in the African American group; that is, the algorithm missed more positive cases among Black people, which constitutes an "equal" bias. "Equal opportunity" (equality in sensitivity) is unacceptable in a clinical context. Importantly, the researchers verified that this disparity was not simply due to fewer Black patients in the data; in fact, they found no correlation between ethnicity and prediction quality within the training set, suggesting that the model was exploiting subtleties in the images correlated with race. This finding aligns with other recent research that has shown the remarkable ability of neural networks to infer a patient's race from X-rays or medical images even without explicit information, leveraging patterns not yet fully understood (Ferro Desideri et al., 2025). Such results have led to questioning naive approaches such as "blinding the algorithm to race," since even by removing this attribute, the model can indirectly reconstruct it and continue to operate in a biased manner (Ferro Desideri et al., 2025).

Examples extend to many other domains. In finance, credit scoring algorithms *have* been found to discriminate based on ethnicity or postal code (a variable highly correlated with race in certain countries). In human resources, CV filtering and candidate ranking systems trained on historical data have been found to replicate racial biases in hiring, disproportionately rejecting minority applicants (Kekez et al., 2025). Even in everyday applications such as search engines and voice assistants, biases in results associated with racial stereotypes have been detected (for example, associating certain names or terms culturally linked to ethnic groups with negative content) (Hafner et al., 2025). A particularly insidious aspect is implicit bias: that which is not explicitly coded but emerges from correlations in the data. For instance, studies in the United States revealed that a seemingly neutral variable like postal code could serve as a *proxy* for race in health risk prediction models, introducing indirect discrimination against patients from predominantly African American neighborhoods (Murphy et al., 2024; Cau et al., 2025). Thus, even without discriminatory "intent" on the part of developers, algorithms can learn and exacerbate historical patterns of exclusion present in the data (Willem et al., 2024; Ferrara, 2025).

and racial biases in AI are widely documented in diverse domains, from computer vision and natural language processing to decision support systems in medicine, finance, and security; (b) typically, these biases manifest as performance or treatment gaps that disadvantage minority or traditionally marginalized groups (lower accuracy, higher error rates, or suboptimal recommendations for these groups); and (c) methodologies and metrics already exist to measure such gaps, as well as experiments with different techniques to mitigate them, although no strategy is infallible and challenges remain in achieving truly equitable AI (Cau et al., 2025; Omar et al., 2025; Ferro Desideri et al., 2025). This body of evidence lays the groundwork for a systematic review that consolidates scattered findings, compares approaches, and derives general lessons.

and ethnic bias in AI is fully justified. Beyond the social and moral relevance of the issue, there are clear scientific motivations: understanding where, why, and to what extent these biases occur is crucial for mitigating them and ensuring that AI benefits everyone equally. Several authors have advocated for expanding research at this intersection of race and artificial intelligence — an area that until recently was nascent or lacked sufficient evidence—in order to inform both technological development and public policy (Murphy et al., 2024; Ferrara, 2025). This review answers that call, providing an updated and rigorous basis on what is known about racial/ethnic algorithmic biases, their consequences, and strategies to address them (Page, McKenzie, Bossuyt, et al., 2021).

Despite the rapid growth in research on this topic, significant knowledge gaps remain. Many studies are one-off or case-by-case—for example, bias analysis of a specific algorithm or a single domain—and their results, while revealing, do not provide a complete picture on their own. How widespread and severe is racial bias in AI when we consider the entire body of evidence? We still lack a clear quantification of the typical magnitude of this bias: some systems have shown enormous differences (e.g., double the error rates for one group versus another), while in other contexts the reported disparities are more subtle. Furthermore, the variability in the direction of the bias has not been sufficiently explored: virtually all documented cases point to disadvantages for racialized groups (African Americans, Latinos, Asians, depending on the context)—which aligns with historical power structures—but it is worth asking whether there are situations where the gap is reversed or fluctuates (for example, could a certain algorithm favor the minority at the expense of the majority group in a particular scenario?). Identifying which groups are systematically disadvantaged or benefited is crucial for understanding the dynamics of the bias, and the literature has not yet synthesized this information in general (Cau et al., 2025; Omar et al., 2025; Ferro Desideri et al., 2025). In the words of

Ferrara (2025: 228): "Despite growing awareness of algorithmic biases, the field still suffers from a lack of concrete evidence that critically evaluates the functioning and outcomes of these algorithms in real-world contexts."

Another notable gap is the lack of systematization regarding the determinants of bias in different contexts. It is often assumed that the main cause is unbalanced representation in the data (*datasets* with insufficient samples of minorities, or with embedded historical biases), and this is certainly a major factor. However, recent studies suggest that it is not the only one: the way the model is trained, design decisions (which variables are included/excluded), and even the context of use can introduce additional biases. For example, even with balanced data, an algorithm could exhibit bias if it optimizes a metric that does not accurately capture fairness (such as minimizing average error without considering its distribution). Or interaction bias could occur: users' behavior toward the system (e.g., what questions they ask a virtual assistant) could differ by demographic profile, inducing biased responses. The literature lacks a consensus on the most frequent determinants of racial bias in AI—is it predominantly a data problem? A problem of poorly calibrated algorithms? A lack of testing in diverse environments? — , and this review seeks to shed light on this by compiling the available evidence (Sasseville et al., 2025; Willem et al., 2024; Murphy et al., 2024).

Related to the above, there is a gap regarding the effectiveness of mitigation strategies. Numerous techniques have been proposed, as we have seen, but how much empirical evidence is there that they actually work? Many studies present novel fairness approaches *with* evaluations on laboratory *datasets;* however, fewer studies demonstrate successful bias mitigation in applied or industrial-scale scenarios. In fact, Sasseville et al. (2025) emphasize that "there remains a knowledge gap regarding which strategies and methods have been empirically applied to mitigate bias toward diverse groups" in real-world contexts. On the regulatory side, there is also a lack of connection between technical research and regulatory frameworks and concrete policies: for example, how to incorporate fairness criteria into certifications of medical algorithms or external audits. In short, it has not yet been synthesized which techniques have shown solid evidence of reducing racial bias in AI and under what conditions—key information for guiding both developers and policymakers (Sasseville et al., 2025).

