Journal of Posthumanism
2025
Volume: 5, No: 4, pp. 448–466
ISSN: 2634-3576 (Print) | ISSN 2634-3584 (Online)
posthumanism.co.uk

DOI: https://doi.org/10.63332/joph.v5i4.1091

Recalibrating Human–Machine Relations through Bias-Aware Machine Learning: Technical Pathways to Fairness and Trust

Esam Othman[1], Rund Mahafdah[2]

Abstract

Considering the importance of artificial intelligence (AI) in decision-making processes in various fields such as health, law and finance, the concern for bias and fairness of decision making has increased. This paper presents an extensive discussion of bias-aware machine learning(Ml) such as fairness-aware modeling, detection and mitigation. The paper demonstrates aspects of fairness, different forms of algorithmic bias including intersectional bias and how biased systems impact society. The paper turns to appreciation of dentieth, Trust Dynamics, Legal and Regulatory Frameworks And in the Context of Promoting Transparency: Exploring the Role of Explainable AI (XAI). Taking into account the current advances for combatting bias, also pre-processing, in-processing, and post-processing methods, for instance, draw on examples from major domains of interest. Apart from the improvements AIs have achieved, existing challenges involve little attention to relationship among different identities, poor frameworks in place for implementation and operation in other parts of the world, inadequate abuse detection mechanisms among others. Regarding this, we present some of the research questions that focus on the notions of transparency, privacy protected fairness audits, and shared control with the aim of guiding the growth of fair, responsible, and competent AI systems.

Keywords: *Bias-Aware Machine Learning, Algorithmic Fairness, Trust in AI Systems, Intersectional Bias, Explainable AI (XAI).*

Introduction

The very broad use of AI and ML systems across various sectors is the major driving force for concern among people over whether they will enhance or enforce existing societal biases. As these technologies start driving healthcare, criminal justice, hiring, and financial services, it becomes vitally important to restore fairness and trust. The growing recognition of bias in AI applications has spawned the development of resolution strategies for these challenges under both specialized and deep learning methods. Consequently, the commercialization of these systems saw more and more attentive researchers who started to look at possible issues of concern and are now earnestly trying to address them (Mehrabi, 2019). This has seen an upsurge in machine learning algorithms in decision-critical areas like banking, healthcare, recruitment, education and criminal justice. As such, bias and fairness have been the main thrust of published work of late in the realm of machine learning. More generally, the wide-reaching appliance of ML systems in decision-critical applications, such as criminal sentencing and bank loans, has raised a general concern about the risks of unfair outcomes. In response to these concerns, researchers are working on creating algorithms with an eye on fairness and different metrics for measuring fair versus discriminatory outcomes. The most recent research has provided for more

[1] Information Systems Department. College of Computer and Information Sciences. Prince Sultan University, Riyadh, Saudi Arabia, Email: Eothman@psu.edu.sa, (Corresponding Author)
[2] Department of Computer Science. College of Computing and IT, Shaqra University, Shaqra, Saudi Arabia, Email: rundmahafdah@su.edu.sa

sophisticated practices on intersectional bias, which deals with multiple sensitive attributes like race and gender in simultaneity (Gohar, 2023). Trust dynamics in AI systems holding ethical issues remain the subject of a paucity of studies. Though insights may be drawn from trust on automated systems, which mainly concern their accuracy (as in alarm systems for monitoring tasks), regarding trust-building in applications where ethical concerns like fairness are crucial remains an ethical minefield (Langer, 2021). Instituting such trust in fairness in machine learning systems is indeed a profound socio-technical endeavor, owing to the lack of adequate processes and governance practices focused on ensuring fairness in spite of the rapidly increasing use of ML tools (Toreini, 2023). This work attempts to merge existing research on bias-aware machine learning regarding how ongoing work is trying to regain the trust and fair play in intelligent systems. This review be divided under the conceptual foundations, impact and implications of biased systems, methods for detecting and measuring bias, mitigation strategies, application-specific considerations, practical implementation challenges, and emerging research directions. If we analyze these dimensions, we expect to attain a well-rounded perspective on the existing status of research in this rapidly changing field.

Literature Review

Conceptual Foundations of Bias and Fairness in ML

Defining Bias and Fairness

To understand bias and fairness in AI comprehensively, one must analyze real-world cases where bias has been exhibited and identify the variety of sources of bias which impacts AI application. Over the years, here have appeared taxonomies for defining fairness, defining how systems could be built upon bias. In state-of-the-art methods, investigations were made in tracking of fairness outcomes, most times, across domains and sub-domains, to propose various approaches for addressing the issue (Mehrabi, 2019). It is necessary for companies and researchers to implement ML technologies that do not cause social implications or biases based on gender, ethnicity, disability, and so on. The field of biased mitigation and fairness is extended to a level wherein beginner-learners can easily lose direction. The extant literature clustered different strategies for improving ML fairness into well-accepted concepts like pre-processing, in-processing, and post-processing paradigms and 11 sub-method areas (Caton, 2020). Consciousness toward ethical issues has increased due to the Boolean response towards the factorial reality of machine learning systems altering our sociological landscapes and daily encounters. ML fairness, an evolving landscape, holds the intensity to protect social implication, against data- and model-aided unfair treatment for race, gender, disability, or sexual orientation, among others. Recent fairness arguments and methodologies have few constraints in practice, undermining their potential (Oneto, 2020).

Types of Bias and Fairness Definitions

Software for machine learning finds its use in healthcare, transport, finance, HRM, etc. The training of these systems on biased data is might clearly result in biased behavior on the part of the systems. Therefore, most fairness testing techniques in classification-based ML models are described in the literature under Perera et al. (2022). So far, within society, biasing and fairness problems have substantially arisen with algorithmic decision-making in the business world. Many of these can be unfavorable to minority gender, ethnicity, and race. Hence, research has turned, from optimizing toward a fixed objective function, and become focused on fairness and accountability. Fairness in AI in the area of healthcare is important to eliminate an episode of

health disparities and hopefully give patients their desired treatment. The various (often contradictory) definitions of fairness make this crucial endeavor for little of its accord. Meijden (2025) found at least 27 definitions of fairness from the recent literature, depending on use of model, decision type, and distributive justice ethics.

Intersectional Fairness

Now that single attributes no longer receive recognition due to the demand for fairness in an intersectionality context, it needs to be mentioned that the automatic debiasing protocol does involve the discovery of machine learning model bias. Evaluations of debiasing, however, have been carried out for attributes represented in binary terms, such as gender and race, only. The fair methods should acknowledge that certain bias influencer groups favoring the single attributes also intersect with overlapped attributes (Subramanian, 2021). The engineering of machine learning systems with the integration of social science and art research concepts of intersectional critique allows for the pursuit of alternate fairness avenues. Research has indicated that "situating/situated knowledges," "figuration," "diffraction," and "critical fabulation/speculation" may provide theoretical and methodological tools for the design grounded in concepts. Interdisciplinary interventions require understanding intersectional feminist methodologies to reinforce more inclusive, accountable, and contextualized ML system designs (Klumbytė, 2022).

