American Scholarly Journal for Scientific Research (ASJSR)
https://asjsr.org/ Volume 2026, Issue 1004 ISSN: 3143-2999

Bias in Artificial Intelligence
A Narrative Review of AI Bias and Fairness

Author: Arjun Khanna
Institution: Shiv Nadar School, Gurgaon, India
Mentor: Geeti Gupta

ABSTRACT

Artificial intelligence can distribute benefits and errors unevenly across social groups, even when a system does not explicitly use sensitive characteristics. This paper examines how such disparities arise, what selected empirical studies establish, and how institutions can respond. A narrative review of primary research and technical guidance published between 2016 and 2025 connects evidence from facial analysis, healthcare allocation, speech recognition, education, and language models. The review distinguishes errors in prediction from stereotyping and from unequal access to opportunities. Its central argument is that fairness depends on the complete decision process: the problem selected, the data collected, the outcome predicted, the evaluation criteria, and the use of the resulting output. Statistical fairness measures reveal different aspects of this process and can conflict under specified conditions. A hypothetical numerical example illustrates why equal error rates do not necessarily produce equal predictive values. The paper proposes an audit approach combining valid outcome measures, subgroup evaluation, culturally appropriate testing, documentation, meaningful review, and continuing monitoring. These recommendations are a synthesis of the literature rather than results of a new experiment. Responsible AI requires institutions to explain which harms they are reducing and demonstrate whether affected people actually experience improvement.

Keywords: artificial intelligence, algorithmic bias, fairness, discrimination, generative AI, accountability

1. INTRODUCTION

An automated decision can appear objective because it produces a consistent numerical answer. Consistency, however, does not establish that the answer measures the right outcome or treats people fairly. A model may reproduce an institution's earlier decisions accurately while retaining weaknesses in those decisions. Evaluating AI therefore requires attention to the relationship between prediction and purpose: what is being estimated, why that estimate matters, and what happens when someone acts on it.

Bias in AI is often attributed to unrepresentative training data. That explanation captures only part of the problem. Schwartz et al. (2022) distinguish systemic, statistical or computational, and human sources of bias, placing technical systems within their institutional environment. This broader view helps explain how harmful outcomes can occur without a developer deliberately programming discriminatory rules. Choices about measurement, objectives, interfaces, and authority influence whose interests a system serves.

This paper uses the term harmful AI bias to describe systematic patterns in an AI system or its use that create unjustified disadvantage, misrepresentation, or unequal burdens. This is a working definition for the review, not a universal legal test. A difference in outcomes warrants investigation, but its size alone does not establish its cause or ethical acceptability. Conversely, identical aggregate outcomes do not prove that individuals were treated appropriately.

The review focuses on social and demographic bias in prediction, classification, and generated language. It addresses three questions. First, where can bias enter the development and use of AI? Second, what do documented cases reveal about its consequences and the limits of available evidence? Third, how can organizations assess and reduce harm while recognizing that fairness has several competing interpretations?

The argument developed here is that useful fairness evaluation connects technical measurements to an explicit account of the decision being made. Improving a benchmark score is valuable when it addresses a meaningful harm. It is insufficient when the underlying task is poorly defined, the benchmark excludes affected users, or the institution turns an uncertain output into an unquestionable judgment. This distinction also leaves room for AI to improve existing practices when comparisons are credible and its benefits are demonstrated.

2. REVIEW METHOD

This paper is a narrative literature review using purposive source selection. Targeted web searches, completed on 19 November 2026, located original articles, author manuscripts, conference proceedings, and official National Institute of Standards and Technology publications. Searches combined terms such as algorithmic bias, equality of opportunity, healthcare allocation, speech recognition disparities, AI writing detectors, dialect prejudice, and multilingual bias, alongside the titles of relevant papers. Publisher records and primary repositories were used to verify findings and bibliographic details.

The selected literature spans 2016 to 2025. Sources were chosen for their relevance to the research questions, identifiable methods or arguments, and ability to explain a distinct mechanism. Empirical studies supply examples of measured disparities; mathematical papers clarify fairness criteria; conceptual papers and technical guidance inform interpretation and proposed practices. Their evidential roles are kept separate. A reporting framework is not treated as experimental proof that documentation eliminates discrimination.

