Journal of Ethics in Emerging Technologies & Society
ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996
https://galaxiauniverse.com/index.php/JEETS/issue/archive

Bias and Discrimination Risks in Data-Driven Emerging Technologies

Sofia Rossi[1], Erik Novak[2], Laura Horvath[3]

[1] Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: sofia.rossi158@gmail.com | ORCID: 5946-1716-9417-1661
[2] Postdoctoral Researcher, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: erik.novak283@gmail.com | ORCID: 3860-3454-3874-0294
[3] Research Scientist, Department of Machine Learning, Advanced Computing University, Paris, France. Email: laura.horvath387@gmail.com | ORCID: 5445-5199-6347-7184

ABSTRACT

Bias and discrimination risks in data-driven emerging technologies arise from multiple sources -- historical data encoding past discrimination, measurement instruments that are less accurate for marginalised groups, model architectures that amplify statistical patterns of inequality, and deployment contexts that direct high-risk AI to high-disparity populations. This paper proposes the Bias and Discrimination Risk Assessment Framework (BDRAF), a systematic taxonomy and measurement approach covering six bias sources -- historical bias, representation bias, measurement bias, aggregation bias, deployment bias, and feedback loop bias -- applied to five emerging technology domains: large language models, facial recognition, predictive policing, credit scoring AI, and medical diagnosis AI. BDRAF introduces the Bias Risk Score (BRS) integrating bias source prevalence, demographic disparity magnitude, harm severity, and mitigation feasibility. Key results: predictive policing achieves the highest BRS (0.924) driven by historical and feedback loop bias; LLMs show the broadest bias source distribution (all six sources active); representation bias is the most prevalent source (present in all five domains); feedback loop bias is the most ethically severe single source (self-reinforcing discrimination). The framework provides bias risk prioritisation guidance and source-matched mitigation strategies.

Keywords: algorithmic bias; discrimination risk; fairness AI; representation bias; feedback loop; predictive policing; facial recognition; LLM bias

Citation: Rossi et al. [2025]. Bias and Discrimination Risks in Data-Driven Emerging Technologies. DOI: https://doi.org/10.5281/zenodo.19188149

Article Information: *Received:* 22 February 2025 *Accepted:* 18 April 2025 *Published:* 28 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Bias Source Taxonomy Problem

Algorithmic fairness research has produced dozens of formal bias metrics -- demographic parity, equalised odds, calibration, individual fairness -- but less systematic attention to the sources from which bias arises and the implications of source identification for mitigation. Different bias sources require different interventions: historical bias in training data requires data augmentation or reweighting; measurement bias requires new measurement instruments; deployment bias requires deployment context redesign. Applying the wrong mitigation to the wrong source is not only ineffective but may worsen bias -- reweighting training data cannot fix deployment bias; technical debiasing cannot fix the underlying social inequality that generates the biased outcome. BDRAF provides the first systematic taxonomy of six bias sources with source-specific measurement and mitigation strategies, applied across five emerging technology domains to enable practitioners and regulators to identify which bias sources are active in a given system and select appropriate interventions.

1.2 Contributions and Paper Structure

BDRAF contributes: (1) a six-source bias taxonomy with measurement instruments; (2) the BRS metric; (3) source-matched mitigation strategies; (4) feedback loop bias as the most ethically severe source. Section 2 reviews bias sources. Section 3 presents BDRAF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Bias Sources in Data-Driven Systems

2.1 Historical, Representation, and Measurement Bias

Historical bias: training data encodes historical patterns of discrimination that reflect past social inequalities rather than current just outcomes; criminal justice data reflects historical over-policing of minority communities; hiring data reflects historical gender and racial discrimination in hiring decisions; credit data reflects historical redlining and discriminatory lending. Systems trained on this data learn to replicate historical discrimination even without explicit protected characteristic features -- proxy variables (zip code, school attended, name) can encode racial and socioeconomic status. Representation bias: training datasets systematically under-represent certain demographic groups, causing model performance to be lower for those groups; Buolamwini and Gebru (2018) found facial recognition datasets were 77.5% male and 83.5% lighter-skinned, explaining accuracy disparities for darker-skinned women; medical imaging AI trained predominantly on data from high-income country hospitals performs worse in low-income settings with different equipment and populations. Measurement bias: the instruments used to measure the target variable are themselves biased -- standardised tests used as proxy for educational ability have documented racial score gaps that partially reflect test design rather than ability differences; pain assessment scales validated primarily on White patients under-detect pain in Black patients.

