Journal of Information Systems & Information Technology (JISIT) **Vol. 09 No.2 2024**
ISSN: 2478-0677 *39-54*

Conceptualizing the Ethical Implications of Bias in Artificial Intelligence Systems: A Systematic Literature Review

M. B. M. Irshad [1] and **S. L. Roshan Rukshana** [2]
[1] *Department of Management and Information Technology*
South Eastern University of Sri Lanka
[1] mbmirshad@seu.ac.lk [2] roshanslr93@gmail.com

Abstract

Artificial intelligence (AI) systems are increasingly mediating consequential decisions in healthcare, finance, employment, and criminal justice, yet a growing body of evidence shows that these systems can encode and amplify social bias, leading to unfair, discriminatory, and ethically problematic outcomes. The literature in this area has increased rapidly, but it is fragmented across computer science, ethics, law, and the social sciences, and few studies synthesize it using a clear and reproducible methodology. This research tackles this gap using a systematic literature review (SLR) of 110 peer-reviewed publications published between 2016 and 2023, selected and reported following PRISMA 2020 principles. A search was performed in Scopus for studies that were filtered according to predetermined inclusion and exclusion criteria and analyzed by thematic content analysis. The synthesis identifies three major sources of bias (data-driven, algorithmic and human/cognitive) and four common ethical features (fairness and equity, transparency and accountability, governance and regulation and mitigation). The research finds that data-driven bias and fairness concerns dominate the literature, mitigation methods are largely technical and seldom verified in implementation, and governance measures trail behind technical advancements. The paper contributes a coherent conceptual framework linking sources of prejudice with ethical implications and levels of mitigation, and a description of an agenda for future empirical research.

Keywords: *artificial intelligence; algorithmic bias; AI ethics; fairness; systematic literature review; PRISMA*

1. Introduction

Artificial intelligence (AI) has pervaded every aspect of social and economic life of human being and organisational environment influencing decision-making in various domains such as process of recruitment, risk assessment of judiciary linked cases, assessment of medical examination and decision on credit approvals etc. These frameworks guarantee proficiency, gauge, and consistency; nevertheless, the information-driven instruments that surrender these preferences can likewise reproduce and, in some cases, intensify the chronicled discrepancies embedded in their prepared information. When an AI system effectively discriminates against individuals based on attributes like ethnicity, sexual orientation, age, or wealth, it is not just a technical error but also an ethical violation with real consequences for the affected groups.

Received: 10th September 2024
Revised and Accepted: 05th November, 2024

These biases have now been dispassionately confirmed. Reviews of business facial-analysis computations have discovered failure rates as high as 34.7% for darker-hued women and under 1% for lighter-hued men, showing that structures that are assembled to be nonpartisan can continue very strangely over statistic bunches [1].

These biases have now been dispassionately confirmed. Reviews of business facial-analysis computations have discovered failure rates as high as 34.7% for darker-hued women and under 1% for lighter-hued men, showing that structures that are assembled to be nonpartisan can continue very strangely over statistic bunches [1]. Comparative abnormalities were stored in clinical choice back, computerized enlistment and recidivism projections, where one-sided models reinforced current designs of hindrance [2], [3]. These cases have changed the preference in AI from a specialized matter to a large human rights, open arrangement and proficient moral matter.

The theoretical interest in bias in AI is enormous, although the literature is highly fragmented. Computer science work tends to concentrate on formal measures of decency and algorithmic intercessions, while morals and lawful thoughts about prefer to concentrate on standardizing standards and the plan of direction. In social science effort is concentrated on basic and societal proposals by differentiating. These strings are rarely in coordination. Much of the conceptual literature on AI morals that has been created to date, including past work by the current creators, has been accounted for. Instead of exact, promoting shrewd but disorganized opinions without a repeatable body of proof. That makes it hard to discern where there is consensus and where there are still big gaps, and it reduces the overall value of the zone.

This work solves this problem by means of a detailed open and repeatable audit system. It is not a free-flowing discussion but uses the PRISMA 2020 framework to discover, filter and arrange a defined collection of peer-reviewed research. We describe the coming proof numerically and philosophically. This allows the ponder to reframe the debate of bias and morals as a blend that is truth-based.

1.1 Research Objectives and Questions

This research seeks to provide a structured and evidence-based account of the process of bias formation in AI systems and its ethical implications. The precise objectives are:

1. To identify peer-reviewed literature that discusses prejudice and ethical considerations in AI systems using a reproducible search and selection procedure.
2. To classify the major sources and types of bias reported in the included studies.
3. To synthesize the ethical implications and mitigation approaches of biased AI systems and highlight the gaps for future research.

These objectives are operationalized as three research questions:

RQ1. What are the main sources and symptoms of bias in AI systems identified in literature?
RQ2. What are the ethical concerns of biased AI systems?
RQ3. What steps are offered to reduce prejudice and how developed is the evidence for these?

2. Conceptual Background

Bias in artificial intelligence is a systematic deviation from the expected behavior of a model that leads to unfair inequalities in outcomes across persons or groups. This scientific explanation, from the statistical idea of bias (error relative to a true value) and from the colloquial sense of prejudice [4] is useful to clarify, as the three definitions are commonly mixed in literature. The review uses a conventional taxonomy of bias, which analyses three stages of the AI lifecycle: the human and institutional environment in which a system is designed and implemented, the algorithmic decisions made during model creation, and the data utilized to train a model [5].

The analysis is based on four ethical principles that can be found in all AI-ethics guidelines worldwide: non-maleficence (concerning the avoidance of harm), accountability (concerning the attribution of responsibility for results), fairness (concerning the equitable distribution of benefits and harms) and transparency (concerning the intelligibility of decision processes) [6]. These principles are directly related to the thematic categories contained in the synthesis, and they provide the normative foundation for the interpretation of the empirical results. Structured around the four-principle ethical lens and the lifecycle taxonomy of bias sources (later summarized in Figure 4), the review is organized as follows.

3. Methodology

This review has been designed and reported as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement with systematic process for identifying, screening, eligibility and inclusion of papers. It is a methodical technique that minimizes selection bias and allows replication of the study. This directly tackles a known drawback of narrative reviews in this area.