Ethical Concerns and Societal Impact

AI has disrupted many different sectors over a short period of time and is actively disrupting healthcare online, where machine learning could strike the clinical operation field and, in so doing, enhance patient outcomes. Its use within a medical context raises several ethical dilemmas to consider, such as justice, fairness, transparency, patient consent and confidentiality, liability, and the provision of patient-centered, equitable care. Such considerations become even more concerning because AI systems have the embodying capability of soliciting bias from unrepresentative data sets and thus models which are only validated internally. Any AI application in the healthcare industry ultimately creates issues related to bias, lack of transparency, and patient trust (Wiener, 2024). AI has enabled the improvement of healthcare, justice, and commerce to raise ethical battles. Explainable AI (XAI) improves the interpretability of algorithmic decision-making, particularly for the medical domain, in which model opacity can affect patient outcomes. The AI Act of the European Union and other kindred regulations encourage the construction of multidisciplinary approaches that combine innovation and accountability to develop AI systems that respect human rights and build trust. The development of AI in an ethical manner fostering fairness and transparency necessitates interdisciplinary collaboration and regulation flexibilities (Falvo, 2024). Algorithms' biases within the big data brackets over issues such as Fair Information Distribution and AI applications' integrity in Africa. In view, fees for dealing with urgent issues include: addressing biases devaluing the question of information fairness in a credible attempt to win back public trust. Inclusivity, cultural sensitivity, and local community engagement in the development of AI systems remain as crucial elements. Ethical basic principles and transparency lead to bias reduction, trust creation, and equal access to technology. To ignore this reality would be to deepen social disparities, with reduced public trust in the institutions while potentially stagnating the economies (Pasipamire, 2024).

Trust Dynamics in AI Systems

Trust processes in the personnel selection aperture have been investigated in light of breaches of fairness. Thus participants rated the evaluation outcomes for their preselection (collections of preselected actors) with human and automated system designers. In the case that the preselection had gender bias (male-dominated selections for some tasks), participants initially had an initial trust deficit in the automated system against the human decision-maker due to the breaching of trust and, subsequently, an added trust repair intervention (an excuse for bias preselection). Again, the trust erosion following trust assaults in the form of biased selections and trust repairs were less marked for decision support tools-client systems considering the impact of varied ones. Thus the deficiency of such systems in terms of trust had been enlarged. These issues cropped up when some hairline-description shook up perceptions of that system again. The learning process of automation either way sometimes applies to emerging conclusion when trust where trust decisions follow ethical standards. Often, machine learning techniques become challenging because they are somewhat opaque. The trust model builds on the interpretation of the rationale behind predictions, notably in the context of predetermining outcomes. It can consequently impact on the right placement of action precedent on predicting decisions, or deployment considering the standardization set of a model which may enhance a significant number of these models and predictions. The narrative gaze on whether explainability is required in medical machine learning systems restates; others also claim that, given the accuracy and confidence of such systems, some form of epistemic-kind justification is already evident in the system. There might be issues with either choice but not with regard to the epistemic concerns of medical professionals. Theoretical justifications could offer solutions to contextual epistemic issues and explain why medical professionals such as clinicians must use the system for the work. The instructions must be such that they establish claims, what the machine ought to know (Theunissen, 2022). Trust, fairness, and accountability invariably lead to the understanding of the reasoning of AI systems. Explainable AI (XAI) aims to help non-technical users observe the workings of "black box" system designs like deep neural networks. AI decisions have groundbreaking implications in the spheres of healthcare, banking, and law; hence, transparency plays a part in these areas. XAI, together with forcing transparency into models that somehow help members of different teams construct models with ethical principles, equality, and the values of humans on an increasing basis, is seen as ethically acceptable. Once again, the rapidly alternating interpretations of XAI and thereby be adopted have resulted in a model that will be easily interpreted by users to comprehend decision processes in which all parties agree are otherwise Krushnasamy et al. (2025).

Legal and Regulatory Implications

Fairness and bias in AI models (fair-AI) literature is growing rapidly, making it difficult for researchers and practitioners to understand the field. Several policy initiatives, standards, and best practices have been proposed to establish bias and fairness management principles, procedures, and knowledge bases. Current research provides concise surveys of fair-AI methods and resources and the main AI bias policies to guide researchers and practitioners. Álvarez (2024) suggests a dual-layer architecture for policy advice: a Legal Layer (focusing on the EU context) and a Bias Management Layer (addressing bias understanding, mitigation, and accounting). AI's rapid development has raised ethical issues like bias, lack of transparency, and privacy, necessitating ethical governance in AI systems. The Ethical Artificial Intelligence Framework Theory (EAIFT) novel unearths real-time monitoring, open decision-making in making applicable ethical controls, identification and remediation of biases, and matching the

philosophical path behind AI to changing ethical and legal norms. The model suggests "ethical AI watchdogs" and dynamic compliance algorithms, adapting themselves to regulatory changes, to provide automatic surveillance on ethical grounds. This paradigm stimulates transparency and explicability to build user trust and test and remedy bias for fairness (Ejjami, 2024).

Methodology

Detecting and Measuring Bias

Fairness Metrics and Evaluation Frameworks

Being fair in an assessment necessitates the use of different measurements and understanding how the models agree and disagree with one another. Empirical research focusing on multiple fairness assessments, datasets, and associations for different measurements have corroborated the idea that fairness-estimate metrics include positive and negative correlations and also some that are uncorrelated. Conventional fairness measurement methods in this testing setting cannot measure the size of output disparity-an essential criterion for regression models.. For regression-based ML systems, researchers developed new fairness measures to fill this gap. For regression-based machine learning systems, "fairness degree" and search-based fairness testing (SBFT) have been proposed (Perera, 2022). Researchers recommend clinical utility, performance-based metrics (area under the receiver operating characteristic curve), calibration, and statistical parity for medical applications due to fairness limitations. AI developers and assessors can evaluate model fairness and bias mitigation strategies using different metrics depending on the intended use and ethical framework, promoting more equitable AI-based implementations (Meijden, 2025).