Finally, although the volume of publications has increased, the literature remains fragmented across disciplines. Each field (medicine, finance, computer vision, etc.) tends to study biases within its own systems using its own specific metrics, and does not always engage with findings from other domains. This makes it difficult to extract general

principles or transferable solutions. Therefore, there is a need for an integrative review that identifies cross-cutting patterns: for example, determining whether biases favor the majority group in all domains, or whether certain types of models (e.g., deep neural networks) are more prone to bias than simpler models, or whether certain interventions (e.g., adding more minority data) tend to be effective regardless of the context (Kekez et al., 2025; Lastrucci, Iosca, Wandael, et al., 2025).

Based on the available evidence and the Scopus bibliometric map, this systematic review seeks to provide a clear and usable synthesis on ethnic/racial bias in AI/ML systems. Specifically, we aim to: (i) estimate the magnitude of documented disparities between racial/ethnic groups using comparable metrics; (ii) describe their direction, identifying which groups are systematically disadvantaged and under what conditions; (iii) map the application domains in which bias is reported (health, safety, credit, employment, education, platforms, and others), contextualizing where the problem has been most studied and with what results; and (iv) evaluate the effectiveness of reported mitigation strategies (pre-, in-, and post - process, as well as governance measures and data), with particular attention to their performance commitments and external validity.

These objectives align with the results that will be presented in three tables: Table 1, which situates the complete corpus of 526 Scopus documents by descriptors by year, thematic areas and countries; and the Tables Sections 2 and 3 delve into 10 systematic reviews published in 2025, selected for comparative qualitative analysis (from a previous subset of 25 reviews and 171 records from 2025-26). With this, we aim to offer a dual perspective: panoramic (the “state of the field”) and high-quality (the best and most recent), useful for research, practice, and regulation.

3. Method

3.1. Methodology and sample

The review was conducted following the PRISMA 2020 guidelines (Page et al., 2021), incorporating its 27-item checklist, abstract guidelines, and flow diagram, along with the PRISMA-S recommendations, which ensure the traceability and transparency of the search process. From the protocol design stage, the domains of analysis, metrics of interest, eligibility criteria, and synthesis plan were established, with particular attention to the equity approach—that is, how racial or ethnic variables are measured and used, the representativeness of the groups, and the algorithmic justice metrics applied in each study or review.

The literature search was conducted in a single source— Scopus (Elsevier)—using the institutional interface and limiting the query to the TITLE-ABS-KEY fields. The search string used was bias AND "artificial intelligence" AND (racial OR ethni*). The use of the asterisk after ethni* is a common practice in academic searches: it allows retrieval of all variants derived from the same lexeme—for example, *ethnic, ethnicity,* etc.—thus ensuring greater sensitivity and avoiding the omission of relevant works that use different formulations of the term. This truncation strategy, combined with the logical OR operator, broadened the coverage without sacrificing thematic precision.

The initial time frame spanned from 2014 to the search cutoff date (October 13, 2025). However, a review of the results revealed that no publications prior to 2014 met the defined criteria. This can be explained by the fact that discussions about racial or ethnic bias in artificial intelligence are relatively recent: until the mid-2010s, the literature focused on technical aspects of algorithm performance and efficiency, and only from 2014 onward did studies addressing fairness and representation in machine learning systems begin to proliferate.

The search yielded a total of 526 publications, which constitute the core corpus of the review. Scientific output shows a clear upward trend throughout the analyzed period: between 2014 and 2018, only a few isolated works were recorded, reflecting a still incipient interest in the topic. From 2019 onward, sustained growth is observed, with a notable increase in 2021 and 2022, coinciding with the expansion of the debate on algorithmic fairness in areas such as health, security, and public administration. The increase intensifies in the last two years, reaching its peak in 2025, which accounts for 170 documents, almost a third of the total. This trend confirms that the issue of ethnic and racial bias in artificial intelligence has gone from being a marginal matter to becoming a priority line of interdisciplinary research in recent literature.

3.2. Reproducibility and transparency

The protocol for this review—which includes the PICOS criteria, the PRISMA-S strategy, the double screening plan, data extraction and synthesis—along with the supplementary materials (extraction matrix, bias risk templates and code to standardize equity metrics), is available in open access on Mendeley Data (Fernández-Prados, Cara-Fernández and Torres-Haro, 2025): https://doi.org/10.17632/pjfmsy3d9s.1