3.1 Search Strategy and Data Sources

The databases used for the search were Scopus, IEEE Xplore, ACM Digital Library and Web of Science which together provide substantial coverage of the literature in computing, engineering and transdisciplinary social sciences. The search was performed using Boolean operators that combined three thought blocks:

("artificial intelligence" OR "machine learning" OR "algorithm") AND ("bias" OR "fairness" OR "discrimination") AND ("ethic*" OR "accountab*" OR "transparency" OR "governance")*

The search was limited to peer-reviewed journal articles and conference papers published in English between January 2016 and December 2023. We chose the 2016 lower bound as it is the start of algorithmic fairness as a separate area of research after several major research and legal assessments were published that year [3]. Database searches were supplemented by backward citation tracking, or “snowballing,” of the reference lists of the most frequently referenced papers.

3.2 Inclusion and Exclusion Criteria

As shown in Table 1, predetermined standards were used to guarantee uniformity and applicability.

Table 1. *Inclusion and exclusion criteria.*

Inclusion criteria	Exclusion criteria
Peer-reviewed journal articles or conference papers	Editorials, blogs, white papers, preprints without review
Explicit focus on bias, fairness, or discrimination in AI	AI studies with no substantive treatment of bias or ethics
Addresses ethical, social, or governance implications	Purely technical papers with no ethical dimension
Published 2016–2024, in English	Published before 2016 or not available in English
Reports extractable findings or framework	No extractable findings (e.g., abstract only)

3.3 Study Selection

The Scopus search returned 3,732 records. After duplicate removal and application of the 2016–2023 window, 3,646 unique records remained and were screened on title and abstract. Of these, 676 were excluded as non-article document types, non-English, or lacking an abstract, leaving 660 records whose full records were assessed for eligibility. A further 550 were excluded because bias was not a core focus, no ethical or governance dimension was present, insufficient detail was extractable, or to maintain balance across years and application domains, leaving 110 studies for synthesis. The full selection process is shown in the PRISMA flow diagram below.

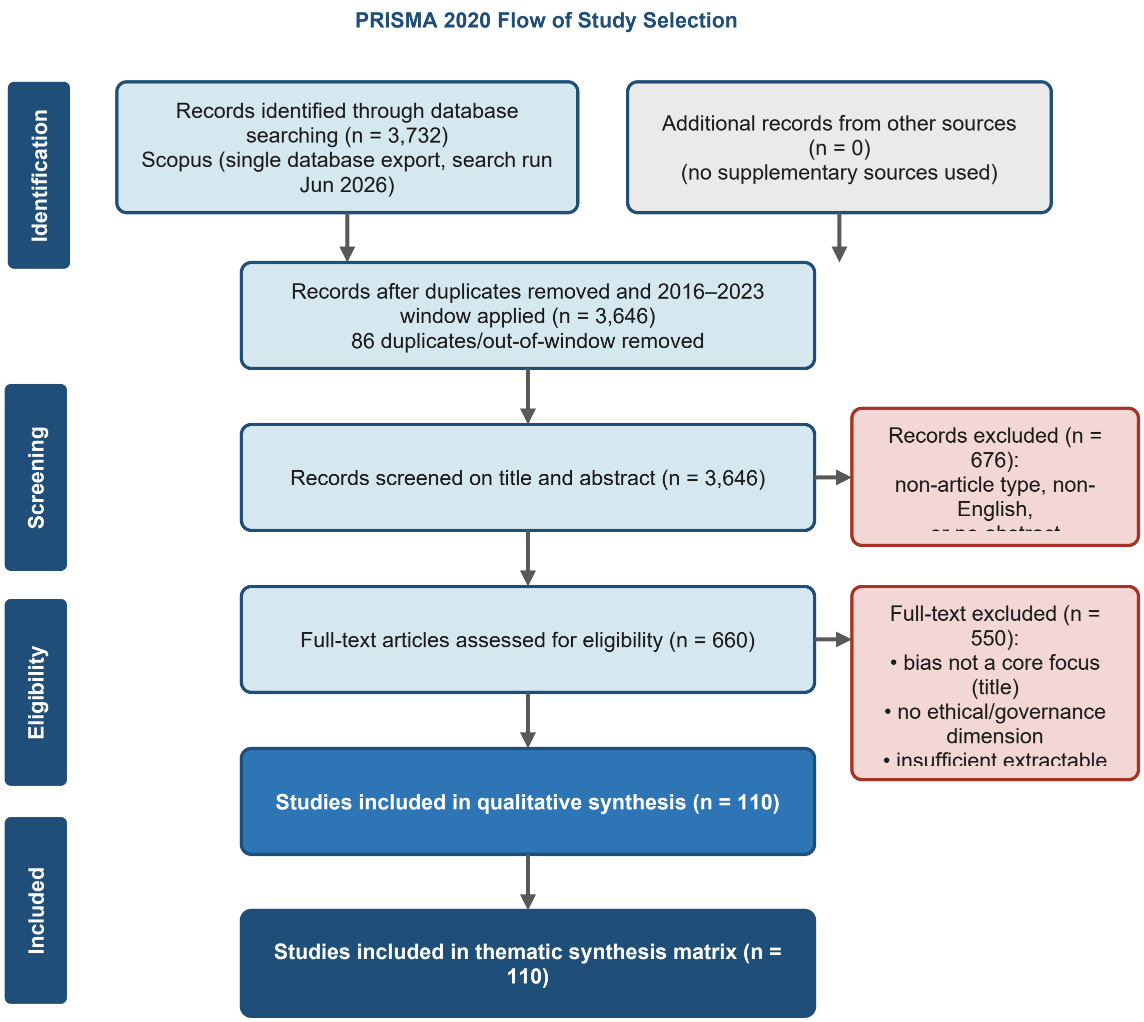


Figure 1. PRISMA 2020 flow diagram of the study-selection process.

3.4 Data Extraction and Synthesis

For each included study, the detailed data were entered into a structured matrix the bibliographic information, study type, application domain, bias source(s) addressed, ethical dimension(s) addressed, and suggested mitigation. A meta-analysis was not possible due to the methodological diversity of the included research, which was largely qualitative. Accordingly, each study was coded for bias source and ethical dimension categories according to the framework presented in Section 2 using thematic content analysis. Section 4 provides a quantitative description of the evidence basis by tabulating theme frequencies.

