VARIÁVEIS UTILIZADAS PARA INVESTIGAR A IMPARCIALIDADE E O VIÉS NA TOMADA DE DECISÕES ALGORÍTMICAS: UMA REVISÃO SISTEMÁTICA E INTEGRATIVA DA LITERATURA

VARIABLES USED TO INVESTIGATE FAIRNESS AND BIAS IN ALGORITHMIC DECISION-MAKING: A SYSTEMATIC INTEGRATIVE LITERATURE REVIEW

DARCI DE BORBA SANTOS JÚNIOR
INSTITUTO DE PESQUISA ECONÔMICA APLICADA - IPEA

HERBERT KIMURA
UNIVERSIDADE DE BRASÍLIA - UNB

Comunicação:
O XIV SINGEP foi realizado em conjunto com a 14th Conferência Internacional do CIK (CYRUS Institute of Knowledge) e com a SPARK (Society for Promoting Academic Research and Knowledge), em formato híbrido, com sede presencial na UNINOVE - Universidade Nove de Julho, no Brasil.

Agradecimento à orgão de fomento:
Este trabalho foi desenvolvido com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo 151855/2025-0, ao qual os autores agradecem pelo suporte concedido à realização da pesquisa.

VARIÁVEIS UTILIZADAS PARA INVESTIGAR A IMPARCIALIDADE E O VIÉS NA TOMADA DE DECISÕES ALGORÍTMICAS: UMA REVISÃO SISTEMÁTICA E INTEGRATIVA DA LITERATURA

Objetivo do estudo
Identificar, classificar e integrar as variáveis utilizadas para investigar equidade e viés na tomada de decisão algorítmica, propondo uma cadeia de mensuração que conecte antecedentes, sistemas algorítmicos, métricas, decisões e consequências sociais e perceptuais.

Relevância/originalidade
O estudo oferece uma síntese transversal inédita das variáveis empregadas na pesquisa sobre equidade algorítmica e integra tradições computacionais, econômicas, comportamentais e organizacionais em um framework capaz de revelar lacunas entre mensuração, decisão e impacto social.

Metodologia/abordagem
Realizou-se uma revisão sistemática integrativa, utilizando o PRISMA 2020 como guia de relato. Foram analisados 56 documentos elegíveis publicados entre 2020 e 2026, codificados quanto a atributos sensíveis, métricas, resultados, métodos e fundamentos teóricos.

Principais resultados
Foram identificadas seis categorias de variáveis. Gênero e raça predominam entre os atributos sensíveis, enquanto paridade estatística, impacto díspar e equalized odds concentram as métricas. Variáveis proxy são frequentes, mas consequências socioeconômicas permanecem pouco mensuradas.

Contribuições teóricas/metodológicas
O artigo propõe a Algorithmic Fairness Measurement Chain, que integra tradições antes desconectadas e organiza as variáveis ao longo do processo decisório. A estrutura oferece orientação metodológica para selecionar medidas, combinar níveis de análise e desenvolver estudos empíricos.

Contribuições sociais/para a gestão
O framework apoia pesquisadores, gestores, desenvolvedores e formuladores de políticas na auditoria de sistemas algorítmicos, identificação de variáveis proxy, escolha de métricas e fortalecimento da transparência, responsabilização e governança responsável da inteligência artificial em decisões de alto impacto.

Palavras-chave: imparcialidade algorítmica, viés algorítmico, atributos sensíveis, variáveis ??proxy, revisão integrativa da literatura

VARIABLES USED TO INVESTIGATE FAIRNESS AND BIAS IN ALGORITHMIC DECISION-MAKING: A SYSTEMATIC INTEGRATIVE LITERATURE REVIEW

Study purpose
To identify, classify, and integrate the variables used to investigate fairness and bias in algorithmic decision-making, proposing a measurement chain that connects antecedents, algorithmic systems, metrics, decisions, and societal and perceptual consequences.

Relevance / originality
The study provides an original cross-domain synthesis of variables used in algorithmic fairness research and integrates computational, economic, behavioral, and organizational traditions into a framework that reveals gaps between measurement, decision-making, and societal impact.

Methodology / approach
A systematic integrative literature review used PRISMA 2020 as a reporting guide. Fifty-six eligible documents published between 2020 and 2026 were coded for sensitive attributes, fairness metrics, outcomes, methods, and theoretical foundations.

Main results
Six variable categories were identified. Gender and race dominate sensitive attributes, while statistical parity, disparate impact, and equalized odds prevail among metrics. Proxy variables are recurrent, whereas downstream socioeconomic consequences remain largely unmeasured.

Theoretical / methodological contributions
The article proposes the Algorithmic Fairness Measurement Chain, integrating previously disconnected traditions and organizing variables throughout the decision process. The framework provides methodological guidance for selecting measures, combining levels of analysis, and designing future empirical studies.

Social / management contributions
The framework supports researchers, managers, developers, and policymakers in auditing algorithmic systems, identifying proxy variables, selecting appropriate metrics, and strengthening transparency, accountability, and responsible artificial intelligence governance in high-stakes decisions.

Keywords: algorithmic fairness, algorithmic bias, sensitive attributes, proxy variables, integrative literature review

VARIABLES USED TO INVESTIGATE FAIRNESS AND BIAS IN ALGORITHMIC DECISION-MAKING: A SYSTEMATIC INTEGRATIVE LITERATURE REVIEW

1. INTRODUCTION

The automation of consequential decisions through machine learning has redefined how organizations allocate opportunities and resources. Credit scoring, hiring, university admissions, clinical risk stratification, and emergency response now rely, at least partially, on predictive models trained on historical data (Kordzadeh & Ghasemaghaei, 2022; Mehrabi et al., 2021). This transformation has generated well-documented efficiency gains, but it has also raised concerns that algorithmic systems reproduce and sometimes amplify historical patterns of discrimination (Barocas & Selbst, 2016; Obermeyer et al., 2019).

Literature distinguishes algorithmic bias, referring to systematic deviations introduced by data or design; unfairness, referring to violations of normative criteria; and algorithmic discrimination, referring to unjustified disparities affecting protected groups. This article uses fairness and bias as the broader designation for systematic and unjustified disparities in algorithmic treatment or outcomes across sensitive groups (Friedman & Nissenbaum, 1996; Mehrabi et al., 2021).

This heterogeneity creates a practical difficulty for empirical projects that aim to estimate the effects of fairness bias on people's lives. Before asking whether biased algorithmic decisions affect employability and income, a researcher must know which variables the literature considers constitutive of fairness bias, how these variables are operationalized, and which measurement traditions can be combined. Reviews of fairness in machine learning exist (Caton & Haas, 2024; Mehrabi et al., 2021; Pessach & Shmueli, 2022), but they are organized around mitigation techniques or fairness definitions rather than around the variables employed in empirical investigation. Domain-specific reviews, such as the systematic review of bias mitigation in credit decisions by De Castro Vieira et al. (2025) and the review of performance, fairness, and explainability in credit scoring by Bahlool et al. (2026), confirm that the choice of sensitive attributes and metrics is narrow and weakly standardized, yet they do not provide a cross-domain taxonomy of investigation variables.

The present article addresses this gap by asking: *which variables are used in the scientific literature to investigate fairness and bias in algorithmic decision-making, and how are they organized across measurement traditions?* To answer this question, a systematic integrative literature review was conducted in accordance with the PRISMA 2020 statement (Page et al., 2021), resulting in a corpus of 56 peer-reviewed documents published between 2020 and 2026. Each study was coded for its sensitive attributes, bias sources, fairness metrics, performance measures, outcome variables, methods, and theoretical foundations. The contribution is fourfold. First, and centrally, the review shows that research on algorithmic fairness is fragmented across measurement traditions that each observe a different segment of the same causal process, and integrates these segments into an algorithmic fairness measurement chain that connects the sources of bias to fairness metrics, decisions, and societal and perceptual consequences. Second, it consolidates a taxonomy of six variable categories that populate this chain. Third, it quantifies the relative attention given to specific sensitive attributes and metrics, revealing systematic blind spots, notably socioeconomic status, disability, intersectionality, and longitudinal socioeconomic outcomes. Fourth, it maps the non-explicit variables that the literature documents as unintentional proxies for sensitive attributes, a mapping of direct value where sensitive data are unavailable or legally restricted. The chain applies to high-stakes algorithmic decision-making in general and offers an operational basis for studies that link algorithmic bias to welfare consequences such as employability and income,

a connection of particular relevance for credit markets in highly unequal countries (Fonseca & Matray, 2024; Kanungo & Gupta, 2021).