3.3. Eligibility criteria

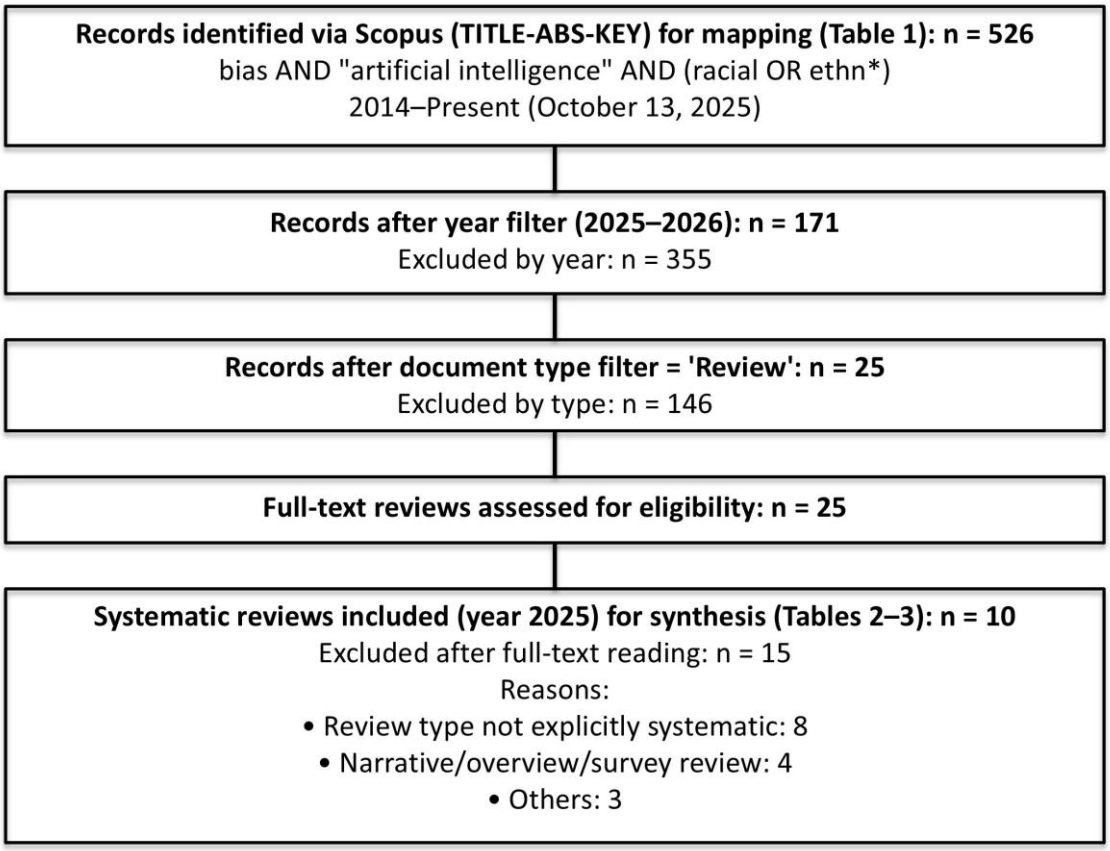
The study selection process followed three successive and complementary filters that allowed for narrowing the final corpus and ensuring its relevance to the review's objectives. First, a temporal criterion was applied, focusing on the years 2025 and 2026, to capture the most recent research on ethnic and racial bias in artificial intelligence. This decision is justified by the exponential growth of publications in this field in recent years and by the emergence, during this period, of audits, reviews, and more sophisticated algorithmic fairness metrics. The filter was applied directly in the Scopus interface, using the year range search function, which reduced the initial universe of 526 records to 171 publications.

The second filter was based on document type, using the Scopus filter to isolate publications classified as reviews. This stage addressed the need to identify systematic or scoping reviews that offered syntheses of evidence and methodological frameworks applied to different AI domains. The result was a set of 25 reviews covering the period 2025–26. However, it was recognized that not all reviews labeled by Scopus met the methodological standards of a systematic review, so a third step of manual cleaning was necessary.

The third criterion involved a manual, expert appraisal of the identified reviews to distinguish those that met the requirements of a systematic review according to the 2020 PRISMA guidelines. This appraisal considered the presence of an explicit protocol, a description of the search strategy, inclusion and exclusion criteria, double peer review, risk of bias analysis, and a structured synthesis of results. Only 10 systematic reviews published in 2025 met these standards and constitute the core of the qualitative synthesis presented in Tables 2 and 3.

The PRISMA flowchart reflects the complete journey: from the initial 526 records to the 10 final revisions included, detailing the intermediate steps, exclusions by year and document type, as well as the reasons for exclusion from full text.

Figure 1. **PRISMA 2020 flowchart for study identification and selection**



Source: Own elaboration.

4. Results

4.1. Bibliometric mapping and trends

The map in Figure 2 confirms a clear shift in focus towards applied fields, with medicine gaining ground year after year. While contributions from medical sciences were minimal between 2014 and 2018, a sustained increase is observed from 2021 onwards, culminating in 2025 when medicine accounts for more than a quarter of the output. This shift is not accidental: it coincides with the introduction of clinical models into practice and with the first large-scale audits documenting racial/ethnic gaps in performance and access.

In parallel, computer science is losing relative centrality. It's not that less research is being done, but rather that its prominence is being "redistributed" as other areas take ownership of the problem. In 2014–2018, almost half of all research projects came from computer science; by 2025, that proportion drops to around 18%. %. This change suggests that the discussion about racial bias is no longer a “technical” matter and is

becoming a public issue, where institutional contexts, regulations, and the effects on people's lives matter.

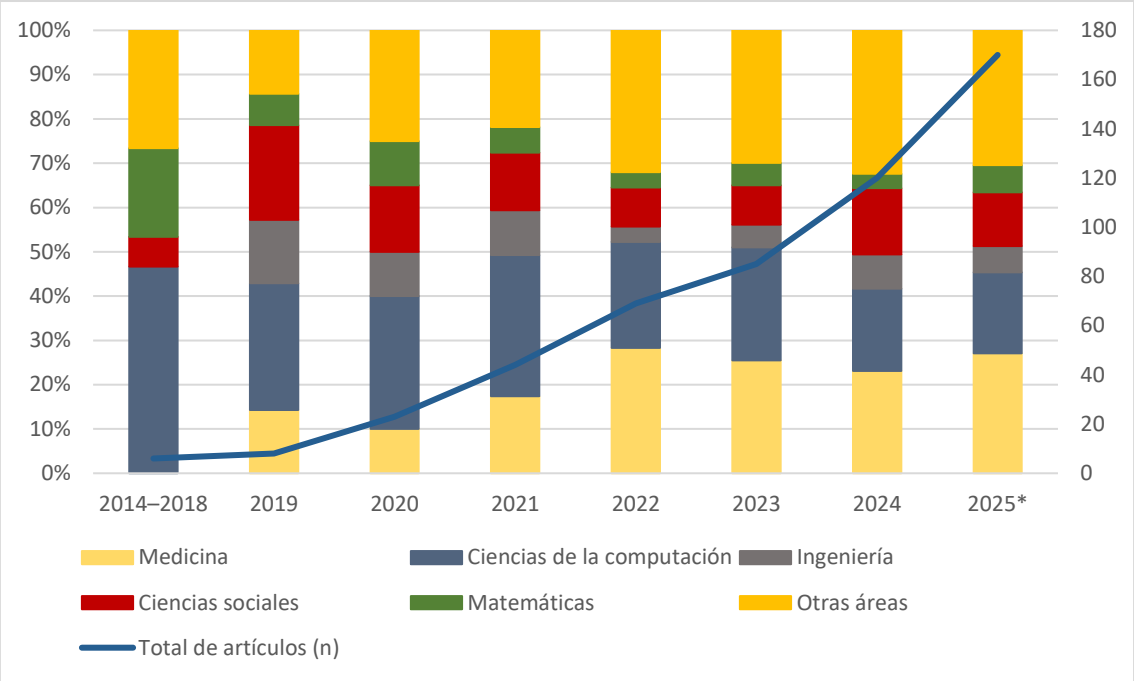
The social sciences maintain a constant presence (around 11 % of the total), although in the last two years it has been exceeding the average. Its contribution is key to framing the bias not only as a statistical error, but as a structural inequality: how race /ethnicity is measured, what categories are used, and what the implications are of deciding on metric thresholds. Engineering and mathematics contribute less volume, but function as methodological "bridges" —validation, modeling, calibration — those cross domains. Finally, the "other areas" category remains high (≈30 % in recent years) and brings together ethics, law, public health and the humanities: a reminder that algorithmic fairness requires both technical solutions and governance, accountability and regulatory frameworks.