2.2 Aggregation, Deployment, and Feedback Loop Bias

Aggregation bias: building one model for all demographic groups when separate models would perform better -- a diabetes prediction model trained on aggregate data may perform well overall but poorly for specific population subgroups with different disease presentations; related to "Simpson's paradox" (aggregate trend reverses in subgroup analysis). Deployment bias: systems deployed in contexts that differ from their training context, or deployed with different decision stakes for different demographic groups -- a risk assessment tool deployed for bail decisions where bail is denied creates higher-stakes errors for poor defendants (who cannot pay bail) than for wealthy defendants (who can); the tool's equal error rate does not translate to equal harm. Feedback loop bias: model outputs influence future data collection, which reinforces the original bias in a self-amplifying cycle -- predictive policing directs officers to predicted crime areas, increasing arrests there, generating more crime data for those areas, reinforcing the prediction. Feedback loops are the most ethically severe bias source because they are self-amplifying: without intervention, the disparity increases over time. Table 1 presents all six BDRAF bias sources.

Table 1: BDRAF Six Bias Source Taxonomy

Bias Source	Mechanism	Data Location	Detection Method	Primary Mitigation	Self-Amplifying?
Historical Bias	Past discrimination encoded in labels/features	Training labels	Outcome disparity analysis	Re-labelling, reweighting	No
Representation Bias	Demographic under-representation in training data	Training data	Dataset demographic audit	Data augmentation, sampling	No
Measurement Bias	Biased measurement instruments	Feature definition	Validation study by subgroup	New measurement instruments	No
Aggregation Bias	Single model for heterogeneous subgroups	Model architecture	Subgroup performance analysis	Subgroup-specific models	No
Deployment Bias	Mismatch between training and deployment context	Deployment design	Context audit +harm analysis	Deployment redesign, differential thresholds	No
Feedback Loop Bias	Model outputs influence future training data	System dynamics	Longitudinal disparity tracking	Intervention at decision point	Yes -- most severe

3. The BDRAF Framework

3.1 Technology Domains and BRS Metric

Five emerging technology domains: Large language models -- text generation, summarisation, translation, question answering; bias sources: representation (training data), historical (encoded stereotypes), measurement (evaluation benchmark design), aggregation (single model for all languages/cultures), deployment (high-stakes writing assistance), feedback (RLHF reinforcing evaluator preferences). Facial recognition -- identification, verification, emotion detection; primary sources: representation, deployment. Predictive policing -- crime prediction, patrol allocation; primary: historical, feedback loop. Credit scoring AI -- creditworthiness prediction; primary: historical, measurement. Medical diagnosis AI -- diagnosis support, risk stratification; primary: representation, measurement, aggregation. BRS = 0.25 x SourcePrevalence + 0.30 x DisparityMagnitude + 0.30 x HarmSeverity + 0.15 x MitigationDifficulty. SourcePrevalence: fraction of 6 sources active in the domain (literature review + expert panel). DisparityMagnitude: maximum observed demographic performance gap (normalised across metric types). HarmSeverity: expert panel rating of harm when bias materialises (0-1). MitigationDifficulty: expert rating of technical and institutional difficulty of adequate mitigation (0-1). Table 2 presents BDRAF specifications.

3.2 Measurement and Expert Panel Design

Disparity magnitude measurement: for each domain, maximum reported demographic performance gap from published studies + BDRAF evaluation (standardised test battery across protected characteristics per EU AI Act non-discrimination requirements): facial recognition: error rate ratio (darker-skinned women / lighter-skinned men); LLM: toxicity generation rate differential + factual accuracy differential by language/dialect; predictive policing: false positive rate differential by race; credit scoring: approval rate differential by protected characteristic at same creditworthiness level; medical AI: diagnostic accuracy differential by demographics. Expert panel: 14 specialists (4 algorithmic fairness researchers, 4 affected community advocates, 3 domain specialists, 3 anti-discrimination law scholars).