2. THEORETICAL BACKGROUND

2.1. FAIRNESS BIAS IN ALGORITHMIC SYSTEMS

The recognition that computer systems can embed bias predates modern machine learning. Friedman and Nissenbaum (1996) distinguished preexisting bias, inherited from social institutions and reflected in data; technical bias, arising from system design constraints; and emergent bias, which appears when a system is used in contexts different from those anticipated. Contemporary scholarship retains this tripartite intuition while relocating its center of gravity to data-driven learning. Barocas and Selbst (2016) showed that disparate impact can emerge from routine data mining practices, such as the definition of target variables, the labeling of training examples, and the selection of features, even in the absence of discriminatory intent. Mehrabi et al. (2021) catalogued more than twenty specific bias mechanisms, from historical and representation bias to measurement and aggregation bias, and demonstrated that these mechanisms propagate along the entire machine learning pipeline.

Two properties make algorithmic bias analytically demanding. First, removing sensitive attributes does not eliminate discrimination when correlated features operate as proxies. Second, small biases may be amplified through model training and feedback loops. Fairness must therefore be assessed explicitly rather than inferred from the absence of sensitive inputs (Barocas & Selbst, 2016; Ferrara, 2024; Incremona et al., 2026).

2.2. FORMAL FAIRNESS CRITERIA AND MITIGATION AS MEASURED INTERVENTIONS

Computational literature has produced a family of formal fairness criteria that operationalize distinct normative intuitions. Group fairness criteria compare statistics across groups defined by a sensitive attribute. Demographic parity, and its operational versions statistical parity difference (SPD) and disparate impact (DI), require similar rates of favorable outcomes across groups (Feldman et al., 2015; Verma & Rubin, 2018). Error-based criteria, such as equalized odds and equal opportunity, require similar error rates conditional on the true outcome (Hardt et al., 2016). Individual fairness requires that similar individuals be treated similarly, formalized through Lipschitz conditions on a task-specific similarity metric (Dwork et al., 2012). Counterfactual fairness requires that predictions remain unchanged in a counterfactual world where the individual's sensitive attribute is altered, grounding fairness in causal models (Kusner et al., 2017).

These criteria are not jointly satisfiable except under degenerate conditions. Chouldechova (2017) proved that calibration and error-rate balance cannot hold simultaneously when base rates differ across groups, and Kleinberg et al. (2017) generalized this impossibility. The practical consequence, extensively documented in the corpus reviewed here, is that empirical studies must select a subset of metrics and justify the selection with reference to the decision context (Criscuolo et al., 2026; Dehghani et al., 2025). A further consequence is the fairness-accuracy trade-off, that improving a fairness criterion frequently degrades predictive performance, generating a Pareto frontier that has itself become an object of measurement (Gunasekara & Saarela, 2026; S. Liu & Vicente, 2022).

Mitigation techniques are conventionally classified as pre-processing, which transforms data before training; in-processing, which modifies learning objectives; and post-processing, which adjusts model outputs. Empirical studies commonly compare fairness metrics before and after these interventions, simultaneously assessing predictive performance and the fairness–accuracy trade-off (De Castro Vieira et al., 2025; Dehghani et al., 2025).

A separate tradition investigates fairness through the perceptions of people affected by algorithmic decisions, drawing primarily on organizational justice theory. Evidence indicates that objective debiasing does not necessarily generate perceived fairness and that judgments vary according to transparency, outcome favorability, decision-agent type, and individual characteristics. This literature contributes variables such as procedural justice, trust, moral judgment, and technology acceptance, which are largely absent from computational research (Choi & Chao, 2026; Y. Li et al., 2025; Newman et al., 2020).

Fairness bias also produces observable socioeconomic consequences. In credit markets, discriminatory approval and pricing restrict access to consumption, entrepreneurship, and human capital investment, while biased systems in health and emergency response may generate unequal service allocation and welfare losses. However, empirical studies rarely connect algorithmic decisions to longitudinal outcomes such as income, employability, or mobility, a gap directly addressed by the present review (Bahamazava, 2026; Bartlett et al., 2022; Greenwald et al., 2024).

3. METHOD

3.1. REVIEW DESIGN AND PROTOCOL

The study adopts a systematic integrative literature review design (Torraco, 2005), using PRISMA 2020 as a reporting guide (Page et al., 2021). Because the research question concerns the diversity of variables used across heterogeneous traditions, the synthesis is qualitative and descriptive, without meta-analytic aggregation of effect sizes.

The review protocol defined, prior to the analysis, the research question, the eligibility criteria, the search strategy, and the coding dimensions. Eligibility required that a document: (a) address fairness or bias in algorithmic or AI-supported decision-making as a central topic; (b) be published in a peer-reviewed journal or, exceptionally, as an indexed editorial or published conference abstract, retained for descriptive mapping with reduced weight in the synthesis; (c) possess verifiable bibliographic metadata, including a DOI; and (d) provide sufficient methodological detail to allow the coding of variables. Preprints without peer review and documents in non-indexed outlets were excluded.

Figure 1 presents the PRISMA 2020 flow of the review. In the identification stage, two sources were combined. The first is a seed corpus of 53 documents assembled in the context of a funded research project on algorithmic justice in credit granting, retrieved from major scientific databases (ScienceDirect, Scopus, and Web of Science) using combinations of the terms *fairness*, *bias*, *algorithmic discrimination*, *machine learning*, and domain qualifiers, complemented by expert curation to ensure coverage of the health, education, recruitment, credit, and public service domains. The second source is an updating search performed in July 2026 on publisher databases and scholarly search engines, aimed at capturing documents published after the assembly of the seed corpus. The updating search combined the query (“algorithmic fairness” OR “fairness bias” OR “algorithmic bias”) AND (“machine learning” OR “artificial intelligence”), restricted to documents dated 2025–2026, was executed on 19 July 2026, and returned 9 additional candidate records. Because the seed corpus combined database retrieval with expert curation, its search strategy cannot be re-executed verbatim record by record; this constraint motivates the integrative designation of the review and is registered as a limitation in Section 5.

Eligible documents addressed fairness or bias in algorithmic decision-making, were published in peer-reviewed or indexed outlets, possessed verifiable bibliographic metadata, and provided sufficient methodological detail for variable coding. Preprints and documents from non-indexed outlets were excluded.

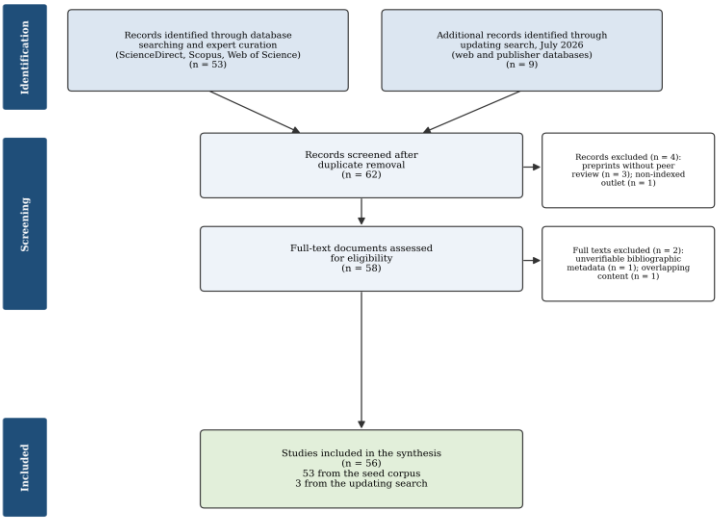


Figure 1: PRISMA 2020 flow diagram of the review, with the number of records processed at each stage

Two atypical indexed documents, one editorial and one published conference abstract, were retained only for descriptive mapping and were not used as primary evidence.