Data Analysis and Interpretation

Testing Approaches for Detecting Bias

Fairness testing has been shown to work on real healthcare data, particularly emergency department wait-time prediction software. Compared to the best methods, search-based fairness testing is 111% and 190% more effective and efficient. Improved fairness measures and testing methods for regression-based ML systems can help software teams make data-driven deployment readiness decisions by assessing prediction fairness. These scientific advances establish fairness standards for emergency department wait-time prediction (Perera, 2022). The issues of ethical slippage are becoming increasingly significant violations for the deep neural networks (DNNs) that are proprietary in themselves. Knowledge of fairness is desired in the DNNs in sociological observations. More scalable methods to detect harmful DNN instances with some lightweight methods such as gradient computation and clustering are more efficient. The testing of the presented methods shows that they can explore the search space for 9 rounds and produce 25 times the number of discriminatory instances in half to 1/7 of the time taken by the older methods (Zhang, 2020).. Deep learning-based recommender systems (DRSs) are widely used in industry, but they suffer from the echo chamber and Matthew effect, which affect fairness. Bias may arise when the system provides lower-quality recommendations to a fraction of its users or when the selection of items become comparatively unfair. Due to the lack of an exhaustive systematic approach in ensuring the fairness of recommendation systems, most existing notations and testing methods that have been used for conventional classifiers are difficult to re-purpose in the setting of DRSs program. To gauge equality on DRS, FairRec uses the metric of model utility and considers two additional fairness metric, item diversity and item popularity. Robustness testing of such large candidate pools can adapt optimization search-based

techniques to uncover groups left unnoticed using the group as a whole. The “Research on industry-level enterprises featuring your own DRSes: what are top companies out there” data report reveals that state-of-art DRSes run by top firms when enhanced significantly boost testing and go up to 95% accuracy completion rates in between half and one-eighth the period of time required for the other testing process (Guo, 2023).

User Understanding and Sensemaking of Fairness Results

Recommender systems, exploratory tools, and dashboards can help users spot machine learning fairness issues. Researchers studied how people understand fairness issues using different de-biasing affordances to design these systems. Quick de-biasing recommendations that lack nuance and "what-if" style exploration that takes time but can lead to deeper understanding and transferable insights are in conflict. Logs, think-aloud data, and semi-structured interviews show that exploratory systems encourage rich hypothesis generation and testing, while recommendations provide quick answers that satisfy participants but reduce information exposure. These findings show ML fairness system design requirements and trade-offs for accurate and explainable assessments (Gu, 2021). AI systems are boundary objects—interdisciplinary artifacts supported by different fields and providing shared discourses. AI system development and operation must be examined to reveal political dynamics and bias introduction points. Hermeneutic reverse engineering permits critical analysis of AI system data and algorithms. This framework analyzes technocultural objects to understand how they create meaning and context. Cultural analysis and speculative imagination of alternative realities are used to identify existing meanings and assumptions, key signification elements, and ways to reassemble meanings. Cultural consideration and technological imagination can unpack AI-created meanings and design innovative approaches for more inclusive AI, allowing critical examination of biases and their effects on different social groups (Shukla, 2025).

Bias Mitigation Strategies

Pre-processing Approaches

It is important to note the intricate and/or fragile balance between the quality of a data input and the extent of how fair the output is. A common flaw on this topic is the mix-up of demographic data errors and data errors not based on any bias towards ethnic or racial groups. Many research projects point out that there is little relevance between quality problems and group differences in terms of missing data from the records. Data cleaning is an example of such a method which very rarely the existing bias within the procedures and the procedures themselves.. When it does, it is more likely to worsen fairness than improve it, especially when cleaning methods are not carefully chosen. Given its implications for many production ML systems, this finding is concerning. Addressing these challenges requires a holistic analysis of data quality, cleaning method effectiveness, and ML model performance across demographic groups. Fairness-aware data cleaning methods and their integration into complex ML-based decision making pipelines should be the focus of future research (Guha, 2023). The historical data used for training can be biased by machine learning algorithms. Discrimination-aware data mining (DADM) and fairness, accountability, and transparency machine learning (FATML) communities have developed computational methods to address these issues, but implementing them is difficult. Organizations may lack sensitive data on gender, ethnicity, sexuality, and disability needed to diagnose and mitigate indirect discrimination-by-proxy like redlining. They may also lack the skills to identify and address fairness issues in complex sociotechnical systems. Several approaches have been proposed to address these knowledge and information gaps: trusted third

parties could store data for discrimination discovery and fairness constraints in a privacy-preserving manner; collaborative online platforms could allow diverse organizations to share contextual and experiential knowledge; and unsupervised learning and interpretable algorithms could develop fairness hypotheses for selective testing. Preparing computational fairness tools incorporates the analysis of difficult, complex and sophisticated situations where those tools are to be used because fairness in machine learning is also a social and cultural phenomenon (Veale, 2017). With the advent of how data flows have been reinvented by artificial intelligence, the focus on bias has also inclined pleasantly. With the help of creative machine learning tools, societal considerations and challenging gender boundaries, algorithms have been designed to discriminate and redressing emergent animal biases. Such effort calls for addressing bias at the root, at the level of data collection, model validation, and decision-making. Moreover, every facet of this model focusing on the incorporation of fairness includes dependable model examination, which helps the system minimize inequity among different societies by employing adaptive learning and fairness aware machine learning methodologies. Besides this, for also adding the training datasets possess a variety of the target groups as well as other will require training them on more than one target groups. And the purpose is to enhance the societal prejudices and moral conduct by the developed AI as well as to develop trust in the use of AI. In fact, such sciences as data science have case studies, which incorporate metrics that evaluate bias reduction and allow conclusion that all these approaches work (Mishra, 2024).

In-processing Methods

Machine learning algorithms frequently increase the biases in data operation, which then result into unfair decisions. Various techniques are being implemented to ensure that algorithms are accurate and fair. Algorithmic fairness is when different methods are used to embed fairness into the machine learning algorithm. This measures bias during training and validation while making learning to decrease the same. Reweighting and adversarial training mitigate unfair weighting by prejudice stance. The emphasis on justice, transparency, and inclusivity significantly shape better choices for the design of machine learning systems by researchers, lending themselves to ethical considerations (Dhabliya, 2024) .Algorithmic bias arises when machine learning models with a reason-able degree of accuracy in-favoring "good" outcomes for one side of a sensitive category (e.g., gender or race) underestimate `good' outcomes for underprivileged minorities. Models optimizing only for accuracy, without in-jecting any fair terms further along these lines, exhibit such a bias. The obvious choice in addressing this challenge is then to factor fairness into the learning objective. The optimization of accuracy and underestimation bias by multi-objective optimization strategies like Pareto Simulated Annealing is the notorious way forward. Both synthetic and actual datasets evidence that it is possible to pick model families that offer varied accuracy/fairness tradeoffs (Blanzeisky, 2021). To address fairness in complex unfairness landscapes, causal Bayesian networks can provide reasoning and intervention. Interestingly, optimal transport theory can apply restrictions that pressure and shape the whole distribution of sensitive attributes; this method is a break from other current approaches that tend to restrict themselves to the most basic quantities. For a widely accepted approach, one may adopt this fairness criterion for the studied methodology for numerous fairness criteria for diverse settings and at the same time produces strong theoretical guarantees. Fair representation learning methods that are attuned to fairly generalize over unseen tasks and methods that understand the legal environment as regards using sensitive attributes to enforce fairness are pivotal and contribute toward better addressing fairness (Oneto, 2020).