***Table 1*. Annual evolution and relative weight by disciplinary area in the production on racial/ethnic bias in AI (Scopus, 2014–2025)**

Area / Year	2014–2018	2019	2020	2021	2022	2023	2024	2025*	Total
Total (n)	6	8	23	44	69	85	120	171	526
Medicine	0.0 %	14.3 %	10.0 %	17.4 %	28.3 %	25.5 %	23.1 %	27.1 %	24.9%
Computer science	46.7 %	28.6 %	30.0 %	31.9 %	23.9 %	25.5 %	18.6 %	18.3 %	22.5%
Engineering	0.0 %	14.3 %	10.0 %	10.1 %	3.5 %	5.1 %	7.7 %	5.9 %	6.8%
Social sciences	6.7 %	21.4 %	15.0 %	13.0 %	8.8 %	8.9 %	15.0 %	12.1 %	11.1%
Math	20.0 %	7.1 %	10.0 %	5.8 %	3.5 %	5.1 %	3.2 %	6.2 %	2.3%
Other areas	26.6 %	14.3 %	25.0 %	21.8 %	32.0 %	29.9 %	32.4 %	30.4 %	27.4%

Source: Prepared by the author using data from Scopus. Note: 2025 indicates the current year, and two publications dated 2026 have been added.

***Figure 2*. Temporal distribution by areas of knowledge in studies on racial/ethnic bias in AI (Scopus, 2014–2025)**



Source: Prepared by the author using data from Scopus. Percentages relative to the annual total. Aggregated areas: Health, Computer Science, Engineering, Social Sciences, Mathematics, and Other.

4.2. Characteristics and analysis of the included reviews

The ten systematic reviews published in 2025 highlight two major areas of concern. The first lies in generative technologies and large language models. A review of text - to - image (TTI) models shows that when we request " neutral " images, professions are predominantly depicted with white features, and dark skin tones are underrepresented. The fundamental question is twofold: how to identify and measure this bias, and what fixes actually work (e.g., rewriting the *prompt,* retraining, or adjusting the model). In parallel, another review examines large language models (LLMs) used clinically and documents that responses or recommendations can change depending on the target demographic (race/ethnicity, gender, or age), a particularly sensitive issue in healthcare settings.

The second focus is on health, with four complementary perspectives. One review proposes that, to make clinical trials more diverse, “digital twins” and equity-conscious recruitment rules should be incorporated. Three other reviews delve into the specifics: in dermatology, the literature on “skin age” warns that dark skin tones are missing from image sets, which could translate into worse outcomes for these individuals; in thyroid cancer, studies with large samples abound but lack complete demographic data, making it impossible to assess whether AI works equally well for everyone; and in

pharmacogenomics of depression, the overall sample is highly unbalanced (too many white and Asian participants, too few Black, Hispanic/Latino, and Native American participants), with the predictable consequence: if this data feeds AI, precision medicine could be less accurate for underrepresented groups.

The set is completed with three panoramas that "broaden focus": one ethical - organizational on large-scale language models in health (audits, transparency and governance); another on science /chemistry education, which reminds us that stereotypes also creep into the classroom; and a map of the use of myographic signals with AI to assess motor disorders, which reveals a basic deficit: performance is almost never reported by ethnicity, so we cannot know if the model treats everyone equally.

Three features are repeated in these reviews: (i) very different ways of "measuring race/ethnicity" (self-reporting, inference, such as country), (ii) little standardization of equity metrics (which hinders comparisons across domains), and (iii) a clear geographic bias toward countries in the Global North and publications in English. In short, we have a vibrant field, with well-formulated questions, but still with inconsistent measurement rules and gaps in sample descriptions. Very clear shortcomings are documented in specific areas: for example, "skin age" studies report underrepresentation of Black people, which reduces the generalizability of the models and can compromise their clinical utility, especially in dark skin.