3.2. DATA EXTRACTION, CODING, AND SYNTHESIS

Each included document was read in full and coded along eight dimensions: bibliographic metadata (authors, year, journal); application domain; research design; theoretical foundations; sensitive attributes and other antecedent variables; fairness and performance metrics; outcome variables; and mitigation strategies, when present. Two additional fields, recorded for methodological transparency, distinguish each study's object of analysis, the abstract focus under investigation, from its object of observation, the empirical materialization of that focus in datasets, populations, or systems. The coding scheme was developed abductively: an initial scheme derived from the fairness literature (Mehrabi et al., 2021; Verma & Rubin, 2018) was iteratively refined as new variable types emerged from the corpus, notably perceptual variables from the organizational justice tradition and welfare variables from the economics tradition. Classification rules were explicit about exclusivity: application domain and research design were coded as mutually exclusive categories, whereas sensitive attributes, fairness metrics, and theoretical foundations were coded as non-exclusive, so that a single study may contribute to several counts; all frequencies reported in Section 4 follow these rules.

Coding results were consolidated into a structured matrix, from which the descriptive statistics, the taxonomy, and the proxy mapping presented in Section 4 were derived. Coding was performed by a single researcher in two full passes separated in time, guided by a structured codebook; in the second pass, ambiguous cases were revisited against the original text, and the final matrix was audited for internal consistency between the coded variables and the reported methods. No independent double coding was performed, and no inter-coder agreement statistics could therefore be computed, a reliability constraint acknowledged in Section 5.

4. DATA ANALYSIS

4.1. CORPUS OVERVIEW

Publication activity is strongly concentrated in the period 2024–2026, which accounts for 52 of the 56 documents and reflects the acceleration of fairness research after the consolidation of the field's benchmark tools and datasets. Health care and biomedicine constitute the most frequent application domain (11 studies), followed by general-purpose machine learning research (10), software engineering (5), recruitment and human resources (5),

and education (4). Emerging technical subfields are visible as distinct clusters: fairness in large language models (Y. Li et al., 2025; Lu et al., 2025; Radaideh et al., 2025), in graph neural networks (C. Li et al., 2026; S. Liu et al., 2026; Zhen et al., 2025), and in federated learning (Kim et al., 2025; Z. Wang, Sun, et al., 2026; Z. Wang, Yan, et al., 2026). Credit and finance, the domain of immediate interest for research on socioeconomic consequences, is represented by three documents (Bahlool et al., 2026; De Castro Vieira et al., 2025; Greenwald et al., 2024).

Methodologically, the corpus is dominated by algorithmic experimentation and benchmarking (32 studies), in which fairness bias is investigated by training models on benchmark datasets and measuring fairness metrics under different mitigation conditions. Empirical studies using secondary real-world data (7), reviews and editorials (7), conceptual or design science contributions (4), behavioral experiments (3), mixed-methods studies (2), and practitioner surveys (1) complete the distribution. This distribution is itself a finding: the

investigation of fairness bias remains predominantly a laboratory activity conducted on a recurrent set of public datasets, with comparatively few studies measuring bias in operating systems or in populations affected by deployed algorithms.

4.2. A TAXONOMY OF VARIABLES USED TO INVESTIGATE FAIRNESS BIAS

The coding yielded six categories of variables, summarized in Table 1 and integrated into the Algorithmic Fairness Measurement Chain shown in Figure 2. The chain links antecedent variables, algorithmic systems, fairness measurement, decision outcomes, and societal and perceptual consequences, while contextual variables moderate these relationships. The taxonomy classifies the variables, whereas the framework represents their position in the decision process.

Table 1: Taxonomy of variables used in the investigation of fairness bias

Category	Typical variables	References
1. Sensitive (protected) attributes	Gender/sex; race/ethnicity; age; socioeconomic status; nationality; religion; disability; intersectional combinations	(De Castro Vieira et al., 2025; Gunasekara & Saarela, 2026; S. Liu et al., 2026; Uddin et al., 2026)
2. Data-level bias variables	Representation and sampling bias; selection bias; label noise; proxy features; missing or incomplete sensitive attributes; distributional distance (e.g., Earth Mover's Distance); bias symptoms	(Canalli et al., 2024; d'Aloisio et al., 2025; Pessach & Shmueli, 2021; Uddin et al., 2026; Z. Wang, Sun, et al., 2026)
3. Fairness measurement variables	Statistical parity difference; disparate impact; equalized odds; equal opportunity difference; individual fairness; counterfactual fairness; calibration; distributional indices (Gini); equality-equity vectors; content-level metrics for generative models	(Dam et al., 2026; Dwork et al., 2012; Gonzalez-Sendino et al., 2025; Hardt et al., 2016; Lu et al., 2025)
4. Performance and trade-off variables	Accuracy; F1-score; AUC; utility; fairness-accuracy (Pareto) trade-off; runtime and scalability; fairness improvement percentage	(Dehghani et al., 2025; Gunasekara & Saarela, 2026; Incremona et al., 2026; Zhen et al., 2025)
5. Decision and societal outcome variables	Approval/selection rates; rankings and exposure; resource allocations; error disparities; income; employability; health outcomes; social welfare losses	(Bahamazava, 2026; Fabris et al., 2026; Greenwald et al., 2024; Mersha et al., 2025)
6. Perceptual, behavioral, and contextual variables	Perceived (procedural) fairness; trust; moral judgment; transparency; outcome favorability; organizational ethical climate; practitioner practices; decision stakes; human-AI interaction factors	(Choi & Chao, 2026; Miazek & Bocian, 2025; Newman et al., 2020; Saiyed & Kumar, 2025; Song et al., 2025; Y. Wang et al., 2024)

Note: systematic reviews, surveys, and editorials cited in this table are listed because they document the corresponding variables or techniques across primary studies; they do not themselves apply the techniques or measure the variables.

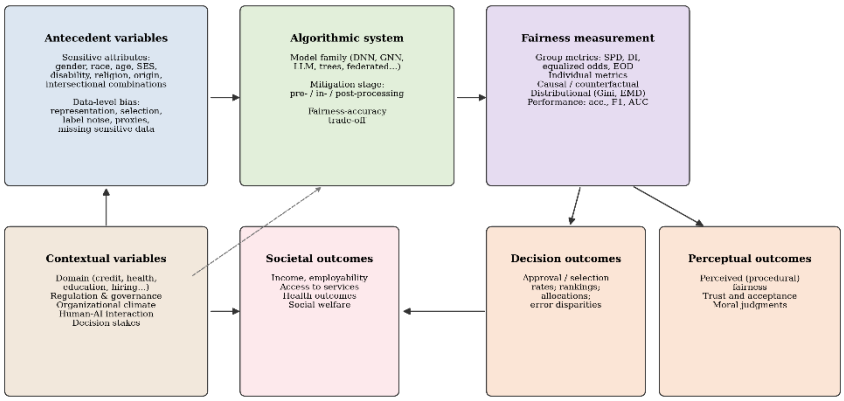


Figure 2: The algorithmic fairness measurement chain connecting the six categories of variables identified in the review

Notes: the algorithmic fairness measurement chain connecting the six categories of variables identified in the review. Solid arrows represent the dominant measurement chain in the corpus; the dashed arrow indicates contextual moderation of the algorithmic system.