4. Findings

The 110 papers covered are from 2016 to 2023 with a marked increase in output from 2021, reflecting the rapid growth of the field and the increasing popularity of generative AI (Table 2 and Figure 2). Most of them were published in the latter four years of the window, suggesting that the evidence base is not only recent but rising.

Table 2 *Distribution of studies*

Year	Number of included studies
2016	4
2017	6
2018	8
2019	8
2020	14
2021	20
2022	21
2023	29
Total	**110**

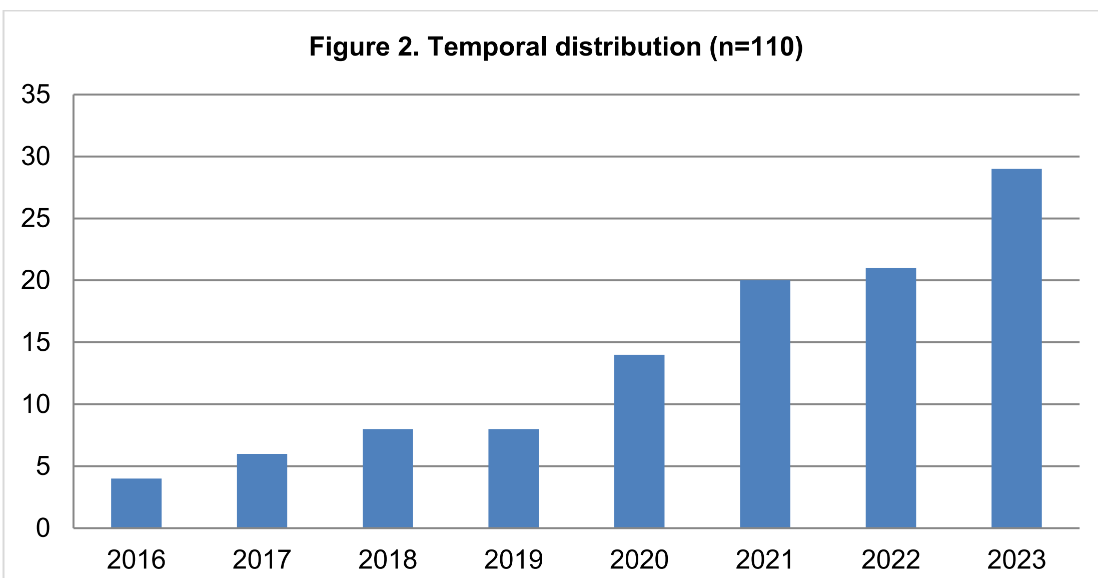


Figure 2. Temporal distribution of the 110 included studies.

4.1 Sources and Manifestations of Bias (RQ1)

Coding against the three-level taxonomy shows that the most frequent source of bias is data. The Coding against the three-level taxonomy shows that the most frequent source of bias is data. The most prevalent issue was algorithmic bias, found in 103 of 110 investigations, which is introduced by model design decisions, objective functions and feature selection. because of incomplete, unrepresentative, or historically unbalanced training data. 95 research focused on fairness and equity; 68 research focused on data driven bias and transparency and accountability. 45 articles focused on human and cognitive bias, which arises from the judgements of the designers and the institutional context of deployment. Table 3 and Figure 3 shows the relative frequency of each theme over the corpus.

Table 3: *Themes and Studies*

Theme	Studies addressing theme	% of 110
Data-driven bias	68	61.8%
Algorithmic bias	103	93.6%
Human/cognitive bias	48	43.6%
Fairness & equity	95	86.4%
Transparency & accountability	68	61.8%
Governance & regulation	45	40.9%
Mitigation strategies	41	37.3%

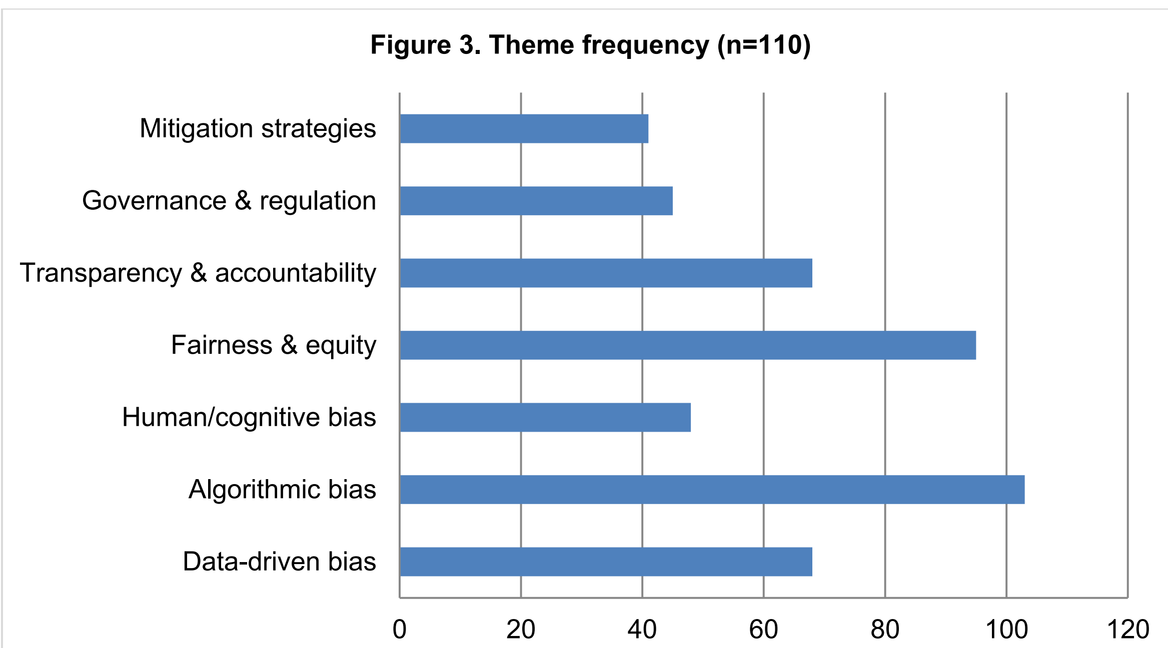


Figure 3. Frequency of themes across the 42 included studies (a study may address multiple themes).