***Table 2*. Key characteristics of systematic reviews (scope, sources, period, and inclusion criteria)**

Citation	Research question	Sources and period	Included studies	Inclusion criteria
Elsharif et al., 2025	How to detect and mitigate cultural bias in text - to - image.	Scopus, WoS, arXiv, Google Scholar (2000 – 2024).	58	Papers on TTI with bias assessment /mitigation; peers; English.
Omar et al., 2025	What demographic biases do medical LLMs present and how they are measured/mitigated.	PubMed, Embase, WoS, PsycInfo, Scopus (2018 – 2024).	24	Paired studies evaluating biases (gender, race/ethnicity, age) in clinical LLMs.
Bull et al., 2025	Development, performance and equity of models in child protection.	Multiple bases (2020 – 2024).	11	Predictive ML with future results; full text; validations.
Tubbs & Álvarez-Vázquez, 2025	How digital twins and AI can improve diversity in trials.	PubMed, IEEE, Scopus and others (≈ 2015 – 2025).	90	Studies with DT/IA that address

				diversity in trials; peers.
McMullen et al., 2025	ML precision for "skin age" and its limits.	Embase, Medline, IEEE, ACM (until 2024).	27	Multivariable ML in humans; pairs; English.
Fareed et al., 2025	Ethics of LLMs in health: biases, privacy, transparency, governance.	ACM, Springer, Wiley, PubMed (2017 – 2025).	27	Peers, full text access; focus on LLMs and ethics.
Jackson et al., 2025	Racial / ethnic representation in depression pharmacogenomics.	Multiple (detailed in methods).	390	PGx studies * with extractable race/ethnicity data.
Erümit & Özdemir, 2025	Use and effects of AI in science/chemistry education (incl. GenAI).	WoS and Scopus (2014–2024).	18	Primary research in English that evaluates the effectiveness of educational AI.
Ramchandani et al., 2025	Overview of AI in thyroid cancer and representativeness.	EMBASE, PubMed, Google Scholar (until 2024).	197	AI models with reported demographic data.
Son and al., 2025	AI and myographic signals for motor disorders.	Scopus and PubMed (2014 – 2024).	111	Humans, English; AI studies and myography with reported results.

Source: Prepared by the authors based on the included reviews. * PGx, Pharmacogenomics

The findings converge on a clear message: disparities exist and, when reported, they often disproportionately affect historically racialized groups. In generative visuals, persistent stereotypical associations and "whitewashing" are observed; adjusting *prompts* or fine-tuning models helps, but it involves quality exchanges and does not eliminate the root problem. In large clinical language models, most studies detect differences in responses or recommendations based on race/ethnicity; corrections through *prompting* or *fine - tuning are not sufficient*. Cultural studies show disparate results and demand validation outside of Western contexts.

In child protection, group equity is rarely assessed and almost never with standardized metrics, making it impossible to judge whether models treat racialized families equally. In clinical trials, ethnic underrepresentation is recognized as a structural problem, and end-to-end equity audits of the trial cycle are proposed. In fields such as dermatology or thyroid medicine, the conclusion is straightforward: when dark skin or minority groups are missing from the data, differential performance becomes apparent or cannot be assessed; the practical consequence is a greater risk of error for those who are underrepresented in training sets.

Other reviews in other fields provide indirect but relevant evidence: ethical maps of large language models that emphasize auditing by subgroups, and pharmacogenomics reviews that reveal significant racial/ethnic participation asymmetries which, if transferred to AI tools, would bias precision medicine. Finally, some fields—such as myography—acknowledge that they do not yet measure equity, which is in itself a finding: without measurement, there is no possible correction.

***Table 3*. Signals of racial/ethnic bias and equity metrics by domain; mitigation strategies and observed effects**

Appointment	General findings	Specific signs of racial/ethnic bias
Elsharif et al., 2025	Lack of standards and incomplete mitigations in TTI; Western bias in *benchmarks* or test benches.	Underrepresentation of dark skin and "whitening" *of* professions; partial mitigation with *prompting* / fine - *tuning*.
Omar et al., 2025	Medical LLMs frequently exhibit demographic biases; irregular mitigation.	10/11 studies with racial bias; differences in clinical recommendations for underrepresented groups.
Bull et al., 2025	Heterogeneity and low transparency in child protection; few equity assessments.	Only 4/11 measure equity by race/ethnicity; lack of comparable metrics (TPR/FPR, *equalized odds).*
Tubbs & Álvarez-Vázquez, 2025	DT/IA can improve equitable recruitment if metrics and governance are integrated.	Proposes equity and demographic parity audits / *equalized odds* in the trial cycle.
McMullen et al., 2025	Acceptable performance in "skin age"; risk of bias due to small samples.	Underrepresentation of dark skin in classic datasets; potential impact on diagnosis.
Fareed et al., 2025	Ethical agenda for LLMs: audits, traceability and clinical supervision.	Demographic biases are identified as a priority, but many studies lack quantification by race/ethnicity.
Jackson et al., 2025	Strong imbalance in PGx (whites/ Asians over - represented).	Risk of generalization bias in AI for precision medicine if the representation is not corrected.
Erümit & Özdemir, 2025	Educational AI is growing; benefits and risks (hallucinations, reliability, biases).	Evidence of racial bias in images and content; calls for literacy and an ethical approach.
Ramchandani et al., 2025	Partial alignment with epidemiology; demographic and socioeconomic gaps.	Discrepancies between global prevalences and samples by ethnicity; difficult to assess TPR/FPR due to lack of variables.
Son and al., 2025	and classification tasks predominate; improvements with signal fusion.	It does not report metrics by race/ethnicity; absence of equity assessment and recruitment bias.

