IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/47AI/2026/4711926/
Vol. 2026 (12), Article ID 4711926

Mitigating Algorithmic Bias in AI-Driven Decision Systems for Financial Services*

Dr. Peter BUSCH and Trang 'Tracy' NGUYEN

School of Computing, Macquarie University, New South Wales 2109
Australia

Correspondence should be addressed to: Peter BUSCH, peter.busch@mq.edu.au

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

Artificial intelligence (AI) and machine learning (ML) systems are increasingly embedded in financial services, powering credit scoring, lending, fraud detection, and risk management. While these technologies enhance efficiency and predictive performance, they also introduce algorithmic bias that can undermine fairness, regulatory compliance, and public trust. Despite rapid growth in fairness-aware machine learning research, limited clarity exists on how bias mitigation approaches are conceptualised and operationalised within financial contexts. This study addresses this gap through a systematic literature review of 32 scholarly and industry publications published between 2015 and 2025. Using NVivo-assisted thematic coding, word frequency analysis, and cross-keyword co-occurrence analysis, the study synthesises how algorithmic bias is defined, measured, mitigated, and governed in financial AI systems. Five dominant themes emerged: data nature and sources of bias, fairness metrics and evaluation, mitigation methods, ethical and regulatory governance, and organisational implementation. The findings reveal a strong concentration on data-centric and technical mitigation strategies, while ethical and governance dimensions remain comparatively under-integrated in operational discourse. In response, the paper proposes a socio-technical bias mitigation framework that conceptualises fairness as an emergent outcome of interactions between technical systems, organisational governance structures, and regulatory environments. The study contributes by bridging technical fairness research with organisational and regulatory implementation, offering structured guidance for responsible AI deployment in financial institutions.

Keywords: Algorithmic bias; Fairness in machine learning; Financial services AI; Ethical governance

Introduction

Context and Research Aim

Automated decision systems powered by ML now represent core infrastructure in modern financial services, and are used for lending and credit scoring, fraud detection, underwriting, anti-money-laundering, and personalised financial advice; this shift has materially improved operational efficiency and scale, but it has also exposed consumers and organisations to novel harms when models reflect, amplify, or introduce unfair patterns (Hall et al., 2021; Moldovan, 2022). Recent domain surveys document both the widespread adoption of AI in finance and the many kinds of algorithmic biases that can arise across financial applications. Algorithmic bias in finance is not merely a technical shortcoming; it has regulatory, reputational, and social consequences. Regulators and sector bodies have begun to scrutinise AI use in financial services more closely, treatment of protected groups in credit decisions, transparency obligations for model-driven decisions, and requirements for risk management are already present in guidance from regulators and industry risk reports (Fritz-Morgenthal et al., 2021). In Australia, the Australian Securities and Investments Commission (ASIC) have made it clear that financial firms must embed governance and risk controls for AI deployments (ASIC, 2024). Similar concerns are echoed internationally: the

Cite this Article as: Dr. Peter BUSCH and Trang 'Tracy' NGUYEN Vol. 2026 (12) "Mitigating Algorithmic Bias in AI-Driven Decision Systems for Financial Services " Communications of International Proceedings, Vol. 2026 (12), Article ID 4711926, https://doi.org/10.5171/2026.4711926

European Union (EU) AI Act (2024) imposes obligations on high-risk financial AI systems; U.S. regulators such as the Federal Reserve, Government Accountability Office (GAO) and Federal Deposit Insurance Corporation (FDIC) have jointly underscored the importance of governance and fairness in AI-driven banking (2024); and global organisations such as the Organisation for Economic Cooperation and Development (OECD) have issued principles for transparency, accountability, and ongoing monitoring. At the same time, technical research has developed a wide range of bias mitigation methods. Pre-processing strategies modify data (e.g., re-weighting, transformations); in-processing approaches constrain model objectives or use fairness-aware learning; and post-processing techniques adjust outputs to meet fairness metrics. While these approaches show promise in controlled studies, their translation into operational finance remains limited.

Research Question and Problem

The core research question guiding this paper is: how is algorithmic bias in AI-driven financial services being mitigated in the literature, and what gaps remain between technical mitigation methods and practical, organisational adoption? Although fairness-aware and ethical AI and ML have been widely studied, a gap persists between technical advances and practical adoption. This gap highlights several focused research problems:

- Contextual effectiveness: How do different bias mitigation approaches compare when applied to distinct financial tasks, for example, credit scoring versus fraud detection?
- Practical integration: What challenges do financial organisations face when aligning fairness considerations with existing risk management and compliance processes?
- Regulatory and ethical governance: What regulatory frameworks and ethical AI guidelines can support ongoing monitoring of fairness, ensuring models remain compliant and socially responsible as data and risk profiles evolve?

By diving into these issues, this paper positions itself at the intersection of ML, financial regulatory frameworks and organisational practices. At the end of the paper, the following outcomes are presented:

- A structured review of bias types and mitigation strategies in AI-driven financial services.
- A conceptual map linking mitigation methods to specific financial applications.
- Identification of gaps between academic research and organisational adoption.
- Recommendations drawn from regulatory frameworks and directions for further applied research.

The remainder of this paper systematically explores key components of algorithmic bias in financial AI systems. Section 2 reviews the AI/ML landscape in financial services, outlining bias sources, fairness metrics, mitigation strategies, and regulatory and ethical frameworks, while highlighting organisational implementation challenges. Section 3 outlines the methodology used to review and synthesise the literature, while sections 4 and 5 present the main findings and discussion. The paper concludes with reflections on the practical implications of bias mitigation and potential directions for further research.

Background

The current AI/ ML landscape in financial services

The adoption of AI and ML in the financial sector can be traced to early experimental deployments in the 1980s (El Hajj & Hammoud, 2023). Over the past decades, advances in computational power, data availability, and algorithmic sophistication have enabled AI/ML systems to move from niche applications to mission-critical infrastructure for financial institutions. Today, AI/ML has taken over a wide spectrum of financial operations, including credit scoring, lending and underwriting, fraud detection, algorithmic and high-frequency trading, robo-advisory services, and customer service automation (Buchanan & Wright, 2021). These technologies promise efficiency gains, cost reductions, and enhanced risk management. However, their integration also raises significant challenges related to data privacy, regulatory compliance, transparency, and fairness (El Hajj & Hammoud, 2023). For example, credit decision-making represents a particularly sensitive and socially consequential application. Access to affordable credit is central to household financial resilience, enabling individuals to manage income volatility, invest in education, start businesses, and accumulate assets. The increasing use of AI-driven models in this domain intensifies the need for rigorous bias mitigation, as discriminatory outcomes - whether intentional or emergent - can systematically disadvantage certain groups (Ahmed et al., 2022). Since 2015, global research output in AI for financial services has accelerated, led by contributions from the U.S., China, and the United Kingdom (Ahmed et al., 2022). While substantial work has gone towards algorithmic accuracy in domains such as credit scoring and consumer banking adoption (Hentzen et al., 2021), gaps remain in cross-disciplinary studies that integrate technical performance with regulatory, ethical, and behavioural dimensions.