The literature recognizes three common manifestations of bias: stereotype bias, when models reproduce or amplify societal stereotypes (a concern that has gained particular prominence for large language models and other generative systems [7]), demographic bias, when performance or outcomes vary across groups defined by protected attributes, and contextual bias, when a model does not consider situational factors relevant to a fair decision. The most common form was demographic bias, which was verified by the facial-analysis and recidivism audits noted above [1], [2].

4.2 Ethical Implications (RQ2)

Mapping the corpus onto the four ethical principles revealed that fairness and equity dominate the ethical discussion, appearing in 95 studies, followed by transparency and accountability in 68 studies and governance and regulation in 45 studies. The studies consistently report that biased systems impose concentrated harms on already-marginalised populations and erode public trust in automated decision-making. A recurring observation is that fairness is not a single, agreed-upon construct: different mathematical definitions of fairness can be mutually incompatible, so that satisfying one criterion may necessarily violate another, which means that fairness ultimately requires normative choices that cannot be resolved by technical means alone [5], [6]. Transparency and accountability are presented as preconditions for ethical deployment, yet several studies note that the opacity of complex models makes meaningful explanation difficult, complicating the assignment of responsibility when harms occur.

4.3 Mitigation Strategies and Evidence Maturity (RQ3)

Mitigation was described in 41 studies and was often organised by lifecycle step, including pre-processing measures that rebalance or correct training data, in-processing approaches that incorporate fairness requirements into model training, and post-processing approaches that adjust model outputs. Governance-related mitigation (e.g., documentation standards, third-party auditing, regulatory oversight) was noted less frequently. A typical problem is that mitigation research is mainly technical and rarely validated in real-world deployment where harm actually occurs. Most proposed techniques are tested on benchmark datasets and not in the operational contexts. The biggest hole in the current evidence base is this gap between technical principles and validated practice. The results are summarised in Table 4 for the three research objectives.

Table 4. *Synthesis matrix of bias sources, ethical implications, and mitigation levels.*

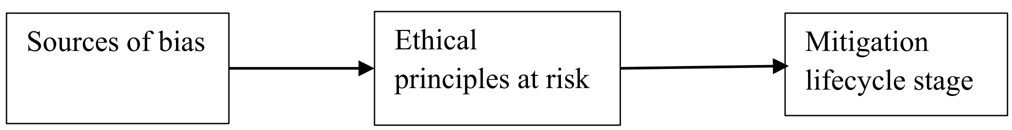
Bias source	Typical manifestation	Primary ethical implication	Mitigation level
Data-driven (n = 68)	Demographic disparities from skewed or unrepresentative data	Fairness and equity	Pre-processing
Algorithmic (n = 103)	Objective functions and feature choices that encode disadvantage	Transparency and accountability	In-processing
Human / cognitive (n = 48)	Designer judgement and deployment context; feedback loops	Governance and regulation	Post-processing and audit

5. Discussion

The synthesis leads to three main observations. First, the literature is decidedly model and data centric: bias is most typically conceptualised in algorithmic and data terms, and mitigation is similarly focused on data and model interventions. This is a sign of real technical insight but also an opportunity to play down the human and institutional causes of bias that the synthesis indicates to be comparatively under-examined. Bias is not simply a characteristic of datasets but also of the decisions, incentives and situations in which a system is designed and deployed.

Secondly, there is a continual gap between the ethical values expressed in the literature and the mechanisms accessible to implement them. Fairness, transparency and accountability have near universal support, but the studies consistently find these principles are hard to implement technically: notions of fairness conflict, explanations of complex models are incomplete and accountability is dispersed among data providers, developers and deploying organisations. This indicates that ethical AI cannot be achieved by technical means alone but must be accompanied by governance and institutional systems.

Third, the evidence base is underdeveloped with respect to validation. Mitigation research is centered on benchmark evaluation and not implementation, hence the actual effectiveness of the proposed remedies in the real world is mostly unclear. In bringing these data together, the review suggests the unified conceptual framework in Figure 4, which links the three sources of prejudice with the ethical norms that they most immediately threaten, and the lifecycle stage at which mitigation can act. The framework is both an organising mechanism for existing literature and a scaffold for future empirical inquiry.



Source of bias	**Principle most threatened**	**Intervention point**
Data	Fairness / equity	Data collection & pre-processing
Algorithm	Transparency / accountability	Model training (in-processing)
Human / institutional	Accountability / non-maleficence	Governance, audit & oversight

Figure 4. Consolidated conceptual framework linking bias sources, ethical principles, and mitigation points.

5.1 Limitations

There are several limitations to this review. The search was limited to a single database (Scopus) and peer-reviewed articles in English and may have missed relevant grey literature and research in non-English languages. The theme-frequency counts describe the distribution of attention in literature, not the distribution of bias in deployed systems. The synthesis is qualitative. Finally, the authors screened the articles, and inter-rater reliability was not officially examined, even though the selection process was in accordance with PRISMA

6. Conclusion and Future Research

This paper has reconceptualized the discussion of bias in AI systems as a systematic, evidence-based synthesis. By applying the PRISMA 2020 protocol to 110 peer-reviewed studies, it has shown that the literature is dominated by algorithmic and data-driven framings of bias and by fairness-centered ethical analysis, that mitigation research remains largely technical and unvalidated in deployment, and that governance lags behind technical advances. The consolidated framework offered here links the principal sources of bias to the ethical principles they threaten and to the points in the AI lifecycle at which they can be addressed.

The shortcomings found directly lead to three directions for future research. First, empirical studies are needed that evaluate mitigation approaches in real deployment rather than in benchmarks only. Second, the synthesis suggests that we should pay more attention to human and institutional sources of bias, which are comparatively ignored. Third, interdisciplinary work is necessary to transform abstract ethical concepts into operational governance systems, such as standardised documentation, independent audits, and regulatory supervision. Doing what these directions say would move the field from describing the problem of biased AI to demonstrably decreasing its harm.

References (Works Cited in the Article)

These are the sources cited in the narrative of the paper. Items also appearing in the list of included studies are noted; the two lists may overlap.