For each empirical example, the synthesis considers the task, population or dataset, measured outcome, principal finding, and limits on generalization. Results are compared thematically rather than combined statistically because the studies measure different quantities. A word error rate cannot be pooled meaningfully with an essay detector's false-positive rate or a language model's stereotyping score.

No new participants were recruited, models trained, surveys administered, or original datasets analyzed. The numerical example in Section 6 is constructed solely to explain fairness metrics. The review is not systematic or exhaustive: it does not report a database-wide screening process, calculate a pooled effect, or claim to represent every current AI product. Its contribution is an analytical synthesis and a proposed evaluation approach.

3. UNDERSTANDING THE SOURCES AND FORMS OF BIAS

Bias Across the Development Process

Suresh and Guttag (2021) identify seven sources of potential harm across the machine learning life cycle. Table 1 summarizes mechanisms that can coexist.

Table 1: Sources of bias across the machine learning life cycle

Source	Central Concern
Historical bias	Records preserve existing disadvantage.
Representation bias	Samples inadequately cover intended users.
Measurement bias	Features or labels misrepresent intended concepts.
Aggregation bias	A common model overlooks population differences.
Learning bias	Training choices distribute errors unevenly.
Evaluation bias	Tests overlook relevant users or conditions.
Deployment bias	Use departs from development assumptions.

Note. Summary adapted from Suresh and Guttag (2021); categories do not rank risks.

The distinction has practical consequences. Consider a hypothetical scholarship model trained to reproduce previous award decisions. Adding more records could improve its imitation of those decisions without establishing whether earlier selection recognized students with fewer opportunities. Before collecting additional data, the institution would need to decide whether its target should be prior selection, demonstrated achievement, financial need, or another clearly justified outcome. A technically improved answer to the wrong question can remain unsuitable.

Different Harms Require Different Evidence

The review uses the allocative and representational harms discussed by Suresh and Guttag (2021): unequal access to resources or opportunities, and harmful portrayals or erasure. It separately considers unequal service quality, including reliability, usability, and the effort needed to obtain assistance. These consequences can overlap. A transcription error becomes an allocation problem if an institution uses the transcript to judge an applicant.

The distinction also changes what an audit should measure. A count of offensive expressions may illuminate one aspect of a chatbot's behavior while revealing little about whether its advice is equally useful. Similarly, equal selection rates do not establish whether a system's descriptions are respectful. The proposed approach is to identify the harm first and then choose observations capable of detecting it.

An additional concern is change over time. In a hypothetical recruitment workflow, a model trained only on people previously hired cannot directly observe how rejected applicants would have performed. If its recommendations shape future hiring, later training records may reflect the model's own earlier choices. This illustrates why an institution should examine how its data are produced rather than treating a growing dataset as independent confirmation that the original decision rule was sound.

4. EMPIRICAL EVIDENCE ACROSS APPLICATION AREAS

Intersectional Disparities In Facial Analysis

Buolamwini and Gebru (2018) evaluated three commercial systems that classified facial images using binary gender labels. Their benchmark contained 1,270 people. Across systems, error rates for darker-skinned female-labelled faces ranged from 20.8% to 34.7%, compared with 0.0% to 0.8% for lighter-skinned male-labelled faces. Examining intersecting characteristics exposed disparities that an overall accuracy figure could obscure.

The study evaluated gender classification, not identification of a person's identity through face recognition. Its skin-tone categories are not equivalent to racial categories, and binary labels cannot fully describe gender identity. The reported results concern systems evaluated in 2017 and published in 2018; they do not establish the performance of current versions. The methodological lesson is to test relevant intersections and state precisely what a model was asked to predict.

An Unsuitable Proxy In Healthcare Allocation

Obermeyer et al. (2019) examined a commercial algorithm used to identify patients for additional care management. At comparable risk scores, Black patients had greater illness burdens than White patients. The algorithm predicted expenditure as a proxy for health need, although spending reflected unequal access to care. Race was excluded from its inputs.