Post-processing Techniques

Due to training data biases, healthcare machine learning classifiers often reproduce or worsen health disparities. Post-processing methods adjust model predictions for fairness without interfering with learning or requiring access to the original training data, preserving privacy and allowing application to any trained model. State-of-the-art debiasing methods in the post-processing family are compared across synthetic and real-world healthcare datasets using performance and fairness metrics to determine their strengths and weaknesses. To mitigate bias, such experiments examine trade-offs between group fairness and predictive performance, as well as between different definitions of group fairness, and analyze impacts on untreated attributes. These evaluations reveal how to balance accuracy and fairness in healthcare debiasing (Dang, 2025). In ethical areas like healthcare and parole, fairness isn't enough; contentious decisions must be auditable, understandable, and defensible. Attention mechanisms can ensure fairness and explain decision-making by attributing features. Due to attention interventions and weight manipulation, attention-based models can identify performance and fairness features. Research has shown that post-processing bias mitigation strategies work for tabular and textual data (Mehrabi, 2021). Machine Learning (ML) decision-making software may favor certain groups based on gender or race. Many mitigation methods promise to automatically fix fairness issues, but they sacrifice accuracy. New search-based methods for repairing ML-based decision-making software aim to improve fairness and accuracy. Compared to state-of-the-art bias mitigation methods using different fairness measurements, multi-objective search approaches for binary classification methods increased accuracy and fairness in 61% of test cases, while traditional methods decrease accuracy when reducing bias. These advances help software engineers improve fairness without sacrificing accuracy, a major issue in fairness-critical applications (Hort, 2024).

Comparative Analysis of Mitigation Approaches

In the grand scheme of things that determine the shape in which the society will be propelled, as a field of data science and machine learning, fair algorithmic decision-making systems hold utmost importance. It is essential that the frameworks go through the entire lifecycle of a data science project to inject into their training data scientists and practitioners with tools to identify and handle bias and fairness in real-world situations. In the present time, available resources with bias mitigation focus more on ML training and optimization but provide hardly any confirmation to how this can be applied in making good decisions. The ideal training process is based on building an understanding of bias and then learning to mitigate bias in real-life data science. It has been through an effective learning program that participants can get their heads around different forms of bias or have lucid conversations about bias, while giving decision-makers some awareness on which metrics apply in evaluating various options involving trades of risks (Ghani, 2023) . Within this learning scenario, one less complicated thing can be that one can evaluate project scoping factors that might affect fairness outcomes and act on the model predictions. This would add so many considerations, and these would be algorithmic fairness metrics, definitions, case studies, data bias understanding, and most importantly, model bias mitigation using tools such as the Aequitas toolkit to bridge theory with internal practices. Eminently necessary are algorithms created to eliminate multiple types of non-privacy-compromising biases. Some classes of algorithms improve fairness and trustworthiness but these require sensitive attributes to evaluate. With bias-mitigation-algorithm variants being developed to increase project final quality or to progress into some academic research and later to industrial research, the subset seeking a fairer design by modifying the algorithm itself must be

substantially addressed. There will be more processing that will come later, concerning sensitive attributes, and that is why the problem we will introduce from the beginning will take us through testing with regard to the sensitive attribute-data-for-measure's fairness before we've taken steps one must be cautioned with-because if regeneration of sensitive attribute data needs to be performed in order to achieve fairness, then it appears partial information available to at least some degree will allow one attendance.. Disparate impact remover is the least sensitive bias mitigation strategy to inference accuracy levels, according to studies. Bias mitigation algorithms with reasonably accurate inferred sensitive attributes outperform standard models in fairness and balanced accuracy. Inferred rather than actual sensitive attributes may improve fairness in black box AI systems using bias mitigation strategies (Wang, 2025).esearchers have tested bias-constrained models (new to NLP) and extensions of iterative nullspace projection techniques that can handle multiple identities to address intersectional biases (Subramanian, 2021). While valuable, human feedback in algorithmic decision-making can introduce biases if not controlled. Researchers have found ways to detect and correct evaluator biases in pairwise ranking tasks. Evaluators' pairwise rankings may reflect both the items' latent quality scores and their biases against or in favor of certain groups. Novel methods extending classic models by adding bias parameters for each evaluator can detect and correct these biases, producing rankings that match true latent quality scores. These methods use alternating optimization to optimize log-likelihood for items' latent scores and evaluators' biases. These methods can reconstruct evaluator biases and outperform non-trivial competitors in producing rankings closer to unbiased standards, according to synthetic and real-world data experiments (Ferrara, 2024). Table 1 compares bias mitigation strategies by implementation type, application domain, and practical considerations to help understand their strengths, weaknesses, and appropriate contexts.

Approach	**Type**	**Key Features**	**Context**	**Strengths**	**Limitations**	**Citation**
Evaluator Bias Detection in Ranking	Post-processing	Corrects ranking outputs using evaluator-specific bias parameters	Recommender Systems, Peer Reviews	Improves ranking fairness without retraining	Requires historical pairwise comparison data	Ferrara (2024)
Disparate Impact Remover	Pre-processing	Removes correlation between input features and protected attributes	Privacy-Constrained Settings	Robust to inferred attributes	May affect accuracy with noisy inference	Wang (2025)

Approach	Type	Key Features	Context	Strengths	Limitations	Citation
Bias-Constrained NLP Models	In-processing	Adds fairness constraints to NLP model training	Text Classification (Intersectional)	Handles multiple identities in input	Still emerging; limited domain generalization	Subramanian (2021)
Aequitas Toolkit	Post-hoc Analysis	Fairness auditing, group fairness metrics	Industry & Policy	User-friendly; supports real-world deployment	Limited to observed fairness outcomes	Ghani (2023)
Inferred Attribute Mitigation	Any (Hybrid)	Uses predicted sensitive attributes to enable bias mitigation	Black-box AI Systems	Works when true attributes are unavailable	Depends on inference accuracy; may introduce new biases	Wang (2025)

Table 1: Comparison of Bias Mitigation Approaches Based on Context and Application Constraints

Significance of Results

Application-Specific Considerations

Criminal Justice and Recidivism Prediction

Quantitatively predicting recidivism by assessing criminal defendants' likelihood of committing future crimes is helping criminal justice officers manage criminal populations. What matters more than prediction accuracy is whether these algorithms make fair decisions. Gender, race, age, ethnicity, and unemployment affect ML system fairness, according to research. Supervised ML algorithms' recidivism predictions on Greek female prison records have been examined for fairness based on age at release and employment status during first imprisonment. Statistical analysis of ML-based predictions has revealed fairness issues in this sensitive area (Bentos, 2024). Multiple machine learning best practices exist without a consensus on standards. Fairness standards have little practical guidance. Fairness in errors (both false negatives and positives) makes weighting, tradeoffs, and judging models with different error types across races difficult. In justice settings, higher false positive rates for one racial group and higher false negative rates for another have been studied, demonstrating the limits of computational approaches for tradeoff resolution. Beyond technical fixes, leadership, line workers, stakeholders, and affected communities may need courageous conversations to address systemic issues (Russell, 2020).