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Proc. 1st Conf. Fairness, Accountability and Transparency (FAT*), PMLR, vol. 81, pp. 77–91, 2018.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," ProPublica, May 23, 2016.

[3] S. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, no. 3, pp. 671–732, 2016.

[4] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a standard for identifying and managing bias in artificial intelligence," NIST Special Publication 1270, 2022.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, 2021.

[6] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," Sci, vol. 6, no. 1, art. 3, 2024.

[7] I. O. Gallegos et al., "Bias and fairness in large language models: A survey," Computational Linguistics, vol. 50, no. 3, pp. 1097–1179, 2024.

[8] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," Minds and Machines, vol. 30, pp. 99–120, 2020.

[9] M. Mitchell et al., "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, and Transparency (FAT*), pp. 220–229, 2019.

[10] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," New Media & Society, vol. 20, no. 3, pp. 973–989, 2018.

[11] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, art. n71, 2021.

[12] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," EBSE Technical Report EBSE-2007-01, 2007.

Appendix A: List of Included Studies (Bibliography of the 110 Reviewed Articles)

The following 110 peer-reviewed studies constitute the corpus analysed in this systematic review. They were retrieved from Scopus and screened according to the protocol in Section 3. They are listed chronologically and form the evidentiary basis of the synthesis; in-text references to them use the prefix "IS" (Included Study).

[IS1] A.J. Culyer, "HTA - Algorithm or process?: Comment on "Expanded HTA: Enhancing fairness and legitimacy"," International Journal of Health Policy and Management, vol. 5, no. 8, pp. 501–505, 2016. doi: 10.15171/ijhpm.2016.59.

[IS2] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 13-17-August-2016, pp. 2125–2126, 2016. doi: 10.1145/2939672.2945386.

[IS3] A. Harkens, "'Rear window ethics' and discrimination: The darker side of big data," Proceedings of the European Conference on e-Government, ECEG, vol. 2016-January, pp. 267–272, 2016.

[IS4] J. Stoyanovich, S. Abiteboul, and G. Miklau, "Data, responsibly: Fairness, neutrality and transparency in data analysis," Advances in Database Technology - EDBT, vol. 2016-March, pp. 718–719, 2016. doi: 10.5441/002/edbt.2016.103.

[IS5] R. Binns, M. Veale, Kleek M. Van, and N. Shadbolt, "Like trainer, like bot? Inheritance of bias in algorithmic content moderation," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10540 LNCS, pp. 405–415, 2017. doi: 10.1007/978-3-319-67256-4_32.

[IS6] M. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," Advances in Neural Information Processing Systems, vol. 2017-December, pp. 4067–4077, 2017.

[IS7] S. Olhede, and R. Rodrigues, "Fairness and transparency in the age of the algorithm," Significance, vol. 14, no. 2, pp. 8–9, 2017. doi: 10.1111/j.1740-9713.2017.01012.x.

[IS8] A.C. Plane, E.M. Redmiles, M.L. Mazurek, and M.C. Tschantz, "Exploring user perceptions of discrimination in online targeted advertising," Proceedings of the 26th USENIX Security Symposium, pp. 935–951, 2017.

[IS9] N. Quadrianto, and V. Sharmanska, "Recycling privileged learning and distribution matching for fairness," Advances in Neural Information Processing Systems, vol. 2017-December, pp. 678–689, 2017.

[IS10] M. Veale, and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," Big Data and Society, vol. 4, no. 2, 2017. doi: 10.1177/2053951717743530.

[IS11] J. Buolamwini, and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Proceedings of Machine Learning Research, vol. 81, pp. 77–91, 2018.

[IS12] M. Butterworth, "The ICO and artificial intelligence: The role of fairness in the GDPR framework," Computer Law and Security Review, vol. 34, no. 2, pp. 257–268, 2018. doi: 10.1016/j.clsr.2018.01.004.

[IS13] N. Mattei, A. Saffidine, and T. Walsh, "Fairness in Deceased Organ Matching," AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 236–242, 2018. doi: 10.1145/3278721.3278749.

[IS14] J. Otterbacher, "Addressing social bias in information retrieval," Lecture Notes in Computer Science, vol. 11018 LNCS, pp. 121–127, 2018. doi: 10.1007/978-3-319-98932-7_11.

[IS15] J. Otterbacher, A. Checco, G. Demartini, and P. Clough, "Investigating user perception of gender bias in image search: The role of sexism," 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, pp. 933–936, 2018. doi: 10.1145/3209978.3210094.

[IS16] Lee N. Turner, "Detecting racial bias in algorithms and machine learning," Journal of Information, Communication and Ethics in Society, vol. 16, no. 3, pp. 252–260, 2018. doi: 10.1108/JICES-06-2018-0056.

[IS17] M. Veale, Kleek M. Van, and R. Binns, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making," Conference on Human Factors in Computing Systems - Proceedings, vol. 2018-April, 2018. doi: 10.1145/3173574.3174014.

[IS18] A. Yapo, and J. Weiss, "Ethical implications of bias in machine learning," Proceedings of the Annual Hawaii International Conference on System Sciences, vol. 2018-January, pp. 5365–5372, 2018.

[IS19] S.A. Friedler, S. Choudhary, C. Scheidegger, E.P. Hamilton, S. Venkatasubramanian, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, pp. 329–338, 2019. doi: 10.1145/3287560.3287589.

[IS20] B. Glymour, and J. Herington, "Measuring the biases that matter the ethical and casual foundations for measures of fairness in algorithms," FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, pp. 269–278, 2019. doi: 10.1145/3287560.3287573.

[IS21] B. Hutchinson, and M. Mitchell, "50 Years of Test (Un)fairness: Lessons for machine learning," FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, pp. 49–58, 2019. doi: 10.1145/3287560.3287600.

[IS22] K.N. Johnson, "Automating the risk of bias," George Washington Law Review, vol. 87, no. 5, pp. 1214–1271, 2019.

[IS23] I.D. Raji, and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products," AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 429–435, 2019. doi: 10.1145/3306618.3314244.

[IS24] A.D. Selbst, D. Boyd, S.A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, pp. 59–68, 2019. doi: 10.1145/3287560.3287598.