Understanding algorithmic bias

ML models can unintentionally pick up and repeat social patterns found in their training data, even when developers do not purposely include them; this is known in computer science as algorithmic bias (Johnson, 2020). Algorithmic bias in AI systems emerges from multiple sources beyond just biased training data, challenging the misconception that models are merely passive reflectors of dataset biases (Hooker, 2021). Research reveals that bias stems from a lack of diversity in development teams, rushed deployment timelines, and inadequate testing phases that fail to represent diverse societal groups (Xivuri & Twinomurinzi, 2023). The problem extends beyond data issues to encompass model design choices, which can amplify existing biases and create disparate impacts on different demographic groups (Hooker, 2021). To support this narrative, Nazer et al. (2023) provide a structured account of how bias can manifest throughout the entire AI lifecycle - from data collection and preprocessing to model design, training, validation, and even in deployment; the authors map potential bias sources at each stage of development. Although their analysis is healthcare-focused, the principle that bias occurs throughout the full lifecycle is transferable to financial AI systems, where fairness and accountability are equally critical; this broader framing cautions against reducing bias to a single technical step, instead positioning it as a systemic issue requiring governance. For information systems (IS) research, because it connects technical considerations with organisational governance and accountability, viewing bias as an algorithmic phenomenon encourages the development of policies, controls, and monitoring frameworks that align AI deployment with social responsibility, risk management, and compliance requirements. In high-stakes domains such as banking, lending, and fraud detection, the consequences of algorithmic bias are profound. Biased decisions can lead to financial exclusion, reputational damage, regulatory sanctions, and systemic risk. Translating the abstract definition of algorithmic bias into operational terms highlights why financial institutions must actively embed fairness considerations into data governance, model governance, monitoring, and ongoing evaluation; this socio-technical framing positions bias not only as a technical issue but as a strategic and ethical challenge for organisations.

Fairness metrics and evaluation

Understanding bias is only the first step; equally important is the ability to measure fairness through robust metrics. The study of fairness in ML has gained significant attention over the past decade, particularly in domains where algorithmic decisions have high stakes, such as finance, healthcare, etc. (Moldovan, 2023). A taxonomy of fairness issues has been proposed, mapping diverse approaches to address these challenges (Jui & Rivas, 2024). In finance, studies have evaluated the effectiveness of bias mitigation methods across different fairness metrics, highlighting the trade-offs between fairness, accuracy, and profitability (Moldovan, 2023). A substantial portion of research has focused on defining fairness and developing metrics to quantify bias in ML models. Moldovan (2023) evaluated 12 bias mitigation methods across 5 fairness metrics in credit scoring, while Chen et al. (2022) conducted a comprehensive study of 17 bias mitigation methods, finding that their effectiveness varies depending on tasks, models, and metrics used. van Giffen et al. (2022) identified eight distinct ML biases and 24 mitigation methods, emphasizing the importance of human judgment in algorithm design. Studies in credit scoring and loan approval have demonstrated that metrics often conflict, meaning achieving fairness along one dimension may worsen outcomes in other aspects (accuracy and profitability) (Moldovan, 2022); this highlights the challenge of translating abstract fairness definitions into actionable strategies for institutions.

Bias mitigation techniques and treatment

Caton and Haas (2024) frame fairness interventions in ML through a tripartite taxonomy: pre-processing, in-processing, and post-processing methods, each corresponding to distinct phases of the AI development lifecycle. Pre-processing approaches modify the training data before model development to reduce bias. Several techniques, especially re-sampling, re-weighting, and causal data repair are effective and widely adopted for reducing bias in ML (Werner et al., 2022). They offer a strong balance between fairness and predictive performance, and ongoing research continues to refine these methods for diverse applications (Chen et al., 2024, Huang et al., 2024). In financial services, this might involve correcting imbalances in correcting imbalances in credit data and adjusting for historically biased labels, which are essential for fair lending (Kumar, Hines & Dickerson 2022). In-processing methods incorporate fairness constraints directly within the model training process. Examples are fairness-aware loss functions, constrained optimization, and adversarial debiasing that offer robust bias mitigation, but require careful design to balance fairness and accuracy for each application (Wan et al. 2022, Yang et al. 2023); these techniques are especially valuable in financial risk models, where maintaining predictive accuracy while achieving fairness is critical. Lastly, Caton et al. (2024) classify post-processing as a key fairness intervention that operates independently of original training data, while Barda et al. (2020) empirically demonstrate its capacity to reduce subpopulation calibration variance by up to 98.8%. Similarly, Cui et al. (2023) validate a model-agnostic framework that achieves fairness with minimal impact on classification accuracy. However, despite strong

evidence across domains such as healthcare and risk assessment, effectiveness validation for post-processing in financial systems remains limited.

Additionally, González-Sendino et al. (2024) introduce a novel causal-model-based approach to generate fairer datasets: by adjusting cause-and-effect relations within a Bayesian network, their method produces mitigated training data that maintains sensitive attributes while enhancing both transparency and explainability. Specifically, within credit decision-making, Vieira et al. (2025) report that pre-processing strategies dominate empirical research: among 34 carefully reviewed studies, gender and race are the most frequently considered sensitive attributes, with statistical parity emerging as the most common fairness metric. However, they also highlight that challenges persist - especially around the need for standardised metrics, improved explainability, and addressing legal and ethical dimensions of algorithmic decision-making (de Castro Vieira et al., 2025). By synthesising causal-data generation, fairness-aware training, and post-processing corrective mechanisms, the literature suggests a layered mitigation strategy tailored for financial systems - and underscores the importance of aligning methodological rigor with domain-specific regulatory, ethical, and operational constraints.

AI regulatory and ethical frameworks

The growing deployment of AI systems has prompted the development of ethical principles and regulatory frameworks aimed at ensuring fairness, transparency, and accountability. The EU AI Act (2024 - Annex 3) legally designates credit scoring as "high-risk," demanding rigorous bias monitoring, impact assessments, and human oversight, with severe penalties for non-compliance. Similarly, the OECD AI Principles (OECD, 2024), endorsed by more than 40 countries, promote human-centred values such as fairness and explainability, encouraging mechanisms like human oversight and impact assessments in AI life cycles. Likewise, the United Nations Educational, Scientific and Cultural Organization (UNESCO's) Recommendation on the Ethics of AI, adopted in 2021 and is applicable to all 194 member states of UNESCO, states that AI stakeholders must take all necessary steps to prevent and mitigate bias throughout the AI system's life cycle, while ensuring effective remedies are available for discriminatory or unfair algorithmic outcomes. While these regulatory frameworks highlight the urgency of ethical AI algorithms; however, they do not guarantee organisational adoption. Financial institutions must translate high-level principles into actionable governance structures, technical standards, and operational practices - which often require cultural change, cross-functional collaboration, and new compliance investments. The challenge is not a lack of frameworks and techniques, but whether organisations can integrate them into daily decision-making, system design, and accountability processes.

Challenges in implementation for organizations

***Data quality and** availability*

While bias mitigation techniques offer promising pathways to enhance fairness in financial AI systems, organizations face multiple practical challenges when attempting to implement them. One key issue is data quality and availability. Many financial institutions rely on legacy datasets that are incomplete, inconsistent, or biased due to historical practices. For instance, credit scoring models may perpetuate historical lending discrimination against certain demographic groups, requiring careful preprocessing interventions to prevent inequitable outcomes (Moldovan, 2022). ML biases can stem from unrepresentative datasets, inadequate models, and human stereotypes, leading to unfair decisions and financial losses (van Giffen et al., 2022). Representation bias in data particularly affects minorities due to historical discrimination and sampling biases in data acquisition methods (Shahbazi et al., 2022). Critically, automated data cleaning, which commonly used in production ML systems, can worsen fairness outcomes rather than improve them, especially when cleaning techniques are not carefully selected (Guha et al., 2024); these findings underscore that without high-quality, representative data and thoughtful preprocessing, even advanced mitigation methods may fail to produce equitable outcomes.