Source: Authors' own elaboration. * EMG, *Electromyography*

5. Discussion

Our corpus reveals a pattern that is difficult to ignore: when empirical evidence exists, racial/ethnic disparities consistently emerge in generative and predictive systems, although with variations depending on the domain and the quality of the available data. In text-to-image models, the systematic review by Elsharif and colleagues (58 studies) confirms a persistent bias that "whitens" occupations and underrepresents dark skin tones; furthermore, it notes that there is no fully effective mitigation strategy and that benchmarks tend to be "Westernized" (the bias re-emerges with new versions of the model) (Elsharif et al., 2025). In the field of clinical LLMs, Omar et al. synthesized 24 studies and found biases in 22 of them; Nine out of ten studies that examined race/ethnicity reported disparities, with practical effects on diagnostic or therapeutic recommendations for subgroups, and mitigation, when it exists, relies primarily on *prompting* with inconsistent results (Omar et al., 2025). Literature outside the healthcare field paints a similar picture: in software engineering recruitment scenarios, LLMs prefer male and Caucasian profiles and generate images with stereotypical body types and age; that is, bias seeps into both the text and the visuals (Bano, Gunatilake, & Hoda, 2025). Even in narrative experiments with ChatGPT-4 in Spain, an overrepresentation of young, heterosexual, and Hispanic characters is observed, with ethnic minorities or those from low socioeconomic backgrounds relegated to the margins of the narrative (Gabino-Campos, Baile, & Padilla-Martínez, 2025). The educational landscape reinforces this same message: the systematic review in science and chemistry education (2014–2024) documents pedagogical benefits, but also recurring risks—demographic biases, hallucinations, and reliability issues—that call for AI literacy and deliberate ethical use in the classroom (Erümit & Özdemir Sarıalioğlu, 2025). Furthermore, the findings of this review expand upon previous work focused on gender biases (Fernández-Prados & Lozano Díaz, 2025), demonstrating that algorithmic inequalities cut across multiple dimensions of identity.

However, the root of the problem rarely lies solely in "the algorithm." Several 2025 reviews focus on the data pipeline: how, when, and by whom measurements are taken. In pharmacogenomics of depression, Jackson et al. quantified 390 studies and showed an overall composition of 57.3% white participants, 36.4% Asian, 1.7% Black, and 3.5% Hispanic/Latino; only 16.2% of the studies included Black or Hispanic/Latino patients. In other words, the promise of AI/ML applied to PGx is built on samples that do not reflect the populations they are intended to serve, with the consequent risk of bias in future precision medicine tools (Jackson et al., 2025). In thyroid cancer, the systematic review by Ramchandani et al. (197 studies; 248,896 people) identifies similar

mismatches and the frequent absence of race/ethnicity variables in the modeling, which prevents the assessment of equity metrics and reveals an underlying representation bias (Ramchandani et al., 2025). Other areas repeat the pattern: in skin age estimation using ML, the lack of ethnic diversity and small sample sizes increase the risk of bias and compromise generalizability (McMullen et al., 2025); and in the assessment of motor impairments with myographic signals, the mapping by Sohn et al. shows that many studies do not even report metrics stratified by ethnicity, with samples dominated by healthy males and no equity analysis, thus remaining blind to potential performance differences by subgroups (Sohn, Sohn, & Son, 2025).

A third common thread is methodological: we are evaluating late, poorly, or in ways that are difficult to compare. In child protection, for example, only 4 of 11 reviewed models included per-group metrics (i.e., compared performance across subpopulations) using race/ethnicity as the protected attribute. Furthermore, transparency and reproducibility are low: the TRIPOD+AI guidelines (for reporting clinical prediction models clearly and completely) and the PROBAST tool (for judging the risk of bias in prediction models) often indicate uncertain or high risk. In addition, the discipline still too frequently resorts to "blindness by design" (failing to consider sensitive variables) instead of reporting equity metrics such as TPR/FPR parity—TPR: true positive rate or sensitivity; FPR: false positive rate—or equalized Odds /equality of odds (requiring similar TPR and FPR across groups) (Bull et al., 2025). In text-to-image, *Elsharif* et al. emphasize that there are no common standards and that there is reliance on biased benchmarks and prompts, *which* prevents reliable and replicable comparisons of different mitigation strategies (Elsharif et al., 2025). Finally, there is an almost ignored problem: the temporality of demographic data. Gosciak and colleagues show, using clinical records of more than 5 million patients, that delays in recording race/ethnicity distort disparity audits—especially at the visit and state levels—and that routine imputations (estimating race/ethnicity using indirect methods) do not correct the bias in time to operationally address inequities (Gosciak et al., 2025).

What has worked—at least according to empirical evidence—to mitigate this? The literature offers two types of signals. The first, of a technical nature, points to data strategies: Marchesi et al. show that conditional synthetic generation (CA-GAN) can alleviate representational bias and improve the fairness of clinical models in Black subpopulations and women, maintaining the original distribution and improving performance on downstream predictive tasks; this is proof of concept that data diversification, if done rigorously, can have measurable effects on *fairness* without sacrificing utility (Marchesi et al., 2025). The second of an organizational nature, situates

fairness throughout the life cycle: Tubbs and Álvarez-Vázquez propose integrating group audits (demographic parity, *equalized odds)* and inclusive data practices in trial design and operation, using digital twins and AI to simulate and correct recruitment biases from the outset (Tubbs & Álvarez-Vázquez, 2025). In contrast, interventions focused on *prompt* Command engineering offers variable results in medicine, and evidence suggests it is insufficient without changes in data, external validation, and deployment governance (Omar et al., 2025). A broader ethical synthesis agrees that technical safeguards (audits, federated learning, *debiasing)* must be accompanied by organizational mechanisms—traceability, clinical oversight, and *network- teaming*. (adversarial testing)—to avoid harm in real-world contexts (Fareed, Fatima, Uddin, Ahmed, & Sattar, 2025). Finally, even when models and versions are refined, biases persist or mutate: recent comparisons between large language model families (LLMs) show that *instruction-tuning* reduces overtly biased expressions, but can introduce censorship or confusion and leaves subtle biases intact (e.g., in disability), with relevant differences between versions and providers; this calls for robust and transparent fairness testing before sensitive adoptions (Gupta, Marrone, Gargiulo, Jaiswal, & Marassi, 2025).

6. Conclusions

Racial or ethnic bias is not an isolated problem. It is observed in: (1) systems that generate images from text descriptions; (2) large language models used in healthcare; (3) models for diagnosing or predicting outcomes; and (4) assessments in education and employment. In all these areas, disadvantages for historically racialized groups are repeated, and biases tend to reappear with each new technological generation (Elsharif et al., 2025; Omar et al., 2025; Bano et al., 2025; Gabino-Campos et al., 2025; Erümit & Özdemir Sarıalioğlu, 2025).