4.2.1. Sensitive attributes and their non-explicit proxies

Sensitive attributes are the pivotal antecedent variables of the field, every group fairness metric is computed conditional on them. Figure 3 shows their frequency in the corpus. Gender or sex is the most frequently operationalized attribute (16 studies), followed closely by race or ethnicity (14). Age and socioeconomic status appear in five studies each, while nationality or national origin, religion, disability, and explicitly intersectional combinations appear only marginally. This distribution replicates, at cross-domain scale, the concentration documented by De Castro Vieira et al. (2025) for credit scoring, where gender and race were also the most studied attributes and statistical parity the most common metric.

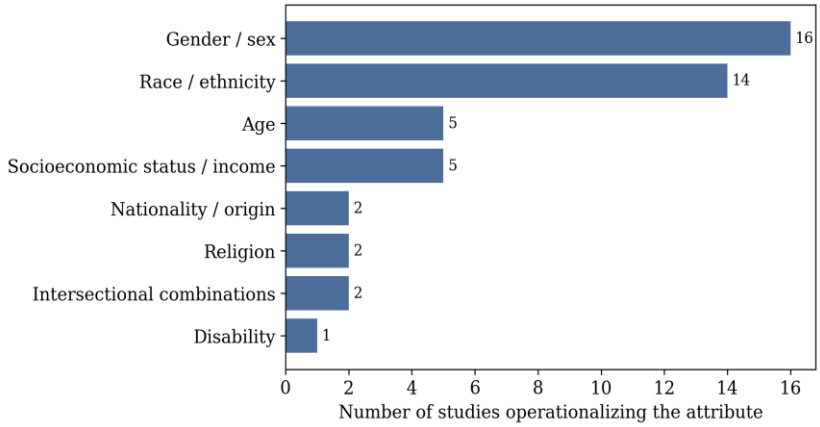


Figure 3: Frequency of sensitive attributes explicitly operationalized in the corpus

Note: n = 56; multiple attributes per study possible.

The explicit attributes, however, tell only part of the story. A central and recurring finding of the corpus is that fairness bias operates through non-explicit variables that encode sensitive attributes unintentionally: features that are legitimate on their face but statistically dependent on race, gender, or class, and that allow models to reconstruct group membership even when sensitive inputs are removed (Barocas & Selbst, 2016; Incremona et al., 2026; Mehrabi et al., 2021). Table 2 consolidates the proxy variables documented across the reviewed studies, the sensitive attributes they encode, and the mechanism through which the encoding occurs. The mapping shows that proxies are not exotic artifacts but ordinary administrative and

behavioral variables: residential location inherits the geography of segregation and redlining (Bartlett et al., 2022); surnames combined with geolocation predict race well enough to support regulatory algorithms, and badly enough to distort measured disparities (Greenwald et al., 2024); occupation, schooling, and insurance type carry the imprint of stratified labor markets and health systems (Radaideh et al., 2025; X. Wang et al., 2024); the very absence of credit history marks the financially invisible, who are concentrated in disadvantaged groups (De Castro Vieira et al., 2025; Kozodoi et al., 2022); and in networked data, homophily makes graph topology itself a proxy, since who one is connected to predicts who one is (C. Li et al., 2026; S. Liu et al., 2026; Zhen et al., 2025).

Table 2: Non-explicit variables documented as unintentional proxies for sensitive attributes

Proxy variable	Attribute(s) encoded	Mechanism of encoding	References
Residential location, ZIP code, neighborhood	Race/ethnicity; socioeconomic status	Historical segregation and redlining persist in geocoded records, making location predictive of group membership	(Bartlett et al., 2022; Greenwald et al., 2024)
Surname (esp. combined with geolocation)	Race/ethnicity; national origin	Onomastic regularities support algorithmic race prediction (e.g., BISG), with systematic errors for minority groups	(Greenwald et al., 2024)
Occupation and job titles	Gender; socioeconomic status	Gendered and class-based occupational segregation is reflected in records and in language models' associations	(Barocas & Selbst, 2016; Radaideh et al., 2025)
Educational background (attainment, school type)	Race/ethnicity; socioeconomic status	Unequal access to schooling makes educational variables systematically different across groups	(Barocas & Selbst, 2016; Y. Wang et al., 2024)
Health care costs and utilization	Race/ethnicity	Unequal access to care makes spending a biased measure of health need	(Obermeyer et al., 2019)
Health insurance type	Race/ethnicity; socioeconomic status	Coverage type correlates with income, employment formality, and group membership	(Y. Wang et al., 2024)
Credit history length; thin or absent files	Race/ethnicity; socioeconomic status; age	'Financially invisible' consumers are concentrated in disadvantaged groups, so file thinness marks group membership	(De Castro Vieira et al., 2025; Kozodoi et al., 2022)
Digital access, device, and social media activity	Socioeconomic status; age; minority status; rurality	The digital divide biases who appears in behavioral data and at what intensity	(Griffin et al., 2024; K. Wang et al., 2025)
Housing type and territorial exposure (e.g., flood-prone areas)	Socioeconomic status; minority share	Territorial vulnerability correlates with demographic composition of communities	(K. Wang et al., 2025)
Network structure (connections, homophily)	Race; gender; religion	Homophily makes graph topology predictive of node group membership, biasing GNN representations	(C. Li et al., 2026; S. Liu et al., 2026; Zhen et al., 2025)
Names and origin cues in profiles and text	Gender; national origin; ethnicity	Linguistic and onomastic cues trigger demographic inference by models and human evaluators	(Fabris et al., 2026; Lu et al., 2025)
Any feature statistically dependent on a sensitive attribute	All	Mutual information between features and attributes allows indirect learning of bias after attribute removal	(Incremona et al., 2026; Mehrabi et al., 2021)

Note: systematic reviews, surveys, and editorials cited in this table are listed because they document the corresponding variables or techniques across primary studies; they do not themselves apply the techniques or measure the variables.

Three refinements complete the picture of this category. First, several studies treat the sensitive attribute not as given but as measured with error or partially missing: Greenwald et al. (2024) showed that proxy-based race prediction distorts measured lending disparities, Wang, Sun, et al., (2026) developed methods for fairness under incomplete sensitive attributes in

federated settings, and Li et al. (2025) avoided sensitive labels altogether for privacy reasons. Second, a small group of studies operationalizes intersectionality, defining groups by combinations such as gender and race or gender and income (C. Li et al., 2026; Pham et al., 2025). Third, one study inverts the logic of attribute selection: rather than assuming which attributes are sensitive, Wang et al. (2025) mined 35 socioenvironmental variables to identify empirically which community characteristics, such as minority share and road density, require fairness protection. This data-driven identification of sensitive attributes, combined with the proxy mapping of Table 2, is directly transferable to contexts where structural inequality does not map neatly onto legally protected categories.

4.2.2. Data-level bias variables

A second category of variables characterizes the data-generating process rather than the individuals. Studies in this category measure or, in the case of surveys, systematically document representation and sampling bias (Uddin et al., 2026; Yang et al., 2024), selection bias induced by the decision process itself, as in hiring data where labels exist only for accepted candidates (Pessach & Shmueli, 2021), label noise as a source of unfairness (Canalli et al., 2024), and statistical "bias symptoms" computable from datasets before any model is trained (d'Aloisio et al., 2025). Uddin et al. (2026) illustrate the scale that such audits can reach, quantifying gender imbalance in 74 open-access disease prediction datasets through the Earth Mover's Distance and linking dataset bias to downstream fairness violations.

The theoretical significance of this category lies in its causal position. Data-level variables are antecedents of fairness bias, and interventions on them constitute the mitigation family with the widest adoption, from fair data generation via causal models (Gonzalez-Sendino et al., 2024) to synthetic data produced by large language models (Voria et al., 2026) and the broad family of preprocessing techniques whose predominance across primary studies is documented by surveys and domain reviews (De Castro Vieira et al., 2025; Tawakuli & Engel, 2025). Measuring these variables early is also economical, d'Aloisio et al. (2025) show that bias-inducing variables can be predicted from dataset statistics before any model is trained, sparing the computational and social cost of training biased systems that must later be corrected.