Model complexity

A further challenge arises from the complexity of models once fairness constraints are introduced. In many cases, these methods require adding fairness objectives to the training process or adjusting the way models learn internal representations (Wan et al., 2022). However, deciding which fairness definition to apply is not straightforward, as different approaches can be contradictory or difficult to reconcile (Ryan et al., 2023). These adjustments increase the technical burden, reduce interpretability, and often demand expertise that may not be readily available within financial institutions. For instance, a bank that integrates fairness metrics into its credit scoring model may achieve more equitable outcomes, but the model itself becomes harder for compliance teams and regulators to audit and require extra resources for development; this trade-off between fairness, transparency, and practicality illustrates why organizations often hesitate to adopt more advanced fairness techniques in production systems.

Resources constraints and trade-offs

Implementing fairness-aware AI systems often requires significant financial and organizational resources. Developing new pipelines for data collection, preprocessing, and model retraining demands both capital investment and skilled personnel. Small and medium-sized enterprises (SMEs) face particular barriers in adopting ethical AI approaches due to lack of knowledge, skills, and resources (Crockett et al., 2023). Despite numerous published toolkits designed to support ethical AI practices, there is limited grassroots implementation, especially among SMEs, with few practical application examples and low uptake rates (Crockett et al., 2023). The software engineering community shows that fairness remains a "second-class quality aspect" in AI system development, requiring specialized methods, development environments, and automated validation tools (Ferrara et al., 2023). For example, post-processing adjustments to fraud detection systems may reduce false positives for minority groups, but can also increase overall fraud risk exposure, raising concerns among risk management teams; these tensions highlight why the costs of implementing fairness are considered not merely technical but also economic and organizational, requiring firms to weigh short-term efficiency against long-term reputational and compliance benefits. Li & Li (2024) identify a triangular trade-off between robustness, accuracy, and fairness in deep neural networks, noting that achieving all three simultaneously presents substantial challenges for AI systems. De-Arteaga et al. (2022) emphasize that while fairness is crucial for legal compliance and social responsibility, unfair systems can threaten organizational survival and competitiveness, though they argue the utility-fairness trade-off assumption is often mistaken or short-sighted. Ryan et al. (2023) highlight practical implementation challenges, noting that fairness definitions can be contradictory or incompatible, creating complex engineering obstacles that require interdisciplinary collaboration between HCI and ML experts.

Regulatory and organizational alignment

Even as global frameworks such as the EU AI Act, OECD AI Principles, and UNESCO's Ethics of AI recommendation provide guidance on fairness and accountability, many firms struggle to translate these high-level principles into concrete practices. Morley et al. (2021) identify a critical gap between AI ethics principles and practical implementation, noting that existing translational tools are either too flexible (enabling ethics washing) or too rigid. Kelley (2022) identifies eleven organizational components necessary for effective AI principle adoption, including management support, enforcement mechanisms, and interdisciplinary approaches, highlighting the complexity of genuine ethical implementation. Within organizations, misalignment between compliance teams, data scientists, and business units further slows progress. For instance, compliance officers may emphasize transparency and auditability, while technical teams focus on predictive accuracy, leading to fragmented implementations. Overall, these challenges demonstrate that achieving fairness in financial AI is not solely a technical endeavour. Organizations must integrate data governance, algorithmic design, and compliance considerations into a holistic strategy, ensuring that bias mitigation efforts are both effective and sustainable in real-world financial systems.

Methodology

This paper employs a qualitative research design based on a systematic literature review (SLR) and thematic analysis, chosen to rigorously synthesize interdisciplinary insights on algorithmic fairness in financial AI systems.

Systematic Literature Review

Relevant studies were sourced from Google Scholar, Scopus, IEEE Xplore, and leading industry journals in information systems, finance, and computer science. We focused on publications published between 2015 and 2025, with particular attention to recent contributions (2021–2025) to capture state-of-the-art developments. To ensure replicability, the search strategy employed combinations of key terms such as:

"algorithmic bias" AND "financial services"

"algorithmic bias" AND "machine learning" OR "artificial intelligence"

"fairness in machine learning" AND "credit scoring"

"bias mitigation" AND "lending" OR "fraud detection"

"AI ethics" AND "financial regulation"

"fairness metrics" AND "machine learning" OR "artificial intelligence"

Only publications written in English were considered to maintain consistency and ensure accessibility for detailed analysis. The search process followed a structured pipeline:

- Initial retrieval of all articles based on keyword searches;
- Screening by title and abstract to exclude irrelevant works (e.g., purely technical papers without financial application, or studies outside the finance domain, such as healthcare-only bias research);
- Full-text review to identify studies that directly examined algorithmic fairness, bias, or mitigation techniques in the context of financial decision-making;
- Final inclusion of peer-reviewed journal articles, conference papers, and authoritative industry reports.

Thematic Coding and Synthesis

The final corpus of literature was imported into NVivo, a computer-assisted qualitative data analysis software that facilitated the systematic organization, coding, and visualization of qualitative data (Dhakal, 2022). NVivo supports multiple file formats and enables researchers to create thematic and case nodes, apply hierarchical coding, classify data with metadata, and visualize relationships across datasets. The analysis began with auto-coding and word frequency analysis. NVivo's auto-coding algorithm segments text at the paragraph level and identifies recurring lexical units based on semantic similarity; this process generated preliminary nodes representing key conceptual categories such as "data," "bias," "fairness," and "decision." A word frequency query was then run to detect high-salience terms across the dataset, using stemmed words and a minimum length threshold (typically >= 3 characters) to ensure linguistic precision. These outputs served as the foundation for initial code development. Following the automated stage, a manual coding refinement process was undertaken to improve contextual accuracy. NVivo allows direct inspection of coded references, enabling the researcher to merge overlapping nodes, rename concepts, and create parent-child hierarchies that capture conceptual depth (e.g., Bias as a parent node for Data Bias and Sampling Bias). Following this, an iterative review and cross-comparison process was conducted manually to identify overlaps and refine thematic alignment; this resulted in the consolidation of the 27 nodes into five higher-order themes that captured the dominant technical, organizational, and regulatory concerns across the literature (tables 1 and 2). For the final analysis, matrix coding queries and cluster analysis were employed to examine relationships between nodes. NVivo's matrix query function enabled us to cross-tabulate co-occurring themes (e.g., bias x data, fairness x decision), revealing how technical and ethical constructs intersect within the literature; this step directly supported the paper's aim of identifying socio-technical divides in AI ethics discourse within financial services, providing a robust foundation for our socio-technical framework.

Results

Overview of Data and NVivo Analytical Process

The qualitative analysis comprised 32 chosen scholarly and industry papers focusing on algorithmic fairness and bias in artificial intelligence within financial services. These sources were imported into NVivo 14, where automated thematic coding was initially executed to identify lexical and conceptual patterns across the dataset. Auto coding feature was applied at the paragraph level to ensure contextual coherence of extracted meaning units. The procedure produced 27 initial nodes, each representing recurring conceptual categories such as "data," "bias," "model," and "fairness" (see table 1). The 27 nodes generated via NVivo auto-coding highlight key concepts across the dataset. "Data" (32 files, 407 references), "fairness" (27 files, 401 references), and "model" (32 files, 359 references) were the most referenced, reflecting the centrality of technical design and equity concerns. Bias ("Bias", 27 files, 293 references) and group considerations ("groups", 26 files, 204 references) further emphasize fairness issues, while terms such as "process", "methods", and "algorithms" indicate attention to methodological rigor. "ethical", "predictive", and outcome-focused nodes reveal that discussions balance technical, ethical, and performance dimensions. Overall, node frequency demonstrates a dataset strongly focused on both the operational and normative aspects of AI-powered decision systems. Following an iterative review and comparison from the first author, these nodes were consolidated into five higher-level themes representing the dominant theoretical and operational concerns in the literature: (1) Data Nature and Sources of Bias; (2) Ethical and Regulatory Governance; (3) Mitigation Methods; (4) Fairness Metrics and Evaluation; and (5) Organisational Implementation. This hierarchical structure is visualised in NVivo's node diagram (figure 1), demonstrating the analytical progression from granular codes to synthesised constructs. Word frequency analysis and word cloud visualisation (figure 1, table 2) were further used to triangulate salience patterns across the dataset, thereby enhancing interpretive reliability.