Healthcare Applications

AI systems can reliably assess surgeon skills through intraoperative surgical videos, which could affect surgeon credentialing. All surgeons must be treated fairly by these systems. Surgical AI systems deployed on robotic surgery videos from geographically diverse hospitals (USA and EU) showed underskilling bias (erroneously downgrading performance) and overskilling bias (erroneously upgrading performance) at different rates across surgeon sub-cohorts. TWIX trains AI systems to provide expert-like visual explanations for skill assessments to address biases. TWIX reduces under- and overskilling biases and improves hospital AI system performance, unlike baseline strategies. These findings apply to medical student skill assessments, a prerequisite for AI-augmented global surgeon credentialing programs that treat all surgeons fairly (Kiyasseh, 2023). Clinical utility, performance-based metrics (area under the receiver operating characteristic curve), calibration, and statistical parity are the best group-based fairness metrics for medical applications (Meijden, 2025). To ensure institutional explanations for medical AI systems are effective, assumptions must be disclosed. To avoid biases and failures, experts and end-users must coordinate field functionality, accuracy evaluation metrics, and auditing procedures. This broader explanatory framework is needed for epistemically meaningful post hoc explanations or accuracy scores, allowing medical professionals to use these systems effectively (Theunissen, 2022).

Business and Commercial Applications

Generative AI in business-to-business (B2B) sales processes can improve efficiency, personalization, and prediction, but it also raises ethical issues and bias risks that could damage trust and fairness. The ethical landscape includes data privacy, security, transparency, accountability, and informed consent. AI algorithm bias can affect customer engagement and satisfaction, requiring mitigation strategies. Sustainable business practices require trust in AI systems and fair customer treatment. AI ethics must be constantly updated and learned to create trustworthy B2B sales environments, according to industry leaders. Creating ethical frameworks and guidelines for fair and transparent AI systems ensures AI benefits without compromising ethics (Tadi, 2024). AI systems deployed in cloud infrastructure have transformed many industries but raised ethical concerns about bias and fairness. Quantitative data from commercial deployments shows demographic disparities in facial recognition, hiring, lending, criminal justice, and healthcare error rates by 40+ factors. These disparities cause economic disadvantages, limited opportunities, and public distrust, especially in marginalized communities. Resampling, synthetic data generation, and fairness-aware algorithms reduce bias metrics by 40-70% while maintaining predictive performance. Technical solutions alone are insufficient; governance frameworks are needed. Though AI ecosystem implementation gaps remain, regulatory approaches, certification mechanisms, participatory design, and professional ethics outperform voluntary guidelines. The best approach combines technical debiasing with strong governance, especially regulatory frameworks, which is both ethical and economic as AI influences critical infrastructure and decision-making worldwide (Gupta, 2025).

Discussion

Practical Implementation and Industry Perspectives

Industry Needs and Challenges

Fair ML tools must be designed around real-world needs to improve industry practice. Systematic semi-structured interviews with ML practitioners and anonymous surveys have

found alignment and disconnect between team challenges and fair ML research literature solutions. Future ML and HCI research should better address practitioners' needs in developing fairer ML systems, suggesting that academic research should become more industry-relevant (Holstein, 2018). Academic bias and fairness resources focus on ML training and optimization, leaving practitioners without comprehensive frameworks for making decisions throughout a real-world project lifecycle (Ghani, 2023). Despite growing interest in software fairness in the software engineering community, little is known about fair machine learning engineering, the software engineering process used to develop fairness-critical ML systems. Significant knowledge gaps exist regarding practitioners' fairness awareness and maturity, required skills, and optimal development phases. Professional surveys have revealed how fairness is perceived and managed in practise, highlighting relevant tools and approaches. Fairness remains a second-class quality in AI system development, according to key findings. Building specific methods, development environments, and automated validation tools could help developers reverse this trend and address fairness throughout the software lifecycle (Ferrara, 2023).

Tools and Frameworks for Implementing Fairness

Fairness as a Service (FaaS) protocols compute and verify the fairness of any machine learning model securely, verifiably, and privately as fair machine learning research grows. For privacy, these designs use cryptograms to represent data and outcomes, and zero-knowledge proofs ensure their well-formedness. Any ML model can be audited for fairness using model-agnostic fairness metrics without trusted third parties or private channels. Secure transparency and verification can be achieved by making cryptograms of all input data publicly available for auditors, social activists, and experts to verify. Implementation studies on publicly available datasets with thousands of entries have shown that such approaches are feasible (Toreini, 2023). FaaS is a novel privacy-preserving, end-to-end verifiable solution for ML algorithmic fairness audits. Being model-agnostic and independent of fairness metrics, it can be used by multiple stakeholders as a service, unlike previous designs. It ensures cryptogram well-formedness and protocol provenance with zero-knowledge proofs. Proof-of-concept implementations using off-the-shelf hardware, software, and datasets have shown the protocol's scalability to large-scale auditing scenarios (over 1000 participants) and security against various attack vectors (Toreini, 2023). As machine learning software makes life-changing decisions, algorithmic discrimination concerns have grown in machine learning and software engineering communities. Researchers have found ways to detect and mitigate "algorithmic bias" or "ethical bias" in AI software. Fairway removes ethical bias from training data and models using pre- and in-processing. Results show bias detection and mitigation in learned models without affecting predictive performance. This suggests that bias testing and mitigation should be routine parts of machine learning software development, with Fairway supporting these tasks (Chakraborty, 2020).

Balancing Fairness with Other Objectives

Fairness measurement research has developed technological rules for assessing model fairness and bias mitigation strategies, making AI-based implementations more equitable (Perera, 2022). Fairness implementation is difficult because bias mitigation methods often sacrifice accuracy for fairness. Novel multi-objective search methods optimize both objectives simultaneously. Search-based repair techniques have improved binary classification accuracy and fairness in most test cases, while traditional bias-reduction methods decrease accuracy. This advancement lets software engineers improve fairness without sacrificing accuracy, a major issue in real-world applications (Hort, 2024). Multi-objective optimization strategies like Pareto Simulated

Annealing can optimize accuracy and underestimation bias by adding fairness as a criterion in model training. This method finds model families with different accuracy/fairness tradeoffs, allowing practitioners to choose models that meet their needs (Blanzeisky, 2021).