4.2.3. Fairness measurement, performance, and trade-off variables

The third category comprises the metrics through which fairness bias is rendered observable. Across the corpus, group metrics dominate. Statistical parity difference and disparate impact are the most recurrent, reported in credit (De Castro Vieira et al., 2025), education (Gunasekara & Saarela, 2026; Pham et al., 2025), recruitment (Albaroudi et al., 2026), and health applications of large language models (Lu et al., 2025). Error-based metrics, particularly equalized odds and equal opportunity difference, are standard in studies that frame fairness as parity of mistakes rather than parity of outcomes (Gunasekara & Saarela, 2026; Hardt et al., 2016; Sivarajkumar et al., 2023; Uddin et al., 2026). Individual fairness appears mainly in methodological work that contrasts it with group criteria, revealing that interventions optimizing one frequently destabilize the other (Dwork et al., 2012; Qin & Merlo, 2026).

Beyond these standards, the corpus documents an active production of new measurement variables. Gonzalez-Sendino et al. (2025) proposed a two-dimensional vector separating equality from equity, together with a fairness curve for interpreting disparities. Qin et al. (2025) introduced representation-based measures (computational profile distances) that assess fairness from a model's internal activations. Wang et al., (2026) contributed a statistically guaranteed evaluation framework that distinguishes systematic from heterogeneous bias in federated deployments. For generative models, whose outputs are texts rather than classifications, content-level metrics such as diagnostic recommendation divergence and referral frequency parity have been developed (Lu et al., 2025), and paired-prompt designs quantify bias as the divergence of outputs under counterfactual demographic cues (Radaideh et

al., 2025). Distributional indices, such as the Gini coefficient applied to exposure across topics or groups, extend measurement to ranking and recommendation contexts, through direct application in primary studies and editorial mapping of the subfield (Boratto et al., 2022; Dam et al., 2026). The proliferation of metrics confirms the theoretical impossibility results (Chouldechova, 2017; Kleinberg et al., 2017), because no single criterion captures all normative intuitions, the choice of measurement variables is itself a substantive research decision that must be aligned with the decision context (Criscuolo et al., 2026; Dehghani et al., 2025).

Because fairness interventions can degrade predictive quality, virtually all experimental studies pair fairness metrics with performance variables, accuracy, F1-score, and AUC being the most common (Dehghani et al., 2025; Gunasekara & Saarela, 2026). The trade-off itself is increasingly treated as a measured object: studies report Pareto frontiers between fairness and utility (S. Liu & Vicente, 2022; Zhen et al., 2025), fairness improvement percentages under different mitigation stages (Dehghani et al., 2025; Pham et al., 2025), and even computational costs of fairness, including runtime scaling (Incremona et al., 2026) and energy or carbon footprints of debiased systems (Albaroudi et al., 2026). The empirical regularity emerging from benchmarking studies is that no classifier or mitigation technique dominates across datasets, metrics, and contexts (Criscuolo et al., 2026; Dehghani et al., 2025), which reinforces the need for context-specific variable selection.

4.2.4. Outcome, perceptual, and contextual variables

The fifth category captures the consequences of algorithmic decisions, and it is here that the corpus is thinnest. Within-system decision outcomes are well represented: approval and selection rates, shortlisting outcomes, rankings and exposure (Fabris et al., 2026; Greenwald et al., 2024). Notably, Fabris et al. (2026) demonstrated experimentally that fair exposure in candidate rankings does not translate into fair shortlisting outcomes, because recruiter behavior, interface design, and position bias mediate the link between algorithmic output and final decision. Beyond the decision itself, however, few studies measure welfare consequences. Bahamazava (2026) modeled response-time equity, health outcomes, and social welfare losses from biased emergency systems; Mersha et al. (2025) examined disparities in length of stay and readmission; and the economics literature cited in the corpus links credit discrimination to interest burdens and access restrictions (Bartlett et al., 2022; Kozodoi et al., 2022). Variables such as employability, income trajectories, and mobility, central to the social meaning of fairness bias, are effectively absent from the empirical corpus, appearing only as motivating concerns. This is the most consequential gap identified by the review.

The final category originates in behavioral and organizational literature. Perceived procedural fairness is the anchor variable, measured through validated organizational justice scales and manipulated experimentally by varying the decision agent (Colquitt, 2001; Newman et al., 2020). Determinants of perceived fairness measured in the corpus include outcome favorability, transparency, algorithmic bias level, individual characteristics such as field of study and anxiety (Song et al., 2025), and self-interest, whose egocentric distortion of fairness judgments is attenuated for algorithmic decision-makers relative to human ones (Choi & Chao, 2026; Miazek & Bocian, 2025). At the organizational level, Saiyed and Kumar (2025) measured ethical concerns, ethical climate, and sustainability-oriented human resource outcomes through structural equation modeling, Voria et al. (2025) surveyed 155 practitioners on the frequency, impact, and effort of 28 fairness-aware development practices, and Rai et al. (2026) theorized the organizational, technical, and governance choices through which firms adapt fairness to regulatory mandates and societal norms. Contextual variables also include properties of the human-AI decision configuration, such as evidence and technology effectiveness, human bias awareness, and decision stakes, which (Y. Wang et al., 2024) showed to condition whether decision-support technologies increase or decrease outcome fairness. The presence of this

category demonstrates that fairness bias cannot be fully investigated within the model, identical algorithmic properties produce different fairness outcomes depending on how humans and organizations interact with the system (Fabris et al., 2026; Rai et al., 2026; X. Wang et al., 2024).

Methodologically, the reviewed studies combine conventional statistical techniques, including regression, hypothesis testing, matching, simulation, and structural equation modeling, with machine learning algorithms ranging from classical classifiers to deep neural networks, graph models, large language models, and federated learning. Because these techniques serve primarily as analytical vehicles for investigating the variables identified in the taxonomy, they are not treated as a separate contribution of the review.

4.3. THEORETICAL FOUNDATIONS ACROSS THE CORPUS

The coding of theoretical foundations reveals a stratified field. The computational core relies on formal fairness theory (group, individual, and counterfactual definitions), information theory (Incremona et al., 2026; Y. Li et al., 2026), causal modeling (Gonzalez-Sendino et al., 2024; Kusner et al., 2017; Y. Li et al., 2026), game theory (Manganini et al., 2026), and learning-theoretic constructs such as the Rashomon set of near-optimal models (M. Liu et al., 2024). The behavioral stratum draws on organizational justice theory (Choi & Chao, 2026; Newman et al., 2020), the Computers Are Social Actors framework and egocentric ethics (Miazek & Bocian, 2025), and technology acceptance and stakeholder theories (Saiyed & Kumar, 2025).

A normative and organizational stratum completes the field, invoking distributive justice distinctions between equality and equity (Gonzalez-Sendino et al., 2025), medical ethics (Ho & Asrani, 2025), governance frameworks (Albaroudi et al., 2026; Griffin et al., 2024), design theory for sociotechnical systems (Rai et al., 2026), and structured quality appraisal frameworks for the literature itself (Bahlool et al., 2026). Cross-references between strata are rare, computational studies seldom cite justice theory, and behavioral studies seldom engage formal metrics, a disconnection also visible in the variables each stratum measures. The corpus updated to 2026 shows the first signs of integration, with design-oriented work explicitly linking technical choices to governance and regulation (Bahlool et al., 2026; Rai et al., 2026).