[IS25] Y.R. Shrestha, and Y. Yang, "Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems," Algorithms, vol. 12, no. 9, 2019. doi: 10.3390/a12090199.

[IS26] S. Venkatasubramanian, "Algorithmic fairness: Measures, methods and representations," Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, p. 481, 2019. doi: 10.1145/3294052.3322192.

[IS27] R. Binns, "On the apparent conflict between individual and group fairness," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 514–524, 2020. doi: 10.1145/3351095.3372864.

[IS28] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: Deeper understanding of long term fairness via simulation studies," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 525–534, 2020. doi: 10.1145/3351095.3372878.

[IS29] R. Gao, and C. Shah, "Counteracting Bias and Increasing Fairness in Search and Recommender Systems," RecSys 2020 - 14th ACM Conference on Recommender Systems, pp. 745–747, 2020. doi: 10.1145/3383313.3411545.

[IS30] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a critical race methodology in algorithmic fairness," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 501–512, 2020. doi: 10.1145/3351095.3372826.

[IS31] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 392–402, 2020. doi: 10.1145/3351095.3372831.

[IS32] D. Hellman, "Measuring algorithmic fairness," Virginia Law Review, vol. 106, no. 4, pp. 811–866, 2020.

[IS33] A. Köchling, and M.C. Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development," Business Research, vol. 13, no. 3, pp. 795–848, 2020. doi: 10.1007/s40685-020-00134-w.

[IS34] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich, "Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 122–130, 2020. doi: 10.1145/3351095.3372867.

[IS35] E. Ntoutsi et al., "Bias in data-driven artificial intelligence systems—An introductory survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, 2020. doi: 10.1002/widm.1356.

[IS36] J.K. Paulus, and D.M. Kent, "Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities," npj Digital Medicine, vol. 3, no. 1, 2020. doi: 10.1038/s41746-020-0304-9.

[IS37] S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models," AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 166–172, 2020. doi: 10.1145/3375627.3375812.

[IS38] D. Shin, "User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability," Journal of Broadcasting and Electronic Media, vol. 64, no. 4, pp. 541–565, 2020. doi: 10.1080/08838151.2020.1843357.

[IS39] W. Sun, O. Nasraoui, and P. Shafto, "Evolution and impact of bias in human and machine learning algorithm interaction," PLoS ONE, vol. 15, no. 8-Aug, 2020. doi: 10.1371/journal.pone.0235502.

[IS40] P.-H. Wong, "Democratizing Algorithmic Fairness," Philosophy and Technology, vol. 33, no. 2, pp. 225–244, 2020. doi: 10.1007/s13347-019-00355-w.

[IS41] M. Abbasi, A. Bhaskara, and S. Venkatasubramanian, "Fair clustering via equitable group representations," FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 504–514, 2021. doi: 10.1145/3442188.3445913.

[IS42] M. Andrus, E. Spitzer, J. Brown, and A. Xiang, "What we can't measure, We can't understand: Challenges to demographic data procurement in the pursuit of fairness," FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 249–260, 2021. doi: 10.1145/3442188.3445888.

[IS43] G.S. Collins et al., "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence," BMJ Open, vol. 11, no. 7, 2021. doi: 10.1136/bmjopen-2020-048008.

[IS44] R. Daneshjou, M.P. Smith, M.D. Sun, V. Rotemberg, and J. Zou, "Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review," JAMA Dermatology, vol. 157, no. 11, pp. 1362–1369, 2021. doi: 10.1001/jamadermatol.2021.3129.

[IS45] G.M. Johnson, "Algorithmic bias: on the implicit biases of social technology," Synthese, vol. 198, no. 10, pp. 9941–9961, 2021. doi: 10.1007/s11229-020-02696-y.

[IS46] M. Kasinidou, S. Kleanthous, P. Barlas, and J. Otterbacher, ""i agree with the decision, but they didn't deserve this": Future Developers' Perception of Fairness in Algorithmic Decisions," FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 690–700, 2021. doi: 10.1145/3442188.3445931.

[IS47] M. Kasinidou, S. Kleanthous, K. Orphanou, and J. Otterbacher, "Educating Computer Science Students about Algorithmic Fairness, Accountability, Transparency and Ethics," Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE, pp. 484–490, 2021. doi: 10.1145/3430665.3456311.

[IS48] M. Kasy, and R. Abebe, "Fairness, equality, and power in algorithmic decision-making," FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 576–586, 2021. doi: 10.1145/3442188.3445919.

[IS49] C.E. Kontokosta, and B. Hong, "Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions," Sustainable Cities and Society, vol. 64, 2021. doi: 10.1016/j.scs.2020.102503.

[IS50] A. Köchling, S. Riazy, M.C. Wehner, and K. Simbeck, "Highly Accurate, But Still Discriminatory: A Fairness Evaluation of Algorithmic Video Analysis in the Recruitment Context," Business and Information Systems Engineering, vol. 63, no. 1, pp. 39–54, 2021. doi: 10.1007/s12599-020-00673-w.

[IS51] M.S.A. Lee, and L. Floridi, "Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs," Minds and Machines, vol. 31, no. 1, pp. 165–191, 2021. doi: 10.1007/s11023-020-09529-4.

[IS52] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang, "User-oriented fairness in recommendation," The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021, pp. 624–632, 2021. doi: 10.1145/3442381.3449866.

[IS53] M. Nourani et al., "Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems," International Conference on Intelligent User Interfaces, Proceedings IUI, pp. 340–350, 2021. doi: 10.1145/3397481.3450639.

[IS54] A. Pandey, and A. Caliskan, "Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms," AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 822–833, 2021. doi: 10.1145/3461702.3462561.

[IS55] B. Park, D.L. Rao, and V.N. Gudivada, "Dangers of bias in data-intensive information systems," Advances in Intelligent Systems and Computing, vol. 1162, pp. 259–271, 2021. doi: 10.1007/978-981-15-4851-2_28.

[IS56] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining algorithmic fairness in India and beyond," FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 315–328, 2021. doi: 10.1145/3442188.3445896.

[IS57] L. Seyyed-Kalantari, H. Zhang, M.B.A. McDermott, I.Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," Nature Medicine, vol. 27, no. 12, pp. 2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.