Where do these biases come from? Not so much from an inherent flaw in the algorithm, but from the entire chain that feeds it: limited data diversity and inadequate measurement of race/ethnicity; delays and gaps that prevent timely audits; and a lack of common standards for comparing equity metrics across different fields (Jackson et al., 2025; Ramchandani et al., 2025; McMullen et al., 2025; Sohn et al., 2025; Gosciak et al., 2025; Bull et al., 2025). There are promising, though still incomplete, advances: diversifying the data (and, where appropriate, generating it synthetically with distribution control) improves the equity we can then measure; conducting audits that consider the entire data chain and incorporating group-specific metrics from the study design stage helps prevent bias from creeping into the model. and ethical governance

provides the necessary support for technical solutions to work in practice (Marchesi et al., 2025; Tubbs & Álvarez-Vázquez, 2025; Fareed et al., 2025).

The task ahead is twofold: on the one hand, empirically we need studies that: (a) measure race/ethnicity with clear criteria (ideally, self-reporting) and record when this information is collected so that audits are timely; (b) report performance and calibration by subgroups, including at least differences in sensitivity and false positive rate between groups (TPR/FPR) and the disparate impact ratio (ratio between the rate of the protected group and that of the group with the highest rate); (c) explain the adjustments by subgroup when thresholds or other model parameters are changed; and (d) publish code and data or, if this is not possible, guarantee reproducibility with guidelines such as TRIPOD+IA, PROBAST/PROBAST - IA and PRISMA in reviews (Bull et al., 2025; McMullen et al., 2025).

On the other hand, in practical terms we recommend: (1) continuous equity audits with updated and time-traceable demographic data. Annual audits based on a “snapshot” can be off by 10 points or more at the center or practice level; disparity panels must account for registration delays and avoid imputations that mask the real problem (Gosciak et al., 2025); (2) data strategies that combine informed oversampling, inclusive curation, and, where appropriate, conditional synthesis for minorities, explicitly assessing the equity achieved in real clinical or operational tasks (Marchesi et al., 2025); (3) external validation in non-Western contexts and in underrepresented populations (a priority highlighted by reviews in pharmacogenomics and thyroid). Without it, AI-supported precision medicine risks being “precision for the few” (Jackson et al., 2025; Ramchandani et al., 2025); (4) governance and transparency: end-to-end equity plans in trials and deployments (from recruitment to evaluation), traceability of decisions and publication of model fact sheets with metrics by subgroups, as recommended by recent ethical syntheses (Fareed et al., 2025); and (5) literacy and training in educational and health centers to understand limits, biases and the responsible use of generative and predictive tools (Erümit & Özdemir Sarıalioğlu, 2025).

The biases we observe are not mere statistical anomalies: they are the imprint of historical inequalities that machines learn to reproduce if we don't intervene. Precisely for this reason, the solution cannot be solely technical. We need the social sciences and humanities to problematize how we define "race/ethnicity," what our *proxies imply,* and what harm we cause when we infer identities; we need ethics and law to frame responsibilities; and we need engineering, medicine, and public health to translate that framework into verifiable data, models, and procedures. Where this interdependence has

been tested—for example, by integrating equity audits into study design or by reinforcing representation with controlled synthetic data—tangible improvements appear. Making them the norm, not the exception, is the challenge of the next phase of the AI era: a phase in which algorithmic justice ceases to be a rhetorical flourish and becomes a daily, measurable, and shared practice.

References

Bano, M., Gunatilake, H., & Hoda, R. (2025). What does a software engineer look like? Exploring societal stereotypes in LLMs. In *Proceedings of the IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS '25)* (pp. 173–184). IEEE. https://doi.org/10.1109/ICSE-SEIS66351.2025.00023

Bull, C., Kisely, S., Betts, K., & Hu, Y. (2025). Understanding the development, performance, fairness, and transparency of machine learning models used in child protection prediction: A systematic review. *Child Abuse & Neglect,* 169, 107630. https://doi.org/10.1016/j.chiabu.2025.107630

Cau, R., Pisu, F., Suri, J. S., & Saba, L. (2025). Addressing hidden risks: Systematic review of artificial intelligence biases across racial and ethnic groups in cardiovascular diseases. *European Journal of Radiology, 183,* 111867. https://doi.org/10.1016/j.ejrad.2024.111867

Elsharif, W., Alzubaidi, M., & Agus, M. (2025). Cultural Bias in Text - to - Image Models: A Systematic Review of Bias Identification, Evaluation, and Mitigation Strategies. *IEEE Access, 13,* 122636–122659. https://doi.org/10.1109/ACCESS.2025.3585745

Erümit, A. K., & Özdemir Sarıalioğlu, R. (2025). Artificial intelligence in science and chemistry education: A systematic review. *Discover Education, 4,* 178. https://doi.org/10.1007/s44217-025-00622-3

Fareed, M., Fatima, M., Uddin, J., Ahmed, A., & Sattar, M. A. (2025). A systematic review of ethical considerations of large language models in healthcare and medicine. *Frontiers in Digital Health, 7,* 1653631. https://doi.org/10.3389/fdgth.2025.1653631

Ferrara, E. (2025). Addressing Racial Bias in AI: Towards a More Equitable Future. In: Czarnowski, I., Howlett, RJ, C. Jain, L. (eds) *Intelligent Decision Technologies*.

KESIDT 2024. Smart Innovation, Systems and Technologies (pp. 227–235). Springer, Singapore. https://doi.org/10.1007/978-981-97-7419-7_20

Fernández-Prados, J. S., Cara-Fernández, Y., & Torres-Haro, M. J. (2025). *Racial/Ethnic Bias in AI Systems: PRISMA Systematic Review of Magnitude, Domains, and Mitigation Strategies (2014–2025)* [Data set]. Mendeley Data, V1. https://doi.org/10.17632/pjfmsy3d9s.1

Fernández-Prados, J. S., & Lozano Díaz, A. (2025). *Analysis of ethical challenges and gender biases in generative artificial intelligence*. In JD Barquero Cabrero, E. Ruiz Callejón, Á. Ramos Ruiz and E. López Carrillo (Eds.), *Media and society in transformation: Challenges and narratives of power* (pp. 152–163). Madrid: ESIC Editorial. ISBN 978-84-1192-162-6.