In a simulation removing the health disparity in selection, the Black share of patients automatically identified at the 97th-percentile threshold increased from 17.7% to 46.5%. These percentages describe the composition of the selected group, not the proportion of all Black patients treated. They are a counterfactual estimate, not an observed improvement following implementation. The case demonstrates why a target that is convenient to predict can be inappropriate for the decision it is intended to support.

Unequal Speech Recognition Performance

Koenecke et al. (2020) tested five commercial speech recognition systems using interviews with 73 Black and 42 White speakers. The corpus contained 19.8 hours of audio from five United States cities, matched on speaker age and gender. Average word error rates were 0.35 for Black speakers and 0.19 for White speakers; all five systems showed disparities.

Word error rate measures substitutions, deletions, and insertions relative to the reference transcript. It does not mean that 35% of Black speakers were misunderstood. The findings concern the historical systems and corpus studied, not every accent or current service. For institutional evaluation, the case suggests examining who must repeatedly correct outputs or seek alternative access, as well as whether a system performs well on average.

AI Writing Detectors In Educational Assessment

Liang et al. (2023) assessed seven AI text detectors using 91 TOEFL essays from a Chinese forum and 88 essays by United States eighth-grade students. The journal article reports an average false-positive rate of 61.3% for the TOEFL essays. Human writing was frequently classified as AI-generated.

The modest samples differed in educational stage and writing task as well as linguistic background. The study therefore does not isolate every cause of the disparity or establish an error rate for all non-native English writers. It also does not audit every present-day detector. A defensible institutional response is to require corroborating evidence and a fair review process before treating a detector flag as proof of misconduct. That response is a recommendation derived from the reported error risk, not a further experimental finding.

What The Cases Establish Together

These studies examine different tasks and should not be ranked on a single scale of bias. Their joint value is diagnostic. A disparity may emerge from who appears in evaluation data, what the prediction target represents, whether a service understands different forms of expression, or how an institution interprets a model's output. The appropriate intervention depends on which explanation is supported. A single generic instruction to make the model fairer leaves these decisions unresolved.

5. LANGUAGE MODELS AND CULTURAL CONTEXT

Learned Associations And Consequential Outputs

Caliskan et al. (2017) showed that GloVe word embeddings learned from web text contained associations resembling established human implicit associations, including those involving race and gender. Their Word Embedding Association Test helped demonstrate that social patterns can be encoded through ordinary statistical learning. However, this was research on static word representations. Association scores neither measure a model's beliefs nor directly establish discriminatory behavior by current generative chatbots.

Hofmann et al. (2024) investigated a more consequential level of output. Using dialect comparisons, including texts matched in meaning, they found that tested language models generated more negative character assessments and disadvantageous recommendations about speakers of African American English. The experiments included hypothetical employment and criminal justice scenarios. More favorable responses to explicit racial references coexisted with negative assessments based on dialect. The findings concern model outputs under the study's conditions, rather than observed hiring, convictions, or sentencing. They show why evaluation should examine indirect identity signals as well as explicit group labels.

For an application that produces recommendations, three questions should consequently remain separate: what associations the model encodes, what advice it generates, and whether that advice changes people's opportunities. An evaluator may have strong evidence at one level and limited evidence at another. Keeping these levels distinct prevents both understatement of concerning behavior and overstatement of its demonstrated consequences.

Evidence From Indian Language Evaluation

Tomar et al. (2025) developed BharatBBQ to assess social biases in Hindi, English, Marathi, Bengali, Tamil, Telugu, Odia, and Assamese. The benchmark covers 13 social categories, including three intersectional categories. Evaluation of five multilingual model families found persistent biases, often greater in the Indian languages assessed than in English. The study did not establish the causes of those differences.

This evidence supports culturally and linguistically grounded evaluation. It does not justify assuming identical behavior across all Indian languages, institutions, or models. For a proposed educational or public-service application, the implication is to test the languages and social

contexts relevant to its users. Translation quality, local meanings, and code-switching should be considered explicitly rather than assuming that an English benchmark fully describes the intended setting.

6. MEASURING FAIRNESS AND INTERPRETING TRADEOFFS

Fairness Criteria Answer Different Questions

Fairness is not a single numerical property. Table 2 summarizes common criteria discussed in the fairness literature, particularly Hardt et al. (2016) and Chouldechova (2017). Here, a positive decision means selection or approval; in another setting it could indicate an adverse classification. The meaning must be specified before interpreting the metric.