Word Frequency and Conceptual Density

The word frequency analysis conducted in NVivo identified key terms that dominated the corpus, visualised through a word cloud (figure 1) and detailed in the top 10 ranked terms (table 2). The most recurrent words

included “fair” (n = 4945), “bias” (n = 3600), “model” (n = 3139), and “data” (n = 3082), followed by “using,” “learning,” “algorithms,” and “groups”.

Table.1 Frequency of 27 Automatically Generated Nodes

Node/ Keyword	No. of files mentioned	No. of references in text
data	32	407
fairness	27	401
model	32	359
bias	27	293
groups	26	204
ethical	17	173
process	31	155
decision	26	145
predictive	26	138
research	30	138
learning	28	130
use	31	130
training	26	128
approach	27	122
methods	24	114
algorithms	24	105
credit	21	101
systems	29	100
attributes	21	99
dataset	24	95
rate	25	92
specific	23	92
development	25	88
outcomes	24	87
metrics	22	86
risk	22	86
variable	23	85

(NVivo 14 Analysis, Author’s Own, 2025)

These terms collectively signal the conceptual centre of current discourse on algorithmic decision systems - balancing the technical mechanisms of data modelling and learning processes with normative imperatives of fairness and bias mitigation. The predominance of “fairness” and “bias” highlights the field’s continuing emphasis on achieving fairness in model behaviour while addressing the social implications of automated decision-making, while the prominence of “data” and “models” highlights the technical foundation of these debates. Meanwhile, supporting terms such as “algorithms,” “groups,” and “ethics” reveal that fairness discussions were not purely theoretical, but embedded in methodological and applied contexts. Together, these lexical patterns suggest that algorithmic fairness research is evolving as an interdisciplinary field situated between data-centric innovation and ethical governance.

Emergent Themes & Consolidation Logic

Following NVivo’s automated generation of 27 nodes, an iterative consolidation was conducted to produce analytically coherent higher-order themes aligned with the study’s focus. Nodes reflecting similar concepts or phenomena were merged to reduce redundancy while preserving interpretive nuance. Specifically, constructs describing data structures and inputs, including “data,” “dataset,” “training,” “variable,” and “attributes” were

grouped under Data Nature and Sources of Bias. Technical mechanisms, for example, "methods," "algorithms," or "approach", formed Mitigation Methods, while nodes referring to evaluation, such as "metrics," "fairness," were consolidated as Fairness Metrics and Evaluation. Operational practices, decision-making, and use cases were encompassed within Organisational Implementation, and oversight-related constructs, "ethical," "risk," "development," formed Ethical and Regulatory Governance. Manual inspection of coded references ensured semantic accuracy and provided a structured foundation for the study's analytical framework.

Table 2: Consolidation of 5 major themes from 27 Automatically Generated Nodes

No.	Theme	Original node
1	**Organisational implementation**	use
		process
		decision
		credit
2	**Mitigation methods**	systems
		methods
		learning
		approach
		algorithms
3	**Fairness metrics and evaluation**	rate
		predictive
		outcomes
		metrics
		fairness
4	**Ethical and regulatory governance**	specific
		risk
		research
		ethical
		development
5	**Data nature & sources of bias**	variable
		training
		model
		groups
		dataset
		data
		bias
		attributes

(NVivo 14 Analysis, Author's Own, 2025)

As shown in table 3, 'Data Nature and Sources of Bias' was the most frequently referenced theme (n = 1,670; approximately 40% of total references), underscoring the field's enduring focus on data provenance and bias root causes. This was followed by 'Fairness Metrics and Evaluation' (n = 804; 19%) and 'Ethical and Regulatory Governance' (n = 577; 13%), reflecting the literature's sustained attention to accountability mechanisms and evaluative frameworks. It is important to note that file counts are reported as N/A because NVivo aggregates overlapping sources across sub-nodes. A single paper may contain multiple references under the same thematic cluster, and summing file totals would therefore inflate the count. Consequently, the number of references was used instead, as it more accurately reflects how frequently and deeply each idea was discussed across the papers. Overall, the five themes show an uneven but comprehensive pattern of focus; with the greatest attention given to data-related fairness issues, while newer discussions are beginning to explore ethical governance and practical implementation. Each theme is discussed in detail below.

Data Nature and Sources of Bias

This theme was the most frequently referenced (n = 1,670; table 3), reflecting the field's strong focus on how data characteristics influence algorithmic outcomes. Key sub-nodes such as "dataset", "attributes", "variable," "training" and "bias" form a coherent cluster, showing that much of the literature concentrates on data-centric factors. The high reference counts indicate that issues related to data quality, preparation, and inherent biases remain central in discussions of fairness, highlighting that understanding and managing data is foundational to mitigating algorithmic bias.

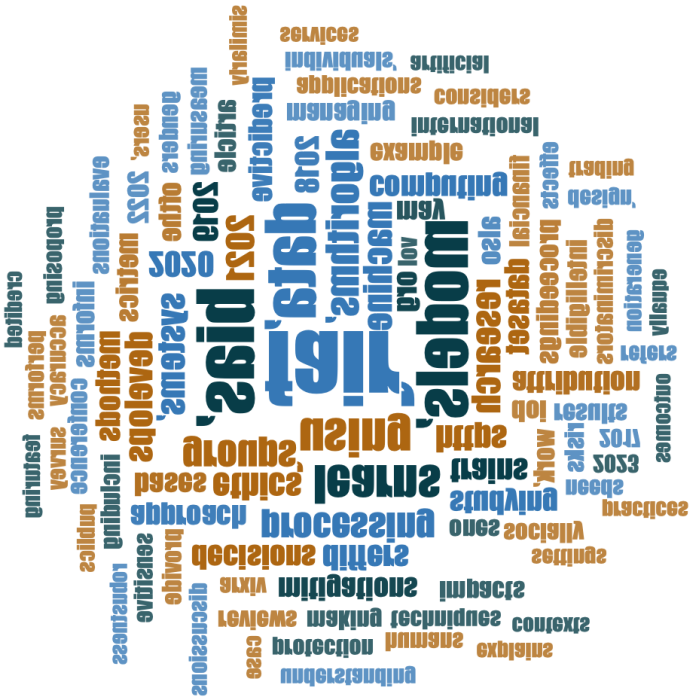

Figure 1 Word cloud illustration across 32 papers

(NVivo 14 Analysis, Author's Own, 2025)

Table 3 Top 10 Most Frequent Words

Word	Length	Count	Weighted Percentage (%)	Similar Words
fair'	5	4945	1,44	fair, fair', fair'', fairly, fairness, fairness'
bias'	5	3600	1,05	bias, bias', bias'', biased, biased', biases, biases', biasing, biasing'
models'	7	3139	0,91	model, modeled, modeler, modelers, modeling, modelled, modelling, models, models'
data'	5	3082	0,90	data, data', data'', data'14, data'18, datae, datas
using	5	2137	0,62	use, used, useful, usefully, usefulness, uses, using
learns	6	2104	0,61	learn, learned, learning, learning', learnings, learns
algorithms'	11	1686	0,49	algorithm, algorithmic, algorithmically, algorithms, algorithms'
groups'	7	1565	0,46	group, grouped, grouping, groupings, groups, groups'
processing	10	1441	0,42	process, processed, processes, processing, processing', processing'stage
machine	7	1348	0,39	machine, machine', machines

(NVivo 14 Analysis, Author's Own, 2025)

Ethical and Regulatory Governance

The second theme (n = 577 references) highlights the increasing attention to ethical principles and regulatory considerations in algorithmic systems. Sub-nodes such as "ethical," "risk," "specific," "development," and "research" cluster around these concerns, showing that the literature engages with ethics guidance and organisational risk management. The references suggest that scholars are connecting technical fairness discussions with broader ethical and compliance frameworks.