Intersectional Approaches to Fairness

Recent research has identified intersectional bias, which combines sensitive attributes like race and gender. Comprehensive reviews of state-of-the-art intersectional fairness approaches have produced taxonomies for notions and mitigation strategies, as well as key challenges and future research directions (Gohar, 2023). Intersectional feminist approaches to machine learning system design are novel. Research has shown workshop frameworks that highlight tensions and possibilities in interdisciplinary ML systems design for more inclusive, contextualized, and accountable systems. Open-ended experimental spaces are needed for critical theoretical concepts to work as design methods. Intersectional knowledge, rooted in history and socio-politics, offers new perspectives on designing fair and accountable systems, promoting equitable AI development (Klumbytė, 2022). Intersectional feminist principles guide equitable, ethical, and sustainable AI research. Researchers have developed AI research principles that address environmental impact and consent while examining and challenging unequal power in data science. These principles account for unequal, undemocratic, extractive, and exclusionary forces in AI research, development, and deployment; identify and mitigate predictable harms before unsafe or discriminatory systems are released; and inspire creative, collective approaches to building more equitable, sustainable AI ecosystems (Klein, 2024).

Explainable AI and Fairness

Explainable AI (XAI) is understanding and explaining how AI models make decisions. Understanding AI systems' reasoning processes is essential for trust, fairness, and accountability as they become more complex, especially machine learning models. XAI unravels the "black box" nature of complex models like deep neural networks, showing how inputs become outputs. This transparency is crucial in high-impact industries like healthcare, banking, and law. Explainability helps identify and mitigate biases, improve model performance, and meet regulations. As AI technologies advance, model accuracy and interpretability must be balanced to keep systems ethical, transparent, and in line with human values. Nimma et al. (2025). In medicine, debate continues over whether ML systems need post-hoc explanations for individual decisions to build trust and ensure accurate diagnoses, or whether high accuracy and reliability are enough. Both approaches have limitations and may not satisfy medical professionals. A different view is that these systems need institutional explanations to convince medical professionals to use them in specific contexts and address users' epistemic concerns (Theunissen, 2022). In complex areas like healthcare and parole, fairness isn't enough; contentious decisions must be auditable, understandable, and defensible. Attention mechanisms can ensure fairness and attribute features to decision-making processes. Attention-based models use weight manipulation and attention interventions to attribute performance and fairness. These methods work for tabular and textual data (Mehrabi, 2021).

Privacy-Preserving Fairness Assessment

Fairness as a Service (FaaS) computes and verifies ML model fairness securely, verifiably, and privately. Zero-knowledge proofs ensure the well-formedness of cryptograms and underlying data, ensuring privacy. This model-agnostic fairness computation method supports multiple metrics without trusted third parties or private channels. Security guarantees and commitments

make input data cryptograms publicly available for auditors and stakeholders to verify. Toreini (2023) found this approach feasible in implementation studies on datasets with thousands of entries. Privacy-preserving fairness auditing systems are model-agnostic and independent of fairness metrics, making them usable by multiple stakeholders. They verify cryptogram well-formedness and protocol step provenance with zero-knowledge proofs. Proof-of-concept implementations using off-the-shelf components have shown scalability to large-scale auditing scenarios (over 1000 participants) and security against various attack vectors (Toreini, 2023). Differential privacy (DP) affects ML models, reducing accuracy and increasing bias, according to fairness research. Security researchers have proposed and tested backdoor attacks to inject bias into NLP models, finding that modern transformer-based models like BERT and RoBERTa are particularly vulnerable, with stealthy attacks generalizing to dynamic triggers at test time. The intersections of trust, privacy, security, and fairness in ML are often studied separately but require integrated approaches (Atabek, 2023)..

Ethical, Legal, and Cognitive Frontiers

Al-Omari et al. (2025) discussed the governance and ethical challenges of AI in higher education, highlighting benefits like improved engagement and efficiency. They emphasized the need for strong governance to address biases and ensure fairness, advocating for international cooperation and robust policies to ensure AI's equitable impact. Hassan et al. (2024) developed an optimized deep learning model for text summarization, improving performance on standard datasets. Their approach is particularly useful for handling complex text, including legal document categorization. Jabbar et al. (2024) examined text-stemming techniques, highlighting their role in improving text preparation for NLP tasks, especially in legal text classification. Ammar et al. (2024) investigated using BERT and GPT models for legal judgment prediction in Arabic, finding transformer models effective in classifying legal texts and emphasizing the importance of fine-tuning AI systems for specific domains like law. Rehman et al. (2025) reviewed facial emotion recognition (FER) techniques, finding that deep learning models, especially CNNs, outperform traditional methods in handling complex image data. They highlighted challenges like lighting and pose variations, suggesting hybrid models that integrate deep learning for better performance. Gaber and Alenezi (2024) examined how serverless computing reduces infrastructure costs and accelerates deployment through FaaS architectures, enhancing scalability, pay-per-use pricing, and developer productivity. However, challenges like vendor lock-in and security risks were noted, concluding that serverless computing is ideal for modern cloud-native applications. Alyousef and Al-Omari (2024) explored AI's role in healthcare, identifying regulatory challenges such as data privacy and algorithm bias. They called for updated global frameworks that balance innovation with patient safety, ensuring ethical AI deployment in healthcare. Semary et al. (2023) used transformer models like RoBERTa for sentiment classification, achieving high accuracy on datasets like IMDb and Twitter. Their hybrid approach showed the potential of deep learning models to handle complex text analysis tasks.

Policy Frameworks and Standardization

Rapidly growing literature on bias and fairness in AI makes it hard for researchers and practitioners to understand the field. Many policy initiatives, standards, and best practices have been proposed for bias and fairness management. Short surveys of fair-AI methods, resources, and policies help researchers and practitioners. NoBIAS architecture provides structured frameworks for addressing challenges, including its Legal Layer (EU context) and Bias

Management Layer (understanding, mitigating, and accounting for bias) (Álvarez, 2024). The Ethical Artificial Intelligence Framework Theory (EAIFT) introduces ethical reasoning to AI systems. Real-time oversight, open decision-making, bias detection, and ethical and legal adaptation are its priorities. The framework recommends "ethical AI watchdogs" that automatically monitor systems and ensure ethical operation and dynamic compliance algorithms that adapt to regulatory changes. This method builds trust and detects and corrects bias to ensure fairness by promoting transparency and explainability. Qualitative methods combining stakeholder interviews, content analysis, and expert commentary found that EAIFT outperforms existing frameworks in proactively reducing biases, increasing transparency, and ensuring ethical standards (Ejjami, 2024).