Table 2: Common group fairness criteria

Criterion	What is compared across groups
Demographic parity	The overall rate of positive decisions.
Equal opportunity	The true-positive rate among people with a positive reference label.
Equalized odds	Both true-positive and false-positive rates.
Predictive parity	The proportion of positive decisions that match positive reference labels.
Group calibration	Within each group, cases assigned probability p have positive outcomes at rate p.

Note. Equal opportunity and equalized odds are statistical definitions; a positive reference label is not automatically a complete or unbiased measure of merit or need.

Chouldechova (2017) demonstrates that, when outcome prevalence differs between groups, predictive parity generally cannot coexist with equal false-positive and false-negative rates in nondegenerate binary classification settings. Perfect prediction and certain boundary cases require qualification. Calibration of continuous scores must also be distinguished from predictive parity after applying a decision threshold. The result does not prove that fair AI is impossible or decide which fairness goal should take priority.

A Hypothetical Numerical Illustration

Suppose a selection system assesses two groups of 100 applicants each. An independent reference assessment classifies 60 applicants in Group A and 40 in Group B as eligible. Assume that the system selects 80% of eligible applicants and 30% of ineligible applicants in each group. The constructed counts are shown in Table 3. They are not observations from an actual institution.

Table 3: ***Constructed example of conflicting fairness criteria***

Quantity	Group A	Group B
Total applicants	100	100
Eligible applicants	60	40
Eligible applicants selected	48	32
Ineligible applicants selected	12	18
Total selected	60	50
True-positive rate	80%	80%
False-positive rate	30%	30%
Positive predictive value	80%	64%

Note. All values are hypothetical. Positive predictive value is eligible applicants selected divided by total applicants selected.

The system meets equalized odds in this example, yet predictive parity differs: 48 of 60 selected applicants in Group A are eligible, compared with 32 of 50 in Group B. Overall selection rates also differ. These calculations illustrate how a system can satisfy one criterion and fail another without any arithmetic inconsistency.

The example assumes that eligibility is validly measured. In an actual application, that assumption would need investigation. Differences in recorded eligibility could reflect unequal preparation opportunities, an inappropriate assessment, or other contextual factors. A mathematical comparison cannot determine their meaning by itself. The institution must justify both the reference measure and its chosen treatment of errors.

Measurement Beyond A Single Score

This paper recommends reporting group sample sizes, error counts, uncertainty, and the practical costs of mistakes alongside fairness metrics. A small subgroup may have an apparently extreme rate based on few observations. A narrow confidence interval, meanwhile, cannot repair a biased reference label. Statistical uncertainty and measurement validity are different problems.

Audits should also compare plausible alternatives. Relevant questions include whether an improved model reduces the absolute burden of errors, whether a simpler method performs adequately, and whether review corrects mistakes consistently. The most useful evaluation explains who benefits from an intervention, whose outcomes worsen, and which consequences remain unmeasured.

7. STRATEGIES FOR REDUCING HARMFUL BIAS

Define A Defensible Purpose And Target

The first proposed step is to specify the decision and the outcome that would make it successful. For example, a student-support system might aim to identify a need for tutoring. Predicting a low grade and predicting benefit from tutoring are different tasks. A team should state which task it can support and avoid presenting one as evidence of the other.

Selbst et al. (2019) argue that technical fairness work can fail when it abstracts away institutions, affected people, and changes caused by introducing a system. This supports evaluating the complete process rather than treating the model as a self-contained solution. In this paper's proposed approach, a project may be redesigned or restricted if its purpose cannot be supported by valid measurements.

Improve Data And Examine Their Provenance

Gebru et al. (2021) propose datasheets that document a dataset's motivation, composition, collection, processing, intended uses, distribution, and maintenance. Such records make sampling choices and limitations available for inspection. Documentation supports evaluation; it does not itself make data representative or ethically acceptable.

An organization applying this approach should distinguish absence of evidence from evidence of adequate coverage. If an important population is missing, the response may require additional, appropriately obtained data or a narrower intended use. Adding synthetic examples should be

evaluated for whether it improves performance on independently collected examples from that population. A larger dataset is useful only when its additional information addresses the problem being investigated.