5. DISCUSSION

5.1. CONVERGENCES AND THE MEASUREMENT CORE

The analysis indicates that, notwithstanding the field's heterogeneity, a measurement core has stabilized. The modal empirical study of fairness bias operationalizes one or two sensitive attributes, almost always gender or race; computes a small set of group metrics, typically statistical parity difference, disparate impact, or equalized odds; pairs them with accuracy-type performance measures; and evaluates a mitigation intervention classified as pre-, in-, or post-processing. This core provides commensurability across domains as diverse as credit scoring (De Castro Vieira et al., 2025), learning analytics (Gunasekara & Saarela, 2026), and clinical prediction (Dehghani et al., 2025), and it supplies ready-made instruments for new empirical projects.

The stability of the core is confirmed by the most recent reviews. Bahlool et al. (2026), synthesizing 43 credit scoring studies, found that performance, fairness, and explainability are predominantly addressed in isolation, with the same concentration of metrics and attributes documented here. For empirical work in any high-stakes domain, lending, hiring, admissions, or clinical triage, the immediate implication is that group metrics conditioned on gender, race, and regional attributes can be computed with established tooling and compared against international benchmarks, while the joint treatment of fairness with explainability and regulatory constraints remains an open frontier.

Read through the measurement chain, the corpus resolves into four measurement traditions whose instruments rarely meet, summarized in Table 3. The computational tradition observes the beginning of the chain and stops at the metrics; the economic tradition begins where the computational one stops, at decisions and welfare; the behavioral tradition observes how the chain is perceived rather than how it operates; and the organizational tradition studies the context that conditions every link. The fragmentation documented in Section 4 is therefore not an accident of topic choice but a structural property of the field, each tradition measures the segment of the chain that its methods can reach.

Table 3: Measurement traditions in research on algorithmic fairness and bias

Measurement tradition	Main variables	Dominant methods	Typical domains	Main blind spot
Computational / algorithmic	Sensitive attributes; data-level bias; fairness metrics; performance trade-offs	Benchmarking; algorithm development; formal analysis	General ML; healthcare; education; graphs; NLP/LLMs; federated learning	Stops at the metrics: decisions and welfare consequences remain unobserved
Economic / econometric	Decision outcomes; measured disparities; income and welfare losses	Econometric regression; structural and welfare modeling; simulation	Credit and finance; emergency services	Rarely engages formal fairness metrics or model internals
Behavioral / organizational justice	Perceived (procedural) fairness; trust; moral judgments; outcome favorability	Laboratory and field experiments; surveys; mixed methods	Human resources; education; consumer contexts	Seldom links perceptions to measured statistical bias
Organizational / design and governance	Governance choices; ethical climate; fairness practices; regulatory context	Design science; SEM surveys; conceptual frameworks	Information systems; organizations; health policy	Weak connection to metric-level and outcome-level evidence

5.2. BLIND SPOTS WITH SUBSTANTIVE CONSEQUENCES

The distribution of variables also exposes blind spots whose correction is theoretically consequential. First, socioeconomic status is operationalized in only a handful of studies, despite theoretical arguments that class-based disadvantage interacts with, and is partially constituted by, algorithmic exclusion (Fonseca & Matray, 2024; Kozodoi et al., 2022). In societies marked by high structural inequality, income, informality, and territorial stigma may function as primary axes of algorithmic disadvantage, not merely as proxies for race or gender. The proxy mapping of Table 2 indicates precisely where such disadvantage hides: in location, schooling, insurance, file thinness, and digital access. The data-driven identification of sensitive attributes demonstrated by Wang et al. (2025) offers a template for discovering these axes empirically.

Second, the near absence of intersectional operationalizations (Y. Li et al., 2026; Pham et al., 2025) limits the field's capacity to detect disparities that materialize only at attribute intersections, a limitation that De Castro Vieira et al. (2025) also flag for credit research. Third, sensitive attribute measurement error is usually ignored, although Greenwald et al. (2024) showed that it can halve measured disparities; studies relying on inferred race or geographic proxies inherit this fragility. Fourth, and most importantly for research on the effects of fairness bias on people's lives, downstream welfare variables are almost never measured. The corpus measures fairness where it is cheapest to observe, inside the model, and rarely where it matters most, in the trajectories of the people subject to algorithmic decisions. Designs that link decision data to administrative records of employment and income would populate precisely the societal outcome cell of the framework in Figure 2, converting a motivating narrative into an estimable causal chain.

5.3. IMPLICATIONS FOR EMPIRICAL RESEARCH DESIGNS

Three design implications follow from the review. First, variable selection should be theory-driven and context-specific: parity metrics suit questions of access equity, while error-based metrics suit questions of diagnostic or predictive harm (Dehghani et al., 2025). Reporting multiple metrics, together with their tensions, should be standard, since fewer than a quarter of credit studies report more than one fairness metric (De Castro Vieira et al., 2025). In parallel, the proxy mapping consolidated here implies that empirical designs should audit candidate features against Table 2 before modeling, treating location, schooling, occupation, insurance, and file thinness as potential carriers of sensitive information rather than as neutral controls (Barocas & Selbst, 2016; Greenwald et al., 2024).

Second, empirical designs should incorporate the human and organizational decision layers. The evidence that fair rankings do not produce fair outcomes (Fabris et al., 2026), that perceived fairness diverges from statistical fairness (Choi & Chao, 2026; Newman et al., 2020; Song et al., 2025), and that fairness adaptation is shaped by organizational and governance choices (Rai et al., 2026) implies that studies confined to model outputs can misestimate the bias experienced by affected populations. Third, the measurement chain should be extended toward societal outcomes. The framework proposed here suggests concrete linkages: sensitive attributes and data bias variables predict fairness metric disparities; disparities materialize as decision outcomes (denial rates, allocations); and decision outcomes accumulate into welfare trajectories (income, employability), moderated by context. Each link is measurable with instruments already available in the corpus, but the full chain has not yet been estimated in any single study.

6. CONCLUSION

This review asked which variables the scientific literature employs to investigate fairness and bias in algorithmic decision-making. The PRISMA-informed analysis of 56 documents published between 2020 and 2026 identified six categories: sensitive attributes, data-level bias variables, fairness measurement variables, performance and trade-off variables, decision and societal outcomes, and perceptual, behavioral, and contextual variables. The field exhibits a stable measurement core, organized around gender and race as sensitive attributes and around statistical parity, disparate impact, and equalized odds as metrics, embedded in a benchmarking methodology that evaluates mitigation techniques by pipeline stage. Around this core, the review documented significant innovation in measurement, from equality-equity vectors and representation-based diagnostics to content-level metrics for generative models, and consolidated a mapping of the non-explicit variables, location, surname, occupation, schooling, insurance, credit file thinness, digital access, and network structure, through which sensitive attributes enter models unintentionally.

The review's central conclusion is theoretical: the literature is fragmented because its measurement traditions observe different segments of the algorithmic fairness measurement chain, and progress now depends less on new metrics than on studies that connect segments of the chain. The chain applies to high-stakes algorithmic decision-making in general, from clinical triage to hiring and admission; credit granting stands out as a particularly consequential application because its decisions accumulate directly into income and employment trajectories.

The review has limitations. The seed corpus, although diverse and recent, was purposively assembled around a research program on algorithmic decision-making, and the updating search, while systematic in its criteria, relied on scholarly search engines rather than on a full re-execution of the original database queries; exhaustive coverage with formal dual screening would likely enlarge the set of behavioral and economics studies. The protocol was not registered, as the review falls outside the scope of health-intervention registries, which reduces external verifiability of the pre-specified criteria. The search strategy of the seed

corpus, in particular, cannot be re-executed verbatim, which is why the review claims PRISMA-informed reporting rather than full PRISMA compliance.

The coding of variables, performed by a single researcher under a two-pass protocol but without independent double coding or inter-coder agreement statistics, involves interpretive judgment, particularly in assigning theoretical foundations and in distinguishing objects of analysis from objects of observation. Finally, the frequency counts reported here describe attention in literature, not the importance of the variables; rarely measured variables, such as disability or income trajectories, may be socially central despite their scarcity in published research.