Conclusion

This literature review has synthesized research on bias-aware machine learning, examining approaches to reconstruct trust and fairness in intelligent systems. The field has evolved from simply identifying bias to developing sophisticated methods for detecting, measuring, and mitigating unfairness across diverse applications. Key insights include the recognition of intersectional bias, the development of domain-specific fairness metrics, and the creation of tools for practical implementation in industry settings. Several themes emerge across the literature. First, fairness is multifaceted and context-dependent, requiring careful consideration of domain-specific needs and stakeholder perspectives. Second, there is a growing emphasis on explanatory mechanisms that provide transparency regarding both the performance and fairness characteristics of models. Third, privacy-preserving approaches for fairness assessment enable organizations to evaluate bias without compromising sensitive data. Finally, balancing fairness with other objectives like accuracy remains challenging but increasingly feasible through innovative optimization approaches. Despite significant progress, important gaps persist. Many studies focus on binary protected attributes, with fewer addressing intersectional concerns. Real-world implementation remains difficult, with fairness often treated as a secondary consideration in development processes. Additionally, the field lacks standardized evaluation protocols and agreed-upon metrics for consistent assessment across different contexts. Future research should focus on developing more comprehensive intersectional approaches, creating standardized tools that integrate seamlessly into development workflows, and establishing clearer connections between technical fairness measures and meaningful social outcomes. As AI systems continue to influence critical decisions across society, bias-aware machine learning will remain essential for ensuring these technologies serve all users equitably and maintain public trust.

Acknowledgments

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication. The authors would like to acknowledge the support of Prince Sultan University for their support making this publication successful.

References

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys. https://doi.org/10.1145/3457607

Atabek, A., Eralp, E., & Gursoy, M. E. (2023). Trust, privacy and security aspects of bias and fairness in machine learning. International Conference on Trust, Privacy and Security in Intelligent Systems and Applications. https://doi.org/10.1109/tps-isa58951.2023.00023

Gohar, U. & Cheng, L. (2023). A survey on intersectional fairness in machine learning: notions, mitigation,

and challenges. International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2023/742

Langer, M., König, C. J., Back, C., & Hemsing, V. (2021). Trust in artificial intelligence: comparing trust processes between human and automated trustees in light of unfair bias. Journal of business and psychology. https://doi.org/10.1007/s10869-022-09829-9

Toreini, E., Mehrnezhad, M., & Moorsel, A. (2023). Fairness as a service (faas): verifiable and privacy-preserving fairness auditing of machine learning systems. International Journal of Information Security. https://doi.org/10.1007/s10207-023-00774-z

Caton, S. & Haas, C. (2020). Fairness in machine learning: a survey. ACM Computing Surveys. https://doi.org/10.1145/3616865

Oneto, L. & Chiappa, S. (2020). Fairness in machine learning. Studies in Computational Intelligence. https://doi.org/10.1007/978-3-030-43883-8_7

Perera, A., Aleti, A., Tantithamthavorn, C., Jiarpakdee, J., Turhan, B., Kuhn, L., & Walker, K. (2022). Search-based fairness testing for regression-based machine learning systems. Empirical Software Engineering. https://doi.org/10.1007/s10664-022-10116-7

Shrestha, Y. & Yang, Y. (2019). Fairness in algorithmic decision-making: applications in multi-winner voting, machine learning, and recommender systems. Algorithms. https://doi.org/10.3390/a12090199

Meijden, S. L. V. D., Wang, Y., Arbous, M. S., Geerts, B. F., Steyerberg, E. W., & Hernandez-Boussard, T. (2025). Navigating fairness in ai-based prediction models: theoretical constructs and practical applications. medRxiv. https://doi.org/10.1101/2025.03.24.25324500

Subramanian, S., Han, X., Baldwin, T., Cohn, T., & Frermann, L. (2021). Evaluating debiasing techniques for intersectional biases. Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2021.emnlp-main.193

Klumbytė, G., Draude, C., & Taylor, A. S. (2022). Critical tools for machine learning: working with intersectional critical concepts in machine learning systems design. Conference on Fairness, Accountability and Transparency. https://doi.org/10.1145/3531146.3533207

Weiner, E. B., Dankwa-Mullan, I., Nelson, W. A., & Hassanpour, S. (2024). Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. PLOS Digital Health. https://doi.org/10.1371/journal.pdig.0000810

Falvo, F. R. & Cannataro, M. (2024). Ethics of artificial intelligence: challenges, opportunities and future prospects. IEEE International Conference on Bioinformatics and Biomedicine. https://doi.org/10.1109/bibm62325.2024.10822112

Pasipamire, N. & Muroyiwa, A. (2024). Navigating algorithm bias in ai: ensuring fairness and trust in africa. Frontiers in Research Metrics and Analytics. https://doi.org/10.3389/frma.2024.1486600

Krushnasamy, V. S., Al-Omari, O., Sundaram, A., & others. (2025). LiDAR-based climate change imaging in geoscience using spatio extreme fuzzy gradient model. Remote Sensing in Earth Systems Sciences. https://doi.org/10.1007/s41976-025-00197-5

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": explaining the predictions of any classifier. North American Chapter of the Association for Computational Linguistics. https://doi.org/10.1145/2939672.2939778

Theunissen, M. & Browning, J. (2022). Putting explainable ai in context: institutional explanations for medical ai. Ethics and Information Technology. https://doi.org/10.1007/s10676-022-09649-8

Álvarez, J. M., Colmenarejo, A., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Lobo, P. R., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in ai. Ethics and Information Technology. https://doi.org/10.1007/s10676-024-09746-w

Ejjami, R. (2024). Ethical artificial intelligence framework theory (eaift): a new paradigm for embedding ethical reasoning in ai systems. International Journal For Multidisciplinary Research. https://doi.org/10.36948/ijfmr.2024.v06i05.28231

Zhang, P., Wang, J., Sun, J., Dong, G., Wang, X., Wang, X., Dong, J., & Dai, T. (2020). White-box fairness testing through adversarial sampling. International Conference on Software Engineering. https://doi.org/10.1145/3377811.3380331

Guo, H., Li, J., Wang, J., Liu, X., Wang, D., Hu, Z., Zhang, R., & Xue, H. (2023). Fairrec: fairness testing for deep recommender systems. International Symposium on Software Testing and Analysis. https://doi.org/10.1145/3597926.3598058

Gu, Z., Yan, J. N., & Rzeszotarski, J. M. (2021). Understanding user sensemaking in machine learning fairness assessment systems. The Web Conference. https://doi.org/10.1145/3442381.3450092

Shukla, N. (2025). Investigating ai systems: examining data and algorithmic bias through hermeneutic reverse engineering. Frontiers in Communication. https://doi.org/10.3389/fcomm.2025.1380252

Guha, S., Khan, F. A., Stoyanovich, J., & Schelter, S. (2023). Automated data cleaning can hurt fairness in machine learning-based decision making. IEEE International Conference on Data Engineering. https://doi.org/10.1109/tkde.2024.3365524

Veale, M. & Binns, R. (2017). Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data & Society. https://doi.org/10.1177/2053951717743530