Auditing can require information about group membership that is itself sensitive. This paper recommends limiting collection to a defined evaluation purpose, restricting access, and setting retention periods. Where practical, demographic data used for auditing should be separated from operational decision inputs. Participation and categories should respect how people describe themselves. If missing data prevent a credible subgroup assessment, the limitation should be disclosed; an organization should not claim that unmeasured disparities have been ruled out.

Apply Technical Interventions To A Specified Objective

Hardt et al. (2016) show how a predictor can be adjusted to satisfy their statistical fairness criterion under specified information and distributional assumptions. Such techniques make fairness requirements operational. Their guarantees concern the chosen criterion, rather than every social consequence of deployment.

For an applied project, this paper recommends comparing candidate interventions on held-out data and stating what each is intended to change. Reweighting examples may address unequal influence in training; revising labels may address a measurement problem; adjusting thresholds may redistribute particular errors. These are candidate strategies, not interchangeable cures. Each requires fresh assessment of overall performance, subgroup outcomes, and operational feasibility. A technique should be retained because its effects are demonstrated in the intended setting, not merely because it is described as debiasing.

Report Model Limitations And Test Realistic Use

Mitchell et al. (2019) propose model cards that describe intended use, evaluation procedures, subgroup performance, and limitations. They help readers understand the circumstances under which a model has supporting evidence. A completed card is a reporting instrument, not a certificate that the system is fair.

For generative AI, Autio et al. (2024) recommend assessing harmful bias across groups and contexts, considering indirect signals such as dialect, and involving affected communities. Their

guidance also directs attention to application outcomes and the limitations of evaluation data. It is technical guidance rather than proof of universal effectiveness.

A practical evaluation should therefore preserve the model version, prompts, settings, and decision rules used during testing. Where outputs vary, repeated trials can show whether an apparent result is stable. Tests should include ordinary user requests and foreseeable difficult cases, with qualified reviewers assessing whether changes improve the intended behavior.

Make Review And Correction Operational

Human involvement should be designed as a process with authority, evidence, and adequate time. A reviewer who cannot question a model, inspect relevant information, or change the outcome supplies little practical protection. The proposed workflow should specify when review occurs, what the reviewer considers, and how disagreements are recorded.

Affected people should receive a usable way to raise errors and provide relevant context. For an educational application, this could include access to the material underlying a concern and an opportunity to explain the work. Institutions should evaluate whether the review process corrects errors consistently and whether obtaining a correction imposes disproportionate burdens. These are design recommendations requiring evaluation, not assumptions that adding a human automatically resolves bias.

8. INSTITUTIONAL ACCOUNTABILITY AND A PROPOSED AUDIT APPROACH

Responsibility Beyond The Developer

Raghavan et al. (2020) examined the disclosed claims and practices of vendors providing algorithmic employment assessments. Their analysis highlights consequential choices about training data, prediction targets, validation, and the meaning of bias mitigation. It supports scrutinizing how a vendor establishes its claims, while not providing a performance verdict on every hiring product.

For procurement, this review recommends asking for evidence that matches the proposed use. A school evaluating an assessment tool needs information relevant to its students, assessment tasks, and decisions. A general accuracy claim is insufficient if it does not identify the evaluation population, reference labels, error types, or product version. Responsibility should be allocated

explicitly between the provider and deploying institution so that uncertainties do not disappear between them.

The NIST AI Risk Management Framework organizes risk work around Govern, Map, Measure, and Manage (Tabassi, 2023). These functions offer a structure for assigning responsibility, understanding context, evaluating risks, and responding to findings. The framework is voluntary guidance; reference to it alone does not demonstrate legal compliance or successful mitigation.

A Practical Sequence For Evaluation

Drawing on the reviewed literature, this paper proposes the following sequence. It is a synthesis intended for adaptation and testing, not a newly validated audit standard.