For research programs concerned with these effects, literature supplies mature instruments for the antecedent and measurement segments, but almost no precedents for the final segment, so studies that integrate algorithmic decision data with longitudinal administrative records occupy a position of high marginal scientific value. The taxonomy, the proxy mapping, and the framework offered here are intended as an operational map for such studies: they specify which variables must be collected, from which traditions their instruments derive, which ordinary features must be treated as carriers of sensitive information, and where the unmeasured links of the chain lie.

DECLARATION OF GENERATIVE AI

Generative AI and AI-assisted technologies were used to support Portuguese-to-English translation and to improve the language, clarity, and readability of the manuscript. All outputs were reviewed and revised by the authors, who take full responsibility for the final content.

REFERENCES

Albaroudi, E., Mansouri, T., Hatamleh, M., & Alameer, A. (2026). HitHire: The future of ethical, fair, and sustainable AI recruitment – A governance framework. *Array*, *29*, 100592. https://doi.org/10.1016/j.array.2025.100592

Bahamazava, K. (2026). Implications of algorithmic bias in AI-driven emergency response systems. *Journal of Economy and Technology*, *4*, 20–34. https://doi.org/10.1016/j.ject.2025.05.003

Bahlool, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. *Journal of Risk and Financial Management*, *19*(2), 104. https://doi.org/10.3390/jrfm19020104

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671–732. https://doi.org/10.15779/Z38BG31

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics*, *143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Boratto, L., Faralli, S., Marras, M., & Stilo, G. (2022). Guest editorial of the IPM special issue on algorithmic bias and fairness in search and recommendation. *Information Processing & Management*, *59*(1), 102791. https://doi.org/10.1016/j.ipm.2021.102791

Canalli, Y., Braida, F., Alvim, L., & Zimbrão, G. (2024). Fair Transition Loss: From label noise robustness to bias mitigation. *Knowledge-Based Systems*, *294*, 111711. https://doi.org/10.1016/j.knosys.2024.111711

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), 166. https://doi.org/10.1145/3616865

Choi, J., & Chao, M. M. (2026). For me or against me? Reactions to AI (vs. Human) decisions that are favorable or unfavorable to the self and the role of fairness perception. *Personality and Social Psychology Bulletin*, *52*(3), 671–691. https://doi.org/10.1177/01461672241288338

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, *86*(3), 386–400. https://doi.org/10.1037/0021-9010.86.3.386

Criscuolo, C., Salnitri, M., & Martinenghi, D. (2026). FAIR-CARE: A comparative evaluation of unfairness mitigation approaches. *Information and Software Technology*, *189*, 107898. https://doi.org/10.1016/j.infsof.2025.107898

d'Aloisio, G., Di Sipio, C., Di Marco, A., & Di Ruscio, D. (2025). Towards early detection of algorithmic bias from dataset's bias symptoms: An empirical study. *Information and Software Technology*, *188*, 107905. https://doi.org/10.1016/j.infsof.2025.107905

Dam, A., Pathak, S., & Mitra, B. (2026). FairRec: An expert recommendation framework for fair and topic-aware influence maximization. *Expert Systems with Applications*, *307*, 131071. https://doi.org/10.1016/j.eswa.2025.131071

De Castro Vieira, J. R., Barboza, F., Cajueiro, D., & Kimura, H. (2025). Towards Fair AI: Mitigating Bias in Credit Decisions—A Systematic Literature Review. *Journal of Risk and Financial Management*, *18*(5), 228. https://doi.org/10.3390/jrfm18050228

Dehghani, F., Paiva, P., Malik, N., Lin, J., Bayat, S., & Bento, M. (2025). Accuracy-fairness trade-off in ML for healthcare: A quantitative evaluation of bias mitigation strategies. *Information and Software Technology*, *188*, 107896. https://doi.org/10.1016/j.infsof.2025.107896

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Fabris, A., Rus, C., Saldivar, J., Gatzioura, A., Biega, A. J., & Castillo, C. (2026). Does fair ranking lead to fair recruitment outcomes? A study of interventions, interfaces, and interactions. *Information Processing & Management*, *63*(3), 104506. https://doi.org/10.1016/j.ipm.2025.104506

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. https://doi.org/10.1145/2783258.2783311

Ferrara, E. (2024). The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, *15*, 100525. https://doi.org/10.1016/j.mlwa.2024.100525

Fonseca, J., & Matray, A. (2024). Financial inclusion, economic development, and inequality: Evidence from Brazil. *Journal of Financial Economics*, *156*, 103854. https://doi.org/10.1016/j.jfineco.2024.103854

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gonzalez-Sendino, R., Serrano, E., & Bajo, J. (2024). Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems*, *155*, 384–401. https://doi.org/10.1016/j.future.2024.02.023

Gonzalez-Sendino, R., Serrano, E., & Bajo, J. (2025). Quantifying algorithmic discrimination: A two-dimensional approach to fairness in artificial intelligence. *Engineering Applications of Artificial Intelligence*, *144*, 109979. https://doi.org/10.1016/j.engappai.2024.109979

Greenwald, D. L., Howell, S. T., Li, C., & Yimfor, E. (2024). Regulatory arbitrage or random errors? Implications of race prediction algorithms in fair lending analysis. *Journal of Financial Economics*, *157*, 103857. https://doi.org/10.1016/j.jfineco.2024.103857

Griffin, A. C., Wang, K. H., Leung, T. I., & Facelli, J. C. (2024). Recommendations to promote fairness and inclusion in biomedical AI research and clinical use. *Journal of Biomedical Informatics*, *157*, 104693. https://doi.org/10.1016/j.jbi.2024.104693

Gunasekara, S., & Saarela, M. (2026). Trade-offs between fairness and performance in educational AI: Analyzing post-processing bias mitigation on the OULAD. *Information and Software Technology*, *189*, 107933. https://doi.org/10.1016/j.infsof.2025.107933

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*, 3315–3323. https://doi.org/10.48550/arXiv.1610.02413

Ho, C. K., & Asrani, S. K. (2025). Ethics, Bias, and Governance in Artificial Intelligence for Hepatology: Toward Building a Safe and Fair Future. *Journal of Clinical and Experimental Hepatology*, *15*(6), 102628. https://doi.org/10.1016/j.jceh.2025.102628

Incremona, A., Pozzi, A., Guiscardi, A., & Tessera, D. (2026). A differentiable and uncertainty-aware mutual information regularizer for bias mitigation. *Neurocomputing*, *669*, 132498. https://doi.org/10.1016/j.neucom.2025.132498

Kanungo, R. P., & Gupta, S. (2021). Financial inclusion through digitalisation of services for well-being. *Technological Forecasting and Social Change*, *167*. https://doi.org/10.1016/j.techfore.2021.120721

Kim, D., Oh, K., Lee, Y., & Woo, H. (2025). Overview of fair federated learning for fairness and privacy preservation. *Expert Systems with Applications*, *293*, 128568. https://doi.org/10.1016/j.eswa.2025.128568

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 43:1-43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. In *European Journal of Operational Research* (Vol. 297, Issue 3, pp. 1083–1094). Elsevier B.V. https://doi.org/10.1016/j.ejor.2021.06.023

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30*, 4066–4076. https://doi.org/10.48550/arXiv.1703.06856

Li, C., Cheng, D., Zhang, G., Li, Y., & Zhang, S. (2026). Toward fair graph neural networks via dual-teacher knowledge distillation. *Neural Networks*, *194*, 108184. https://doi.org/10.1016/j.neunet.2025.108184

Li, Y., Du, M., Song, R., Liu, M., & Wang, Y. (2025). BATED: Learning fair representation for Pre-trained Language Models via biased teacher-guided disentanglement. *Artificial Intelligence*, *348*, 104401. https://doi.org/10.1016/j.artint.2025.104401