Mishra, I., Kashyap, V., Yadav, N., & Pahwa, D. R. (2024). Harmonizing intelligence: a holistic approach to bias mitigation in artificial intelligence (ai). International Research Journal on Advanced Engineering Hub (IRJAEH). https://doi.org/10.47392/irjaeh.2024.0270

Dhabliya, D., Dari, S. S., Dhablia, A., Akhila, N., Kachhoria, R., & Khetani, V. (2024). Addressing bias in machine learning algorithms: promoting fairness and ethical design. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/202449102040

Nimma, D., Al-Omari, O., Pradhan, R., Ulmas, Z., Krishna, R. V. V., El-Ebiary, T. Y. A. B., & Rao, V. S. (2025). Object detection in real-time video surveillance using attention based transformer-YOLOv8 model. Alexandria Engineering Journal, 118, 482–495. https://doi.org/10.1016/j.aej.2025.01.032

Blanzeisky, W. & Cunningham, P. (2021). Using pareto simulated annealing to address algorithmic bias in machine learning. Knowledge engineering review (Print). https://doi.org/10.1017/S0269888922000029

Al-Omari, O., & Al-Omari, T. (2025). Artificial Intelligence and Posthumanism: A Philosophical Inquiry into Consciousness, Ethics, and Human Identity. Journal of Posthumanism, 5(2), 458–469. https://doi.org/10.63332/joph.v5i2.432

Dang, V. N., Campello, V. M., Hernández-González, J., & Lekadir, K. (2025). Empirical comparison of post-processing debiasing methods for machine learning classifiers in healthcare. Journal of Healthcare Informatics Research. https://doi.org/10.1007/s41666-025-00196-7

Mehrabi, N., Gupta, U., Morstatter, F., Steeg, G. V., & Galstyan, A. (2021). Attributing fair decisions with attention interventions. TRUSTNLP. https://doi.org/10.18653/v1/2022.trustnlp-1.2

Hort, M., Zhang, J. M., Sarro, F., & Harman, M. (2024). Search-based automatic repair for fairness and accuracy in decision-making software. Empirical Software Engineering. https://doi.org/10.1007/s10664-023-10419-3

Ghani, R., Rodolfa, K. T., Saleiro, P., & Jesus, S. (2023). Addressing bias and fairness in machine learning: a practical guide and hands-on tutorial. Knowledge Discovery and Data Mining. https://doi.org/10.1145/3580305.3599180

Wang, Y. & Singh, L. (2025). Impact on bias mitigation algorithms to variations in inferred sensitive attribute uncertainty. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2025.1520330

Ferrara, A., Bonchi, F., Fabbri, F., Karimi, F., & Wagner, C. (2024). Bias-aware ranking from pairwise comparisons. Data mining and knowledge discovery. https://doi.org/10.1007/s10618-024-01024-z

Bentos, S., Bailis, E., Oikonomou, F., Spirou, S., Mavrikos, E., Chatzistamatis, S., Kotis, K., & Tsekouras, G. (2024). Evaluation of fairness in machine learning-based recidivism predictions: the case of greek female prison system. 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). https://doi.org/10.1109/iraset60544.2024.10548202

Russell, J. (2020). Machine learning fairness in justice systems: base rates, false positives, and false negatives. International Conference on Machine Learning and Applications. https://doi.org/10.1109/icmla51294.2020.00133

Kiyasseh, D., Laca, J. A., Haque, T. F., Otiato, M. X., Miles, B. J., Wagner, C., Donoho, D., Trinh, Q., Anandkumar, A., & Hung, A. J. (2023). Human visual explanations mitigate bias in ai-based assessment of surgeon skills. npj Digital Medicine. https://doi.org/10.1038/s41746-023-00766-2

Tadi, V. (2024). Navigating ethical challenges and biases in generative ai: ensuring trust and fairness in b2b sales interactions and decision-making. Journal of Artificial Intelligence & Cloud Computing. https://doi.org/10.47363/jaicc/2024(3)e104

Gupta, S. (2025). Ethical considerations in cloud ai: addressing bias and fairness in algorithmic systems. International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/ijarsct-25053

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., & Wallach, H. M. (2018). Improving fairness in machine learning systems: what do industry practitioners need?. International Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3290605.3300830

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & Lucia, A. D. (2023). Fairness-aware machine learning engineering: how far are we?. Empirical Software Engineering. https://doi.org/10.1007/s10664-023-10402-y

Toreini, E., Mehrnezhad, M., & Moorsel, A. (2023). Verifiable fairness: privacy-preserving computation of fairness for machine learning systems. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-54129-2_34

Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). Fairway: a way to build fair ml software. ESEC/SIGSOFT FSE. https://doi.org/10.1145/3368089.3409697

Klein, L. & D'Ignazio, C. (2024). Data feminism for ai. Conference on Fairness, Accountability and Transparency. https://doi.org/10.1145/3630106.3658543

Hassan, A. Q. A., Al-onazi, B. B., Maashi, M., Darem, A. A., Abunadi, I., Mahmud, A. (2024). Enhancing extractive text summarization using natural language processing with an optimal deep learning model. AIMS Press, 2024. DOI: 10.3934/math.2024616

Jabbar, A., Iqbal, S., Tamimy, M. I., Rehman, A., Bahaj, S. A., Saba, T. (2024). An analytical analysis of text stemming methodologies in information transformers. arXiv.

Ammar, A., Koubaa, A., Benjdira, B., Nacar, O., Sibaee, S. (2024). Prediction of Arabic legal rulings using large language models. Faculty of Electrical Engineering Banja Luka. DOI: 10.3390/electronics13040764

Rehman, A., Mujahid, M., Elyassih, A., AlGhofaily, B., & Bahaj, S. A. O. (2025). Comprehensive review and analysis on facial emotion recognition: Performance insights into deep and traditional learning with current updates and challenges. Tech Science Press. DOI: 10.32604/cmc.2024.058036

Al-Omari, O., Alyousef, A., Fati, S., Shannaq, F., & Omari, A. (2025). Governance and ethical frameworks for AI integration in higher education: Enhancing personalized learning and legal compliance. Journal of Ecohumanism, 4(2), 80–86. DOI: 10.62754/joe.v4i2.5781

Alyousef, A., & Al-Omari, O. (2024). Artificial intelligence in healthcare: Bridging innovation and

regulation. Journal of Ecohumanism, 3(8), 10582–10589. DOI: 10.62754/joe.v3i8.5673

Gaber, S., & Alenezi, M. (2024). Transforming application development with serverless computing. International Journal of Cloud Applications and Computing, 14(1), 1-12. DOI: 10.4018/IJCAC.365288

Semary, N. A., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2023). Improving sentiment classification using a RoBERTa-based hybrid model. Frontiers Media S.A., December. DOI: 10.3389/fnhum.2023.1292010.