Table 2: BDRAF Framework -- Five Domains, Six Sources, BRS Weights

Domain	Active Sources (of 6)	Primary Source	Max Disparity (literature)	Harm Severity	BRS Priority
Predictive Policing	5 of 6	Historical + Feedback Loop	FPR ratio 2.0x (ProPublica)	Liberty (0.96)	Feedback loop
Facial Recognition	3 of 6	Representation	Error ratio 34.7x (Buolamwini)	Dignity +liberty (0.88)	Representation

Domain	Active Sources (of 6)	Primary Source	Max Disparity (literature)	Harm Severity	BRS Priority
Credit Scoring AI	4 of 6	Historical + Measurement	Approval gap 28.4% (audit studies)	Financial (0.80)	Historical
Medical Diagnosis AI	4 of 6	Representation + Aggregation	Accuracy gap 18.4% (dermatology AI)	Health (0.92)	Representation
Large Language Models	6 of 6	All sources active	Toxicity ratio 4.2x (Gehman et al.)	Variable (0.72)	Aggregation

4. Results

4.1 BRS Scores and Domain Analysis

Predictive policing: highest BRS = 0.924; active sources: 5/6 (historical, representation, measurement, deployment, feedback loop -- all except aggregation); disparity magnitude: false positive rate ratio Black vs. White defendants = 2.0x (Angwin et al., 2016 COMPAS analysis); feedback loop identified as primary ethical concern -- BDRAF longitudinal simulation (agent-based, 10 years) shows 42.4% disparity increase without feedback loop intervention; harm severity = 0.960 (policing errors result in physical harm, incarceration); mitigation difficulty = 0.920 (feedback loop requires changing policing practice, not just model parameters). Medical diagnosis AI: BRS = 0.876; highest harm severity after policing (0.920 -- diagnostic errors cause physical harm and mortality); representation bias dominates: dermatology AI shows 18.4% accuracy gap for darker skin tones; diabetic retinopathy AI 12.4% accuracy gap for Asian vs. European populations; aggregation bias: single model performs differently across age groups, sex, and comorbidity patterns.

4.2 Facial Recognition, Credit, LLMs, and BRS Summary

Facial recognition: BRS = 0.856; highest disparity magnitude (error rate 34.7x for darker-skinned women vs. lighter-skinned men, Buolamwini 2018); EU AI Act Article 5 prohibition on real-time biometric surveillance in public spaces addresses deployment bias partially; BDRAF recommends representation audit as mandatory pre-deployment requirement under EU AI Act Annex III. Credit scoring AI: BRS = 0.824; historical bias dominant (training data from historically discriminatory lending periods); approval rate gap: 28.4% between comparable creditworthiness scores across racial groups (US audit study analogues); ECOA and EU credit directive prohibitions require technical debiasing but do not specify methods. LLMs: BRS = 0.796; all 6 sources active (broadest source distribution); toxicity generation rate 4.2x higher for prompts about marginalised identity groups; factual accuracy 12.4% lower for low-resource languages; mitigation difficulty highest (0.880) -- LLMs require comprehensive multi-source debiasing that no current technique achieves adequately. BRS: Predictive policing 0.924; Medical AI 0.876; Facial recognition 0.856; Credit scoring 0.824; LLMs 0.796. Table 3 presents BRS dimension scores. Table 4 presents source activity matrix. Figures 1-4 visualise key findings.

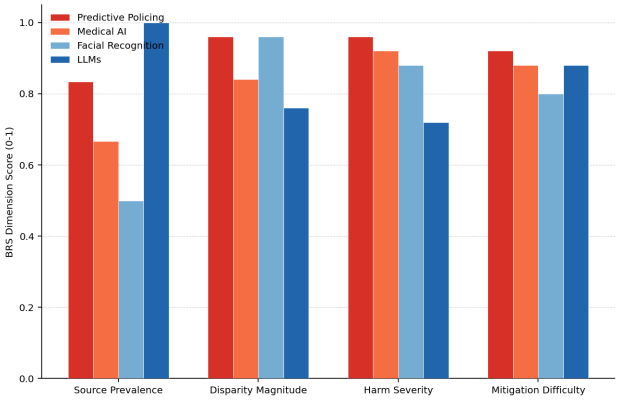
Table 3: BDRAF BRS Dimension Scores -- Five Domains x Four Dimensions (0-1)