Ferro Desideri, L., Kirkpatrick, B., Zinkernagel, M., & Anguita, R. (2025). Bias in predictive models for vitreoretinal diseases: Ethnic and socioeconomic disparities in artificial intelligence. *Eye*. https://doi.org/10.1038/s41433-025-03990-0

Gabino-Campos, M., Baile, J. I., & Padilla-Martínez, A. (2025). Social biases in AI-generated creative texts: A mixed-methods approach in the Spanish context. *Social Sciences, 14* (3), 170. https://doi.org/10.3390/socsci14030170

Gosciak, J., Balagopalan, A., Ouyang, D., Koenecke, A., Ghassemi, M., & Ho, DE (2025). Bias delayed is bias denied? Assessing the effect of reporting delays on disparity assessments. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)* (pp. 1843–1861). https://doi.org/10.1145/3715275.3732123

Gupta, O., Marrone, S., Gargiulo, F., Jaiswal, R., & Marassi, L. (2025). Understanding social biases in large language models. *AI, 6* (5), 106. https://doi.org/10.3390/ai6050106

Hafner, F. S., Hafner, L., & Corizzo, R. (2025). "Slightly disappointing" vs. "worst sh ** ever": Tackling cultural differences in negative sentiment expressions in AI -based sentiment analysis. *Journal of Computational Social Science, 8,* 57. https://doi.org/10.1007/s42001-025-00382-y

Jackson, L., Delaney, K., Bobo, J., Grant, C. W., Hassett, L., Wang, L., Weinshilboum, R., Croarkin, P. E., Moyer, A., Gentry, M. T., & Athreya, A. P. (2025). Quantifying Sample Representation in Global Pharmacogenomic Studies of Major Depressive Disorder: A Systematic Review. *Clinical and Translational Science,* 18, e70256. https://doi.org/10.1111/cts.70256

Kekez, S., Lauwaert, M., & Begicevic, N. (2025). Is artificial intelligence (AI) research biased and conceptually vague? A systematic review from a sustainable and multi -level perspective. *Technology in Society, 81,* 102818. https://doi.org/10.1016/j.techsoc.2025.102818

Lastrucci, A., Iosca, N., Wandael, Y., Barra, A., Ricci, R., Nori Cucchiari, J., Forini, N., Lepri, G., & Giansanti, D. (2025). Transforming breast imaging: A narrative review of systematic evidence on artificial intelligence in mammographic practice. *Diagnostics, 15* (17), 2197. https://doi.org/10.3390/diagnostics15172197

Marchesi, R., Micheletti, N., Kuo, NNI-H., Barbieri, S., Jurman, G., & Osmani, V. (2025). Generative AI mitigates representation bias and improves model fairness through synthetic health data. PLOS Computational Biology, 21(5), e1013080. https://doi.org/10.1371/journal.pcbi.1013080

McMullen, E., Aflaki, R., Khatri, P. J., Metko, D., Storm, K., ... Champagne, T. (2025). Machine learning methods for determining skin age: A systematic review. Journal of Tissue Viability, 34, 100887. https://doi.org/10.1016/j.jtv.2025.100887

Murphy, A., Bowen, K., El Naqa, I. M., Yogarajah, B., & Green, B. L. (2024). Bridging health disparities in the data -driven world of artificial intelligence: A narrative review. *Journal of Racial and Ethnic Health Disparities, 12,* 2367–2379. https://doi.org/10.1007/s40615-024-02057-2

Omar, M., Sorin, V., Agbareia, R., Apakama, DU, Soroush, A., Sakhuja, A., ... Nadkarni, G.N., & Klang, E. (2025). Evaluating and addressing demographic disparities in medical large language models: a systematic review. International *Journal for Equity in Health, 24:*57. https://doi.org/10.1186/s12939 - 025 - 02419 - 0

Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al. (2021). PRISMA Statement 2020: Updated guidelines for publishing systematic reviews (*Spanish translation). PRISMA 2020 Document*. https://doi.org/10.1016/j.recesp.2021.06.016

Ramchandani, R., Guo, E., Biglou, S. G., Sabbah, S.G., Mostowy, M., Mahiny, D., Hurtubise, C., Anicho - Okereke, G., Shorr, R., Caulley, L., Propst, E. J., Wolter, N. E., Wasserman, J. D., Eskander, A., & Siu, J. M. (2025). Representation and Bias in Artificial Intelligence Models for Thyroid Cancer: A Systematic Review. Thyroid. Advance online publication. https://doi.org/10.1177/10507256251372175

Sasseville, M., et al. (2025). Bias mitigation in primary health care artificial intelligence models: A scoping review. *JMIR Publications Inc.*

Sohn, W., Sohn, M. H., & Son, J. (2025). Insights into motor impairment assessment using myographic signals with artificial intelligence: a scoping review. *Biomedical Engineering Letters,* 15, 693–716. https://doi.org/10.1007/s13534-025-00483-7

Tubbs, A., & Álvarez-Vázquez, E. (2025). Digital twins in increasing diversity in clinical trials: A systematic review. *Journal of Biomedical Informatics,* 169, 104879. https://doi.org/10.1016/j.jbi.2025.104879

Willem, T., Fritzsche, M. C., Zimmermann, B. M., & Sierawska, A. (2024). Embedded ethics in practice: A toolbox for integrating the analysis of ethical and social issues into healthcare AI research. *Science and Engineering Ethics, 31,* 3. https://doi.org/10.1007/s11948-024-00523-y