1. **Specify the decision.** Record the intended users, affected people, purpose, alternatives, and consequences of incorrect outputs. Identify who can authorize, restrict, or stop use.
2. **Check the evidence base.** Examine data provenance, coverage, missingness, reference labels, and whether the development setting matches the proposed application. Treat unexplained gaps as unresolved limitations.
3. **Select relevant measures.** Explain which harms each metric detects, justify the groups and intersections examined, and identify consequences that require qualitative assessment.
4. **Evaluate the complete workflow.** Compare model outputs, human interpretation, and final decisions. Use appropriate baselines and independent evaluation data. Report uncertainty and meaningful differences, including unfavorable findings.
5. **Set conditions for use.** Define where evidence supports deployment, where additional review is necessary, and what findings would trigger restriction. Avoid universal cutoffs detached from the importance of the decision.
6. **Monitor and provide remedy.** Track complaints, corrections, subgroup outcomes, and material changes in data or software. Reassess after updates and make someone responsible for acting on deterioration.

For a hypothetical AI writing-screening tool, these steps would separate detection from disciplinary judgment. Testing would examine human writing from the actual student population, review would consider independent evidence, and the institution would monitor whether

particular groups face more false allegations or more difficult appeals. If reliable use cannot be demonstrated, restriction or discontinuation is an available outcome of the audit.

9. DISCUSSION AND LIMITATIONS

The review suggests that three questions should organize decisions about AI fairness: whether the task is justified, whether the measurement is adequate, and whether using the output improves the relevant process. These questions are connected but cannot substitute for one another. A carefully measured model may serve an unsuitable purpose; a worthwhile purpose may be undermined by weak evidence; a technically capable tool may be used through a poor decision procedure.

The analysis also avoids assuming that fairness and accuracy must always conflict. If an apparent conflict depends on a poor target or avoidable errors, redesign may improve both. If a tradeoff remains, decision makers should explain its distribution and rationale. A small gain in aggregate performance may be difficult to justify when its benefits are modest and its costs fall on people who already face substantial barriers. This is an ethical judgment that should be argued openly, rather than hidden inside an optimization setting.

Several limitations constrain the paper's conclusions. Purposive selection may favor prominent studies and omit contradictory or less visible findings. Much of the empirical evidence concerns particular United States populations or English-language tasks. BharatBBQ broadens the cultural scope but cannot represent every relevant language or social category. Model versions change, so historical results establish mechanisms and documented cases rather than a current ranking of providers.

Evidence also varies in strength and level. Controlled model-output experiments help isolate behavior but do not automatically establish changes in institutional outcomes. Observational studies can reveal important disparities while leaving causal questions unresolved. Conceptual frameworks and proposed documentation practices assist interpretation but require evaluation when implemented. The review does not quantify the comparative effectiveness or cost of different mitigation strategies.

Future work should investigate whether interventions improve affected people's outcomes over time. Promising directions include multilingual and intersectional audits, evaluation of accessible appeal procedures, comparisons of human and AI-assisted decisions, and studies of what

happens after a system changes. Researchers should report when an intervention merely shifts errors, when gains fail to transfer, and when withdrawing a tool produces a better result. Such evidence would make fairness claims more actionable and less dependent on a single benchmark.

10. CONCLUSION

Bias in AI can arise from historical records, incomplete representation, inadequate measurement, model development, evaluation, and institutional use. The reviewed cases demonstrate several distinct mechanisms, while also showing why claims must remain tied to the populations, tasks, and model versions studied. Neither overall accuracy nor the exclusion of sensitive attributes is sufficient evidence of fair treatment.

The central conclusion is that fairness requires a justified purpose, appropriate evidence, and accountable decisions throughout deployment. Organizations should examine what their systems predict, compare relevant subgroup outcomes, document uncertainty, and make correction possible. Statistical criteria are useful tools, but selecting among them requires an explicit account of the harm that matters. Progress should ultimately be assessed by whether people receive more reliable services and fairer opportunities, supported by evidence from the settings in which AI is actually used.

REFERENCES

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

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

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 3315–3323). Curran Associates.

https://proceedings.neurips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

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

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (NIST Special Publication 1270). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Article 17, pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tomar, A., Sahoo, N. R., & Bhattacharyya, P. (2025). BharatBBQ: A multilingual bias benchmark for question answering in the Indian context. *Transactions of the Association for Computational Linguistics, 13*, 1672–1692. https://doi.org/10.1162/tacl.a.55