Li, Y., Hu, J., Song, R., & Hu, L. (2026). Enhancing fairness in decision-making of natural language understanding systems: An intersectional bias debiasing model via information theory-based disentanglement. *Expert Systems with Applications*, *306*, 130940. https://doi.org/10.1016/j.eswa.2025.130940

Liu, M., Ning, Y., Ke, Y., Shang, Y., Chakraborty, B., Ong, M. E. H., Vaughan, R., & Liu, N. (2024). FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns*, *5*(10), 101059. https://doi.org/10.1016/j.patter.2024.101059

Liu, S., Qin, Z., Hou, X., Wang, Y., Wang, Z., & Peng, Z. (2026). Fairness-aware graph representation learning through bias disentanglement. *Information and Software Technology*, *193*, 108034. https://doi.org/10.1016/j.infsof.2026.108034

Liu, S., & Vicente, L. N. (2022). Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, *19*(3), 513–537. https://doi.org/10.1007/s10287-022-00425-z

Lu, H., Lin, Y., Li, Z., Yiu, M. L., Gao, Y., & Uddin, S. (2025). Toward fair medical advice: Addressing and mitigating bias in large language model-based healthcare applications. *Artificial Intelligence in Medicine*, *168*, 103216. https://doi.org/10.1016/j.artmed.2025.103216

Manganini, C., Anna Corsi, E., & Primiero, G. (2026). Data speak but sometimes lie: A game-theoretic approach to data bias and algorithmic fairness. *International Journal of Approximate Reasoning*, *190*, 109608. https://doi.org/10.1016/j.ijar.2025.109608

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115. https://doi.org/10.1145/3457607

Mersha, T., Correa-Agudelo, E., & Gautam, Y. (2025). Fairness in precision medicine: Evaluating algorithmic bias in length of stay and readmission for asthma in the All of Us Research Program. *Journal of Allergy and Clinical Immunology*, *155*(2), AB299. https://doi.org/10.1016/j.jaci.2024.12.921

Miazek, K., & Bocian, K. (2025). When AI is fairer than humans: The role of egocentrism in moral and fairness judgments of AI and human decisions. *Computers in Human Behavior Reports*, *19*, 100719. https://doi.org/10.1016/j.chbr.2025.100719

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & others. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pessach, D., & Shmueli, E. (2021). Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings. *Expert Systems with Applications*, *185*, 115667. https://doi.org/10.1016/j.eswa.2021.115667

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys*, *55*(3), 51. https://doi.org/10.1145/3494672

Pham, N., Do, M. K., Dai, T. V., Hung, P. N., & Nguyen-Duc, A. (2025). FAIREDU: A multiple regression-based method for enhancing fairness in machine learning models for educational applications. *Expert Systems with Applications*, *269*, 126219. https://doi.org/10.1016/j.eswa.2024.126219

Qin, Q., Djian, B., Merlo, E., Li, H., & Gambs, S. (2025). Representation-based fairness evaluation and bias correction robustness assessment in neural networks. *Information and Software Technology*, *188*, 107876. https://doi.org/10.1016/j.infsof.2025.107876

Qin, Q., & Merlo, E. (2026). Prune bias from the root: Bias removal and fairness estimation by pruning sensitive attributes in pre-trained DNN models. *Information and Software Technology*, *189*, 107906. https://doi.org/10.1016/j.infsof.2025.107906

Radaideh, M. I., Kwon, O. H., & Radaideh, M. I. (2025). Fairness and social bias quantification in Large Language Models for sentiment analysis. *Knowledge-Based Systems*, *319*, 113569. https://doi.org/10.1016/j.knosys.2025.113569

Rai, A., Tian, J., & Xue, L. (2026). FAIR: A design theory for artificial intelligence fairness. *MIS Quarterly*. https://doi.org/10.25300/MISQ/2026/17971

Saiyed, S., & Kumar, V. (2025). Green AI hiring and ethical risks: Investigating the impact of algorithmic bias on sustainable human capital development and SDGs compliance. *Strategic Business Research*, *1*, 100018. https://doi.org/10.1016/j.sbr.2025.100018

Sivarajkumar, S., Huang, Y., & Wang, Y. (2023). Fair patient model: Mitigating bias in the patient representation learned from the electronic health records. *Journal of Biomedical Informatics*, *148*, 104544. https://doi.org/10.1016/j.jbi.2023.104544

Song, Y., Li, C., Xing, W., Lyu, B., & Zhu, W. (2025). Investigating perceived fairness of AI prediction system for math learning: A mixed-methods study with college students. *The Internet and Higher Education*, *65*, 101000. https://doi.org/10.1016/j.iheduc.2025.101000

Tawakuli, A., & Engel, T. (2025). Make your data fair: A survey of data preprocessing techniques that address biases in data towards fair AI. *Journal of Engineering Research*, *13*(3), 2362–2369. https://doi.org/10.1016/j.jer.2024.06.016

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*(3), 356–367. https://doi.org/10.1177/1534484305278283

Uddin, S., Liang, H., & Guo, H. (2026). Gender-based data bias and model fairness evaluation in benchmarked open-access disease prediction datasets. *Computers in Biology and Medicine*, *203*, 111503. https://doi.org/10.1016/j.compbiomed.2026.111503

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness (FairWare '18)*, 1–7. https://doi.org/10.1145/3194770.3194776

Voria, G., Scala, B., Todisco, L., Venditto, C., Giordano, G., Catolino, G., & Palomba, F. (2026). Fair and square? Evaluating fairness of LLM-generated synthetic datasets. *Information and Software Technology*, *191*, 107980. https://doi.org/10.1016/j.infsof.2025.107980

Voria, G., Sellitto, G., Ferrara, C., Abate, F., De Lucia, A., Ferrucci, F., Catolino, G., & Palomba, F. (2025). Fairness-aware practices from developers' perspective: A survey. *Information and Software Technology*, *182*, 107710. https://doi.org/10.1016/j.infsof.2025.107710

Wang, K., Mihunov, V., Lam, N. S. N., & Sun, M. (2025). Mining socioenvironmental drivers of bias in social media rescue data for fairness-aware modeling: A case study of Hurricane Harvey. *International Journal of Disaster Risk Reduction*, *128*, 105717. https://doi.org/10.1016/j.ijdrr.2025.105717

Wang, X., Wang, Z., Dolata, M., & Nunamaker, J. F. (2024). How credibility assessment technologies affect decision fairness in evidence-based investigations: A Bayesian perspective. *Decision Support Systems*, *187*, 114326. https://doi.org/10.1016/j.dss.2024.114326

Wang, Y., Wang, L., Zhou, Z., Laurentiev, J., Lakin, J. R., Zhou, L., & Hong, P. (2024). Assessing fairness in machine learning models: A study of racial bias using matched counterparts in mortality prediction for patients with chronic diseases. *Journal of Biomedical Informatics*, *156*, 104677. https://doi.org/10.1016/j.jbi.2024.104677

Wang, Z., Sun, H., & Zhao, Z. (2026). Fed-e3: A tri-level federated fairness evaluation framework with statistical guarantees and privacy preservation. *Information Fusion*, *128*, 103949. https://doi.org/10.1016/j.inffus.2025.103949

Wang, Z., Yan, X., & Jin, Y. (2026). Fairness-oriented vertical federated GNNs with incomplete sensitive attributes. *Information Fusion*, *127*, 103871. https://doi.org/10.1016/j.inffus.2025.103871

Yang, Y., Lin, M., Zhao, H., Peng, Y., Huang, F., & Lu, Z. (2024). A survey of recent methods for addressing AI fairness and bias in biomedicine. *Journal of Biomedical Informatics*, *154*, 104646. https://doi.org/10.1016/j.jbi.2024.104646

Zhen, Y., Zhu, Y., Li, J., & Chen, L. (2025). FairDLA: Improving the fairness-utility trade-off in graph neural networks via dual-level alignment. *Knowledge-Based Systems*, *324*, 113768. https://doi.org/10.1016/j.knosys.2025.113768