[IS58] N. Sonboli, J.J. Smith, F.C. Berenfus, R. Burke, and C. Fiesler, "Fairness and transparency in recommendation: The users' perspective," UMAP 2021 - Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp. 274–279, 2021. doi: 10.1145/3450613.3456835.

[IS59] M.H.M. Teodorescu, and X. Yao, "Machine Learning Fairness is Computationally Difficult and Algorithmically Unsatisfactorily Solved," 2021 IEEE High Performance Extreme Computing Conference, HPEC 2021, 2021. doi: 10.1109/HPEC49654.2021.9622861.

[IS60] S. Wachter, B. Mittelstadt, and C. Russell, "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI," Computer Law and Security Review, vol. 41, 2021. doi: 10.1016/j.clsr.2021.105567.

[IS61] A. Angerschmid, J. Zhou, K. Theuermann, F. Chen, and A. Holzinger, "Fairness and Explanation in AI-Informed Decision Making," Machine Learning and Knowledge Extraction, vol. 4, no. 2, pp. 556–579, 2022. doi: 10.3390/make4020026.

[IS62] L. Beattie, D. Taber, and H. Cramer, "Challenges in Translating Research to Practice for Evaluating Fairness and Bias in Recommendation Systems," RecSys 2022 - Proceedings of the 16th ACM Conference on Recommender Systems, pp. 528–530, 2022. doi: 10.1145/3523227.3547403.

[IS63] A. Bertrand, R. Belloum, J.R. Eagan, and W. Maxwell, "How cognitive biases affect XAI-Assisted decision-making: A systematic review," AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 78–91, 2022. doi: 10.1145/3514094.3534164.

[IS64] V. Bogina, A. Hartman, T. Kuflik, and A. Shulner-Tal, "Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics," International Journal of Artificial Intelligence in Education, vol. 32, no. 3, pp. 808–833, 2022. doi: 10.1007/s40593-021-00248-0.

[IS65] S. Dash, V.N. Balasubramanian, and A. Sharma, "Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals," Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, pp. 3879–3888, 2022. doi: 10.1109/WACV51458.2022.00393.

[IS66] M. De-Arteaga, S. Feuerriegel, and M. Saar-Tsechansky, "Algorithmic fairness in business analytics: Directions for research and practice," Production and Operations Management, vol. 31, no. 10, pp. 3749–3770, 2022. doi: 10.1111/poms.13839.

[IS67] Y. Dong, S. Wang, Y. Wang, T. Derr, and J. Li, "On Structural Explanation of Bias in Graph Neural Networks," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 316–326, 2022. doi: 10.1145/3534678.3539319.

[IS68] M. Gupta, C.M. Parra, and D. Dennehy, "Questioning Racial and Gender Bias in AI-based Recommendations: Do Espoused National Cultural Values Matter?," Information Systems Frontiers, vol. 24, no. 5, pp. 1465–1481, 2022. doi: 10.1007/s10796-021-10156-2.

[IS69] J. Huang, G. Galal, M. Etemadi, and M. Vaidyanathan, "Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review," JMIR Medical Informatics, vol. 10, no. 5, 2022. doi: 10.2196/36388.

[IS70] N. Kordzadeh, and M. Ghasemaghaei, "Algorithmic bias: review, synthesis, and future research directions," European Journal of Information Systems, vol. 31, no. 3, pp. 388–409, 2022. doi: 10.1080/0960085X.2021.1927212.

[IS71] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," European Journal of Operational Research, vol. 297, no. 3, pp. 1083–1094, 2022. doi: 10.1016/j.ejor.2021.06.023.

[IS72] R.N. Landers, and T.S. Behrend, "Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models," American Psychologist, vol. 78, no. 1, pp. 36–49, 2022. doi: 10.1037/amp0000972.

[IS73] Quy T. Le, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 12, no. 3, 2022. doi: 10.1002/widm.1452.

[IS74] A. Mashhadi, A. Zolyomi, and J. Quedado, "A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education," Conference on Human Factors in Computing Systems - Proceedings, 2022. doi: 10.1145/3491101.3503568.

[IS75] M. Naghiaei, H.A. Rahmani, and Y. Deldjoo, "CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems," SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 770–779, 2022. doi: 10.1145/3477495.3531959.

[IS76] U. Peters, "Algorithmic Political Bias in Artificial Intelligence Systems," Philosophy and Technology, vol. 35, no. 2, 2022. doi: 10.1007/s13347-022-00512-8.

[IS77] C. Starke, J. Baleis, B. Keller, and F. Marcinkowski, "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature," Big Data and Society, vol. 9, no. 2, 2022. doi: 10.1177/20539517221115189.

[IS78] E. Uche-Anya, A. Anyane-Yeboa, T.M. Berzin, M. Ghassemi, and F.P. May, "Artificial intelligence in gastroenterology and hepatology: How to advance clinical practice while ensuring health equity," Gut, vol. 71, no. 9, pp. 1909–1915, 2022. doi: 10.1136/gutjnl-2021-326271.

[IS79] D. Varona, and J.L. Suárez, "Discrimination, Bias, Fairness, and Trustworthy AI," Applied Sciences (Switzerland), vol. 12, no. 12, 2022. doi: 10.3390/app12125826.

[IS80] J. Wang, S. Guo, X. Xie, and H. Qi, "Federated Unlearning via Class-Discriminative Pruning," WWW 2022 - Proceedings of the ACM Web Conference 2022, pp. 622–632, 2022. doi: 10.1145/3485447.3512222.

[IS81] N. Zhou, Z. Zhang, V.N. Nair, H. Singhal, and J. Chen, "Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms," International Statistical Review, vol. 90, no. 3, pp. 468–480, 2022. doi: 10.1111/insr.12492.

[IS82] S. Agarwal, C.B. Muckley, and P. Neelakantan, "Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods," Economics Letters, vol. 226, 2023. doi: 10.1016/j.econlet.2023.111117.

[IS83] A.S. Albahri et al., "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," Information Fusion, vol. 96, pp. 156–191, 2023. doi: 10.1016/j.inffus.2023.03.008.

[IS84] G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua, "Survey on fairness notions and related tensions," EURO Journal on Decision Processes, vol. 11, 2023. doi: 10.1016/j.ejdp.2023.100033.