Mitigation Methods

This theme (n = 571 references) captures methodological strategies aimed at addressing bias. Nodes including "methods," "approach," "learning," "algorithms," and "systems" reveal that the literature discusses diverse interventions, ranging from algorithmic adjustments to procedural practices. The clustering indicates a focus on how technical solutions can be applied to improve fairness outcomes within computational models.

Fairness Metrics and Evaluation

The fourth theme (n = 804 references) concerns the operationalisation of fairness through measurable criteria. Sub-nodes such as "metrics," "rate," "predictive," "outcomes," and "fairness" show that much of the discourse revolves around quantifying and assessing fairness. The concentration of references suggests that evaluating

algorithmic outputs is a key concern, with scholars exploring multiple approaches to define and measure bias in practice.

Organisational Implementation

The final theme (n = 531 references) reflects practical challenges in embedding fairness into real-world settings. Nodes such as “use,” “process,” “decision,” and “credit” indicate that literature addresses organisational practices, workflow integration, and decision-making processes. The clustering of these sub-nodes suggests that successful implementation depends on aligning technical, ethical, and institutional considerations within organisations.

To further explore the relationships among key concepts, a cross-keyword co-occurrence analysis was also conducted in NVivo and examined how frequently two coded concepts appeared together within the same textual context across the dataset. The resulting matrix (table 4) reveals distinct patterns of association among six dominant keywords: bias, data, ethical, fairness, approach, and decision. The strongest co-occurrence was observed between bias and data (58), highlighting the central role of data quality, representativeness, and preprocessing in shaping algorithmic bias within financial systems. Similarly, data and fairness (48) frequently appeared together, suggesting that fairness considerations are often discussed in connection with data governance and validation processes. Moderate overlaps between fairness and decision (30), and between approach and fairness (29), indicate that fairness evaluation is also linked to decision-making procedures and bias mitigation strategies. In contrast, co-occurrences involving ethical were comparatively low across all pairs (ranging from 7 to 17), suggesting that ethical and regulatory dimensions tend to be addressed separately from technical or operational concerns; this finding points to a socio-technical disconnect, where ethical governance remains conceptually recognised but less embedded in day-to-day operational implementation. Overall, the co-occurrence analysis demonstrates a strong clustering of technical and operational concepts - particularly around bias, data, and fairness, while ethical considerations remain marginal. Such a distribution supports the subsequent development of our socio-technical framework (figure 3) that integrates technical, organisational, and regulatory dimensions to better align ethical intent with practical implementation.

Cross-Keyword Co-Occurrence Analysis

The framework visualizes AI systems as concentric layers, where technical components form the core, encircled by organizational structures, employee engagement, and regulatory guidance; this approach reflects the interplay between technological capabilities, human agency, and institutional contexts, acknowledging that ethical alignment is not a linear process but a multi-dimensional, iterative effort. The framework is motivated by evidence from the co-occurrence analysis, which revealed that technical and operational aspects - particularly bias, data quality, and fairness dominate the discourse, whereas ethical considerations often receive less integrated attention.

Figure 2 Hierarchical clustering of coded themes (NVivo 14 Analysis, Author's Own, 2025)

Table 4 Summary of major themes and sub-nodes

No.	Name	Files	References
1	**Organisational implementation**	**N/A**	**531**
	use	31	130
	process	31	155
	decision	26	145
	credit	21	101
2	**Mitigation methods**	**N/A**	**571**
	systems	29	100
	methods	24	114
	learning	28	130
	approach	27	122
	algorithms	24	105
3	**Fairness metrics and evaluation**	**N/A**	**804**
	rate	25	92
	predictive	26	138
	outcomes	24	87
	metrics	22	86
	fairness	27	401
4	**Ethical and regulatory governance**	**N/A**	**577**
	specific	23	92
	risk	22	86
	research	30	138
	ethical	17	173
	development	25	88
5	**Data nature & sources of bias**	**N/A**	**1670**
	variable	23	85
	training	26	128
	model	32	359
	groups	26	204
	dataset	24	95
	data	32	407
	bias	27	293
	attributes	21	99

(NVivo 14 Analysis, Author's Own, 2025)

Table 5 Cross-keyword co-occurrence analysis (NVivo 14 Analysis, Author's Own, 2025)

	A : bias	B : data	C : ethical	D : fairness	E : approach	F : decision
1 : bias	233	58	8	40	15	17
2 : data	58	320	17	48	26	27
3 : ethical	8	17	112	12	9	7
4 : fairness	40	48	12	304	29	30
5 : approach	15	26	9	29	100	7
6 : decision	17	27	7	30	7	130

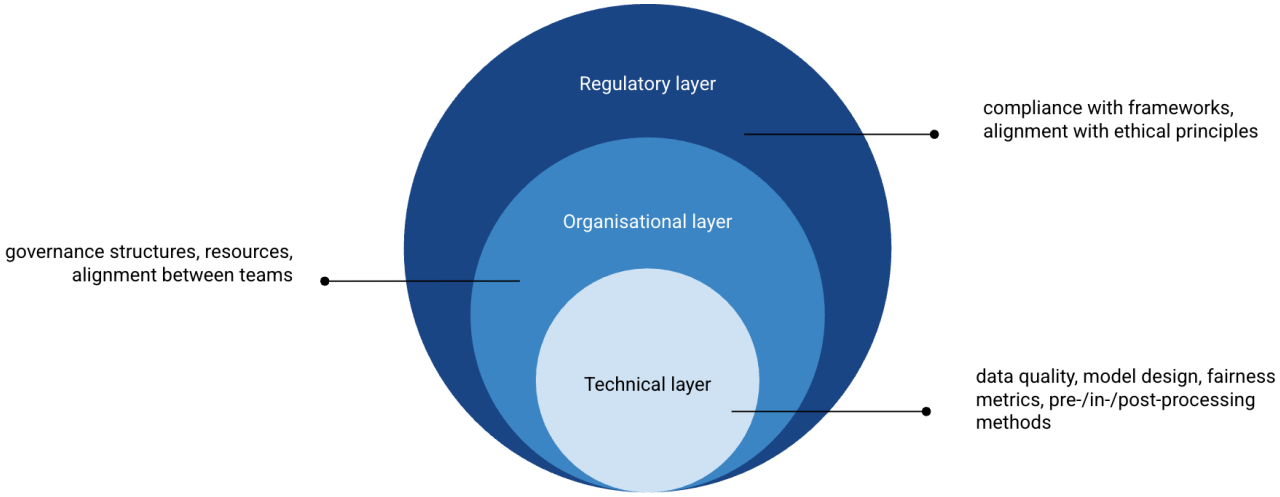

Figure 3 Socio-Technical Bias Mitigation Framework (Author's Own, 2025)

Evidence for this perspective is well-established. Sony et al. (2020) argue that technological systems, such as industry 4.0, are inherently socio-technical, integrating human and non-human elements to achieve shared objectives. This view is further reinforced by Stahl et al. (2023), who introduce the concept of "meta-responsibility" in AI ecosystems, highlighting that intelligent systems cannot be understood in isolation from their broader social context. Similarly, Lu et al. (2022) develop a responsible AI pattern catalogue spanning governance, processes, and design, illustrating the multi-level layer of technological systems and their institutional frameworks.