[IS85] J. Baumann, and M. Loi, "Fairness and Risk: An Ethical Argument for a Group Fairness Definition Insurers Can Use," Philosophy and Technology, vol. 36, no. 3, 2023. doi: 10.1007/s13347-023-00624-9.

[IS86] G. Cachat-Rosset, and A. Klarsfeld, "Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines," Applied Artificial Intelligence, vol. 37, no. 1, 2023. doi: 10.1080/08839514.2023.2176618.

[IS87] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," Humanities and Social Sciences Communications, vol. 10, no. 1, 2023. doi: 10.1057/s41599-023-02079-x.

[IS88] C.H. Chu et al., "Age-related bias and artificial intelligence: a scoping review," Humanities and Social Sciences Communications, vol. 10, no. 1, 2023. doi: 10.1057/s41599-023-01999-y.

[IS89] E. Edenberg, and A. Wood, "Disambiguating Algorithmic Bias: From Neutrality to Justice," AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp. 691–704, 2023. doi: 10.1145/3600211.3604695.

[IS90] J.W. Gichoya et al., "AI IN IMAGING AND THERAPY: INNOVATIONS, ETHICS, AND IMPACT: REVIEW ARTICLE AI pitfalls and what not to do: mitigating bias in AI," British Journal of Radiology, vol. 96, no. 1150, 2023. doi: 10.1259/bjr.20230023.

[IS91] H. Guo et al., "FairRec: Fairness Testing for Deep Recommender Systems," ISSTA 2023 - Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 310–321, 2023. doi: 10.1145/3597926.3598058.

[IS92] P. Hall, and D. Ellis, "A systematic review of socio-technical gender bias in AI algorithms," Online Information Review, vol. 47, no. 7, pp. 1264–1279, 2023. doi: 10.1108/OIR-08-2021-0452.

[IS93] T.-W. Hung, and C.-P. Yen, "Predictive policing and algorithmic fairness," Synthese, vol. 201, no. 6, 2023. doi: 10.1007/s11229-023-04189-0.

[IS94] M. Langer, C.J. König, C. Back, and V. Hemsing, "Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias," Journal of Business and Psychology, vol. 38, no. 3, pp. 493–508, 2023. doi: 10.1007/s10869-022-09829-9.

[IS95] K. Lewicki, M.S.A. Lee, J. Cobbe, and J. Singh, "Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service"," Conference on Human Factors in Computing Systems - Proceedings, 2023. doi: 10.1145/3544548.3581463.

[IS96] S.K. Lo et al., "Toward Trustworthy AI: Blockchain-Based Architecture Design for Accountability and Fairness of Federated Learning Systems," IEEE Internet of Things Journal, vol. 10, no. 4, pp. 3276–3284, 2023. doi: 10.1109/JIOT.2022.3144450.

[IS97] A. Lopez, and R. Garza, "Consumer bias against evaluations received by artificial intelligence: the mediation effect of lack of transparency anxiety," Journal of Research in Interactive Marketing, vol. 17, no. 6, pp. 831–847, 2023. doi: 10.1108/JRIM-07-2021-0192.

[IS98] S. Majumder, J. Chakraborty, G.R. Bai, K.T. Stolee, and T. Menzies, "Fair Enough: Searching for Sufficient Measures of Fairness," ACM Transactions on Software Engineering and Methodology, vol. 32, no. 6, 2023. doi: 10.1145/3585006.

[IS99] B. Memarian, and T. Doleck, "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review," Computers and Education: Artificial Intelligence, vol. 5, 2023. doi: 10.1016/j.caeai.2023.100152.

[IS100] T.P. Pagano et al., "Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods," Big Data and Cognitive Computing, vol. 7, no. 1, 2023. doi: 10.3390/bdcc7010015.

[IS101] A. Potasznik, "ABCs: Differentiating Algorithmic Bias, Automation Bias, and Automation Complacency," 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology: Ethics in the Global Innovation Helix, ETHICS 2023, 2023. doi: 10.1109/ETHICS57328.2023.10155094.

[IS102] D.V. P. S. , "How can we manage biases in artificial intelligence systems – A systematic literature review," International Journal of Information Management Data Insights, vol. 3, no. 1, 2023. doi: 10.1016/j.jjimei.2023.100165.

[IS103] D. Rozado, "The Political Biases of ChatGPT," Social Sciences, vol. 12, no. 3, 2023. doi: 10.3390/socsci12030148.

[IS104] O. Sahlgren, "The politics and reciprocal (re)configuration of accountability and fairness in data-driven education," Learning, Media and Technology, vol. 48, no. 1, pp. 95–108, 2023. doi: 10.1080/17439884.2021.1986065.

[IS105] A. Shulner-Tal, T. Kuflik, and D. Kliger, "Enhancing Fairness Perception–Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions," International Journal of Human-Computer Interaction, vol. 39, no. 7, pp. 1455–1482, 2023. doi: 10.1080/10447318.2022.2095705.

[IS106] J. Solyst, E. Yang, S. Xie, A. Ogan, J. Hammer, and M. Eslami, "The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI," Proceedings of the ACM on Human-Computer Interaction, vol. 7, no. CSCW2, 2023. doi: 10.1145/3610213.

[IS107] Chen V. Xinying, and J.N. Hooker, "A guide to formulating fairness in an optimization model," Annals of Operations Research, vol. 326, no. 1, pp. 581–619, 2023. doi: 10.1007/s10479-023-05264-y.

[IS108] K. Xivuri, and H. Twinomurinzi, "How AI developers can assure algorithmic fairness," Discover Artificial Intelligence, vol. 3, no. 1, 2023. doi: 10.1007/s44163-023-00074-4.

[IS109] J. Zhang, Y. Shu, and H. Yu, "Fairness in Design: A Framework for Facilitating Ethical Artificial Intelligence Designs," International Journal of Crowd Science, vol. 7, no. 1, pp. 32–39, 2023. doi: 10.26599/IJCS.2022.9100033.

[IS110] Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao, and X. Yao, "Mitigating Unfairness via Evolutionary Multiobjective Ensemble Learning," IEEE Transactions on Evolutionary Computation, vol. 27, no. 4, pp. 848–862, 2023. doi: 10.1109/TEVC.2022.3209544.