Discussion

Having conducted qualitative analysis on literature surrounding AI mitigation strategies in the financial domain, various results now enable us to dissect these as a discussion and the development of a framework. Our paper thus proposes a socio-technical layer framework (figure 3) to conceptualize ethical AI implementation within organizations.

Theme-by-Theme Analysis

Technical layer

At the centre of the framework lies the technical layer, which comprises AI system components directly responsible for shaping outputs. Key elements include data quality, module design, fairness metrics, and (pre/in/post-processing methods). Data quality underpins algorithmic reliability, with issues such as incomplete, biased, or unrepresentative datasets often driving inequitable outcomes - a principle explicitly confirmed by Shahbazi et al. (2022) through the "bias in, bias out" concept. Module design, including feature selection, model architecture, and interpretability mechanisms determines how well the AI system can function fairly under diverse conditions, with Hooker et al. (2021) emphasizing that model design itself can amplify algorithmic bias, not merely reflect existing data limitations. Fairness metrics operationalize ethical objectives by providing measurable standards, while pre- and post-processing interventions mitigate residual biases (Zhou et al., 2022; Trigo et al.,

2024). The technical layer embodies the principle that ethical AI begins with robust engineering practices. However, as suggested by the co-occurrence analysis, technical robustness alone does not ensure ethical integration; without organizational or human oversight, fairness and bias mitigation remain theoretical aspirations rather than actionable outcomes.

Organizational layer

Surrounding the technical core is the organisational layer, which mediates how fairness principles are enacted in practice. The moderate overlaps between "fairness" and "decision" (30) and between "fairness" and "approach" (29) (table 4) suggest that organisational processes - decision protocols, governance structures, and alignment strategies play a pivotal but secondary role in shaping fairness outcomes; this finding aligns with Morley et al. (2021), who emphasise that effective AI ethics requires institutional embedding clear roles, cross-team coordination, and accountability pathways that connect data scientists, compliance officers, and senior leadership. In many financial institutions, however, AI governance remains fragmented: fairness initiatives are often led by technical teams with limited integration into organisational oversight structures. The co-occurrence patterns in the dataset reflect this fragmentation, indicating that fairness is discussed in relation to decision-making but not consistently linked to broader organisational governance. Standards such as ISO/IEC 42001:2023 (AI Management Systems) formalise these principles, urging institutions to establish documented processes for risk assessment, model validation, and human oversight. Yet, as the findings imply, the translation of such frameworks into practice is uneven. Cultural and behavioural factors also appear in this layer. Organisational ethics depend not only on rules but also on shared values and competencies among employees. Training, communication, and leadership engagement shape whether fairness and accountability are internalised as everyday practice (Westover et al., 2024). Thus, while the technical layer operationalises fairness, the organisational layer determines its sustainability.

Regulatory layer

The outermost layer - the regulatory dimension represents formal compliance mechanisms and ethical alignment with societal norms. The empirical finding that "ethical" co-occurrences were low across all pairs (table 4) suggests that ethical and regulatory considerations are conceptually recognised but practically detached from day-to-day operations. This pattern resonates with Biddle's (2025) notion of ethics washing, where organisations symbolically endorse ethics and relegate responsibility to compliance departments rather than genuinely empowered. At this level, institutions are influenced by external governance frameworks such as the EU AI Act, OECD AI Principles, and other regulatory bodies. These regulations codify expectations of transparency, accountability, and human oversight. However, compliance alone does not guarantee ethical integrity. Research indicates that organizations frequently adopt ethical systems as a performative exercise, creating what Berjillos et al., (2025) terms a "tick-boxing" approach to ethics. Manning et al., (2020) further emphasizes that moving from a compliance-based to an integrity-based organizational climate requires more than superficial rule adherence Nevertheless, the regulatory layer is indispensable. It sets boundary conditions, ensuring minimum standards of accountability and offering a reference point for harmonisation across sectors. Importantly, it also exerts reflexive pressure on inner layers, motivating organisations to institutionalise governance and invest in technical controls. The challenge, as highlighted by the findings, is to move from rule-based compliance toward principle-based integration - embedding ethical reasoning throughout all layers of the framework.

Analysis of gaps, trade-offs, and tensions

The framework illuminates several gaps and tensions. First, the co-occurrence analysis revealed a clustering of technical and operational concepts, whereas ethical considerations were less integrated into daily discussion; this indicates a potential misalignment between ethical intentions and implementation, a gap the framework addresses by embedding ethical principles across layers. Second, trade-offs emerge between layers; technical solutions may optimize fairness metrics but conflict with organizational priorities, such as efficiency or resource constraints. Similarly, regulatory mandates may impose additional monitoring burdens, creating tensions between innovation speed and compliance. Recognizing these interdependencies encourages a holistic approach, where ethical alignment is negotiated across technical, organizational, human, and regulatory dimensions. Third, the framework identifies the risk of layer isolation. Technical teams, organizational managers, or regulatory officers operating independently can generate blind spots, resulting in ethical compromises. The layer structure emphasizes that no single layer is sufficient; integration and feedback across layers are critical for systemic ethical performance.

Implications for Implementation

The framework offers actionable guidance for institutions seeking to implement ethical AI. At the technical level, organizations should prioritize data governance, fairness metric selection, and bias mitigation strategies. Organizational practices should ensure cross-functional alignment, governance clarity, and adequate resource

allocation to operationalize technical interventions. Employee engagement programs including training, ethical awareness initiatives, and reporting mechanisms - support active oversight and accountability. Finally, organizations should maintain continuous engagement with regulatory developments to ensure compliance and foster trust. Importantly, the framework encourages iterative evaluation, where technical performance, organizational practices, and regulatory adherence are regularly reassessed to maintain ethical alignment. By adopting a layered perspective, institutions can anticipate potential tensions, address gaps proactively, and ensure that ethical considerations permeate all aspects of AI implementation.

Future Directions

While the framework provides a conceptual foundation, several avenues for future research emerge. First, empirical studies could examine the effectiveness of layered interventions across organizational contexts, assessing which combinations of technical, organizational, and human factors most effectively enhance ethical outcomes. Second, research could explore how regulatory evolution influences internal practices, and how organizations adapt to shifting legal and ethical landscapes. Third, studies could track the evolution of ethical alignment over time, exploring how financial organizations maintain responsible practices amid technological advancements, structural changes, and shifting societal expectations. Finally, comparative studies across banking, insurance, and fintech domains could highlight sector-specific challenges and best practices for integrating responsible AI principles into complex socio-technical ecosystems.

Conclusion

This study critically examined the state of algorithmic bias mitigation in AI-driven financial services through a systematic literature review and qualitative synthesis of 32 peer-reviewed and industry sources. The analysis confirms that while technical research on fairness has matured - offering an array of pre-, in, and post-processing techniques, the translation of these approaches into financial institutions remains constrained by organisational, regulatory, and resource barriers. Most research continues to privilege the technical dimensions of fairness, particularly data quality and model design, whereas ethical and governance aspects are often treated as peripheral considerations rather than integral components of system development. The socio-technical bias mitigation framework developed in this paper advances an integrative perspective by situating fairness within the interplay of technical, organisational, and regulatory layers. It underscores that achieving fairness is not a static engineering goal but an evolving organisational process requiring governance, monitoring, and human oversight. For the financial sector, this implies that effective mitigation of algorithmic bias demands embedding fairness considerations into data governance, model validation, and compliance infrastructures, supported by interdisciplinary collaboration among data scientists, ethicists, risk managers, and regulators. Future research should build on this framework by empirically testing fairness interventions in live financial environments, assessing their long-term impact on decision quality, regulatory compliance, and consumer trust. Moreover, comparative studies across jurisdictions could illuminate how differing regulatory cultures influence the operationalisation of ethical AI. Ultimately, the pursuit of fairness in financial AI systems represents not merely a technical challenge, but a defining test of institutional responsibility in an era of algorithmic decision-making.

References

- Arjunan, G 2022, “Enhancing data quality and integrity in machine learning pipelines: Approaches for Detecting and Mitigating Bias,” International Journal of Scientific Research and Management (IJSRM), 10(09):940–945, https://doi.org/10.18535/ijsrm/v10i9.ec04 (accessed 30 October 2025).
- ASIC 2024, ASIC warns governance gap could emerge in first report on AI adoption by licensees | ASIC, https://www.asic.gov.au/about-asic/news-centre/find-a-media-release/2024-releases/24-238mr-asic-warns-governance-gap-could-emerge-in-first-report-on-ai-adoption-by-licensees/ (accessed 20 August 2025).
- Barda, N, Yona, G, Rothblum, GN, Greenland, P, Leibowitz, M, Balicer, R, Bachmat, E & Dagan, N 2020, “Addressing bias in prediction models by improving subpopulation calibration,” Journal of the American Medical Informatics Association, 28(3):549–558, https://doi.org/10.1093/jamia/ocaa283 (accessed 02 September 2025).
- Biddle, JB, Nelson, JP & Olugbade, OE 2025, “How Can We Know if You are Serious? Ethics Washing, Symbolic Ethics Offices, and the Responsible Design of AI Systems,” Canadian Journal of Philosophy, 1–17, https://doi.org/10.1017/can.2025.9 (accessed 30 October 2025).
- Buchanan, BG & Wright, D 2021, “The impact of machine learning on UK financial services,” Oxford Review of Economic Policy, 37(3):537–563, https://doi.org/10.1093/oxrep/grab016 (accessed 14 August 2025).

- Caton, S & Haas, C 2024, "Fairness in Machine Learning: a survey," ACM Computing Surveys, 56(7) :1–38, https://doi.org/10.1145/3616865 (accessed 19 August 2025).
- Chen, F, Wang, L, Hong, J, Jiang, J & Zhou, L 2024, "Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models," Journal of the American Medical Informatics Association, 31(5):1172–1183, https://doi.org/10.1093/jamia/ocae060 (accessed 30 October 2025).
- Crockett, K, Colyer, E, Gerber, L & Latham, A 2021, "Building Trustworthy AI Solutions: A case for practical solutions for small businesses," IEEE Transactions on Artificial Intelligence, 4(4) :778–791, https://doi.org/10.1109/tai.2021.3137091 (accessed 02 September 2025).
- Cui, S, Pan, W, Zhang, C & Wang, F 2023, "Bipartite ranking fairness through a model agnostic ordering adjustment," IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11):13235–13249, https://doi.org/10.1109 /tpami.2023.3290949 (accessed 02 September 2025).
- De Castro Vieira, JR, Barboza, F, Cajueiro, D & Kimura, H 2025, "Towards Fair AI: Mitigating Bias in Credit Decisions—A Systematic Literature Review," Journal of Risk and Financial Management, 18(5) :228-257, https://doi.org/10.3390/jrfm18050228 (accessed 22 August 2025).
- De Los Ríos-Berjillos, A, Rodero-Cosano, ML, Ruiz-Lozano, M & Mesa-Pérez, E 2025, "The institutionalization of an ethical culture through the implementation of compliance systems," Corporate Social Responsibility and Environmental Management, https://doi.org/10.1002/csr.70017 (accessed 30 October 2025).
- De Vargas, VW, Aranda, JAS, Costa, RDS, Da Silva Pereira, PR & Barbosa, JLV 2022, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," Knowledge and Information Systems, 65(1):31–57, https://doi.org/10.1007/s10115-022-01772-8 (accessed 02 September 2025).
- De-Arteaga, M, Feuerriegel, S & Saar-Tsechansky, M 2022, "Algorithmic fairness in business analytics: Directions for research and practice," Production and Operations Management, 31(10) :3749–3770, https://doi.org/10.1111 /poms.13839 (accessed 02 September 2025).
- Dhakal, K 2022, "NVivo," Journal of the Medical Library Association JMLA, 110(2):, https://doi.org/10.5195/jmla. 2022.1271 (accessed 30 October 2025).
- EU Commission 2024, Annex III: High-Risk AI Systems referred to in Article 6(2) | EU Artificial Intelligence Act, https://artificialintelligenceact.eu/annex/3/ (accessed 22 August 2025).
- Ferrara, C, Sellitto, G, Ferrucci, F, Palomba, F & De Lucia, A 2023, "Fairness-aware machine learning engineering: how far are we?," Empirical Software Engineering, 29(1): Article 9, https://doi.org/10.1007/s10664-023-10402-y (accessed 03 September 2025).
- Fritz-Morgenthal, S, Hein, B & Papenbrock, J 2022, "Financial risk management and explainable, trustworthy, responsible AI," Frontiers in Artificial Intelligence, 5(1), 779799. https://doi.org/10.3389/frai.2022.779799 (accessed 15 August 2025).
- Gallegos, IO, Rossi, RA, Barrow, J, Tanjim, MM, Kim, S, Dernoncourt, F, Yu, T, Zhang, R & Ahmed, NK 2024, "Bias and Fairness in Large Language Models: A survey," Computational Linguistics, 50(3) :1097–1179, https:// doi.org/10.1162/coli_a_00524 (accessed 20 August 2025).
- González-Sendino, R, Serrano, E, Bajo, J & Novais, P 2023, "A review of Bias and Fairness in Artificial Intelligence," International Journal of Interactive Multimedia and Artificial Intelligence, 9(1) :5-17, https://doi.org/10.9781 /ijimai.2023.11.001 (accessed 22 August 2025).
- Guha, S, Khan, FA, Stoyanovich, J & Schelter, S 2024, "Automated data cleaning can hurt fairness in machine learning-based decision making," IEEE Transactions on Knowledge and Data Engineering, 36(12) :7368–7379, https://doi.org/10.1109/tkde.2024.3365524 (accessed 05 September 2025).
- Hajj, ME & Hammoud, J 2023, "Unveiling the Influence of artificial intelligence and machine learning on financial markets: A comprehensive analysis of AI applications in trading, risk management, and financial operations," Journal of Risk and Financial Management, 16(10) :1-16, https://doi.org/10.3390/jrfm16100434 (accessed 23 August 2025).
- Hall, P, Cox, B, Dickerson, S, Kannan, AR, Kulkarni, R & Schmidt, N 2021, "A United States Fair Lending Perspective on machine learning," Frontiers in Artificial Intelligence, 4, Article 695301, https://doi.org/10.3389/ frai.2021.695301 (accessed 24 August 2025).
- Hentzen, JK, Hoffmann, A, Dolan, R & Pala, E 2021, "Artificial intelligence in customer-facing financial services: a systematic literature review and agenda for future research," International Journal of Bank Marketing, 40(6) :1299–1336, https://doi.org/10.1108/ijbm-09-2021-0417 (accessed 12 August 2025).
- Hooker, S 2021, "Moving beyond 'algorithmic bias is a data problem,'" Patterns, 2(4) :100241, https://doi.org/10.1016/ j.patter.2021.100241 (accessed 03 September 2025).

- Huang, Y, Guo, J, Chen, W-H, Lin, H-Y, Tang, H, Wang, F, Xu, H & Bian, J 2024, "A scoping review of fair machine learning techniques when using real-world data," bioRxiv (Cold Spring Harbor Laboratory), https://doi.org/10.1101/2024.03.03.24303669 (accessed 03 September 2025).
- Johnson, GM 2020, "Algorithmic bias: on the implicit biases of social technology," Synthese, 198(10) :9941–9961, https://doi.org/10.1007/s11229-020-02696-y (accessed 12 August 2025).
- Jui, TD & Rivas, P 2024, "Fairness issues, current approaches, and challenges in machine learning models," International Journal of Machine Learning and Cybernetics, 15(8) :3095–3125, https://doi.org/10.1007/s13042-023-02083-2 (accessed 27 August 2025).
- Kelley, S 2022, "Employee perceptions of the effective adoption of AI principles," Journal of Business Ethics, 178(4) :871–893, https://doi.org/10.1007/s10551-022-05051-y (accessed 02 September 2025).
- Kim, J-Y & Cho, S-B 2022, "An information theoretic approach to reducing algorithmic bias for machine learning," Neurocomputing, 500, 26–38, https://doi.org/10.1016/j.neucom.2021.09.081 (accessed 12 August 2025) (accessed 05 September 2025).
- Kumar, IE, Hines, KE & Dickerson, JP 2022, "Equalizing Credit Opportunity in Algorithms," AAAI/ACM Conference on AI, Ethics, and Society (AIES 2022), 357–368, https://doi.org/10.1145/3514094.3534154 (accessed 02 September 2025).
- Li, J & Li, G 2024, "The Triangular Trade-off between Robustness, Accuracy and Fairness in Deep Neural Networks: A Survey," ACM Computing Surveys, https://doi.org/10.1145/3645088 (accessed 03 September 2025).
- Lu, Q, Zhu, L, Xu, X, Whittle, J, Zowghi, D & Jacquet, A 2023, "Responsible AI Pattern Catalogue: A collection of best practices for AI governance and engineering," ACM Computing Surveys, 56(7):1–35, https://doi.org/10.1145/3626234 (accessed 30 October 2025).
- Madaio, M, Egede, L, Subramonyam, H, Vaughan, JW & Wallach, H 2022, "Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support," Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1) :1–26, https://doi.org/10.1145/3512899 (accessed 12 August 2025)
- Manning, L 2020, "Moving from a compliance-based to an integrity-based organizational climate in the food supply chain," Comprehensive Reviews in Food Science and Food Safety, 19(3):995–1017, https://doi.org/10.1111/1541-4337.12548 (accessed 02 September 2025).
- Mehrabi, N, Morstatter, F, Saxena, N, Lerman, K & Galstyan, A 2019, "A survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, 54(6) :1–35, arXiv (Cornell University), https://arxiv.org/abs/1908.09635 (accessed 21 August 2025).
- Moldovan, D 2023, "Algorithmic decision making methods for fair credit scoring," IEEE Access, 11(1): 59729–59743, https://doi.org/10.1109/access.2023.3286018 (accessed 2 August 2025).
- Morley, J, Elhalal, A, Garcia, F, Kinsey, L, Mökander, J & Floridi, L 2021, "Ethics as a service: A pragmatic operationalisation of AI ethics," Minds and Machines, 31(2):239–256, https://doi.org/10.1007/s11023-021-09563-w (accessed 02 September 2025).
- Nazer, LH, Zatarah, R, Waldrip, S, Ke, JXC, Moukheiber, M, Khanna, AK, Hicklen, RS, Moukheiber, L, Moukheiber, D, Ma, H & Mathur, P 2023, "Bias in artificial intelligence algorithms and recommendations for mitigation," PLOS Digital Health, 2(6) :1-14, https://doi.org/10.1371/journal.pdig.0000278 (accessed 23 August 2025).
- Orphanou, K, Otterbacher, J, Kleanthous, S, Batsuren, K, Giunchiglia, F, Bogina, V, Tal, AS, Hartman, A & Kuflik, T 2022, "Mitigating bias in Algorithmic Systems—A fish-eye view," ACM Computing Surveys, 55(5) :1–37, https://doi.org/10.1145/3527152 (accessed 12 August 2025).
- Pagano, TP, Loureiro, RB, Lisboa, FVN, Peixoto, RM, Guimarães, G a. S, Cruz, GOR, Araujo, MM, Santos, LL, Cruz, M a. S, Oliveira, ELS, Winkler, I & Nascimento, EGS 2023, "Bias and Unfairness in Machine Learning Models: A Systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," Big Data and Cognitive Computing, 7(1), Article 15, https://doi.org/10.3390/bdcc7010015 (accessed 22 August 2025).
- Palladino, N 2022, "A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices," Telecommunications Policy, 47(5) :102479, https://doi.org/10.1016/j.telpol.2022.102479 (accessed 02 September 2025).
- Ryan, S, Nadal, C & Doherty, G 2023, "Integrating fairness in the software design Process: An interview study with HCI and ML experts," IEEE Access,11(1): 29296–29313, https://doi.org/10.1109/access.2023.3260639 (accessed 04 September 2025).
- Shahbazi, N, Lin, Y, Asudeh, A & Jagadish, HV 2023, "Representation Bias in Data: A survey on identification and resolution techniques," ACM Computing Surveys, 55(13s):1–39, https://doi.org/10.1145/3588433 (accessed 30 October 2025).
- Sony, M & Naik, S 2020, "Industry 4.0 integration with socio-technical systems theory: A systematic review and proposed theoretical model," Technology in Society, 61101248, https://doi.org/10.1016/j.techsoc.2020.101248.

- Stahl, BC 2023, “Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems,” Scientific Reports, 13(1):, https://doi.org/10.1038/s41598-023-34622-w (accessed 30 October 2025).
- Trigo, A, Stein, N & Belfo, FP 2024, “Strategies to improve fairness in artificial intelligence:A systematic literature review,” Education for Information, 40(3):323–346, https://doi.org/10.3233/efi-240045.
- Van Giffen, B, Herhausen, D & Fahse, T 2022, “Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods,” Journal of Business Research, 144, 93–106, https://doi.org /10.1016/j.jbusres.2022.01.076 (accessed 04 September 2025).
- Wan, M, Zha, D, Liu, N & Zou, N 2022, “In-Processing Modeling Techniques for Machine Learning Fairness: A survey,” ACM Transactions on Knowledge Discovery From Data, 17(3) :1–27, https://doi.org/10.1145/3551390 (accessed 02 September 2025).
- Westover, J 2024, “Preparing the workforce for ethical, responsible and trustworthy AI,” Human Capital Leadership., 13(4):, https://doi.org/10.70175/hclreview.2020.13.4.7 (accessed 30 October 2025).
- Xivuri, K & Twinomurinzi, H 2023, “How AI developers can assure algorithmic fairness,” Discover Artificial Intelligence, 3(1):Article 27, https://doi.org/10.1007/s44163-023-00074-4 (accessed 03 September 2025).
- Yang, J, Soltan, A a. S, Eyre, DW, Yang, Y & Clifton, DA 2023, “An adversarial training framework for mitigating algorithmic biases in clinical machine learning,” Npj Digital Medicine, 6(1):55, https://doi.org/10.1038/s41746-023-00805-y. (accessed 23 September 2025).
- Zhou, N, Zhang, Z, Nair, VN, Singhal, H & Chen, J 2022, “Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms,” International Statistical Review, 90(3) :468-480, https://doi.org/10.1111/insr.12492 (accessed 11 August 2025).