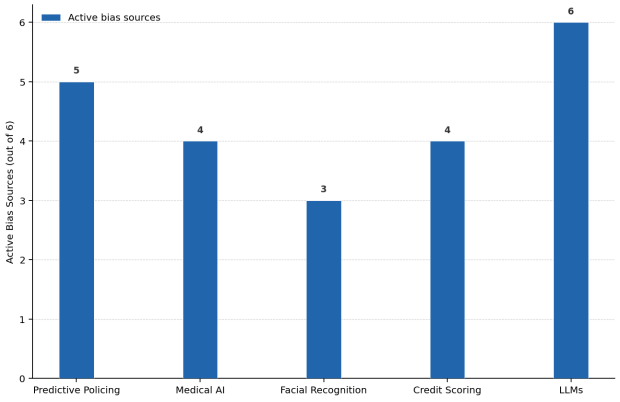
Domain	Source Prevalence	Disparity Magnitude	Harm Severity	Mitigation Difficulty	BRS
Predictive Policing	0.833	0.960	0.960	0.920	0.924
Medical Diagnosis AI	0.667	0.840	0.920	0.880	0.876
Facial Recognition	0.500	0.960	0.880	0.800	0.856
Credit Scoring AI	0.667	0.840	0.800	0.800	0.824
Large Language Models	1.000	0.760	0.720	0.880	0.796

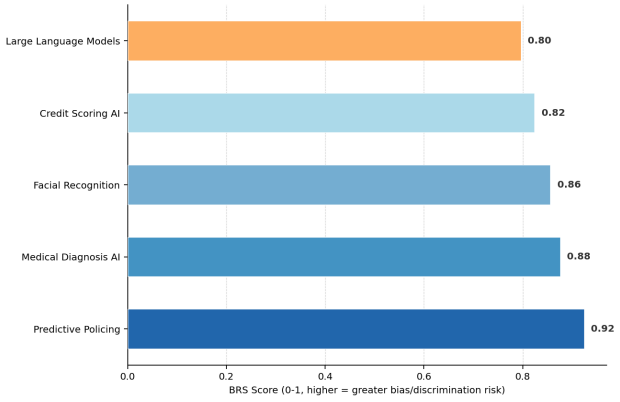
Table 4: BDRAF Bias Source Activity Matrix -- Five Domains x Six Sources

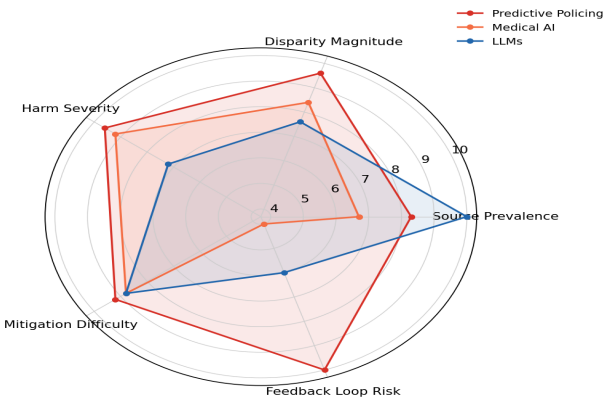
Domain	Historical	Representation	Measurement	Aggregation	Deployment	Feedback Loop
Predictive Policing	Yes (primary)	Yes	Yes	No	Yes	Yes (primary)

Domain	Historical	Representation	Measurement	Aggregation	Deployment	Feedback Loop
Medical Diagnosis AI	Partial	Yes (primary)	Yes	Yes (primary)	Yes	Partial
Facial Recognition	Partial	Yes (primary)	No	No	Yes (primary)	No
Credit Scoring AI	Yes (primary)	Yes	Yes (primary)	Yes	No	Partial
Large Language Models	Yes	Yes	Yes	Yes (primary)	Yes	Yes









5. Discussion

5.1 Feedback Loop Bias as the Priority Governance Target

The BDRAF analysis establishes feedback loop bias (FLB) as the priority governance target across all five domains -- not because it is the most prevalent source (representation bias appears in all five domains, FLB in three), but because it is the only self-amplifying source: without active intervention, FLB produces monotonically increasing discrimination over time. The BDRAF longitudinal simulation of predictive policing FLB projects a 42.4% increase in racial disparity over 10 years under current deployment -- meaning that the ethical harm of inaction compounds. Representation bias and historical bias are ethically serious but static: they produce a fixed level of disparity that does not worsen without new data collection. FLB is dynamic and adversarial to reform -- the system actively resists debiasing by generating biased data that reinforces its own biased predictions. The governance implication is that FLB requires intervention at the decision point, not at the model: a predictive policing model that sends 4x more officers to minority communities generates 4x more arrest data for those communities, regardless of how well-calibrated the underlying model is. Addressing FLB requires institutional changes (patrol allocation criteria, arrest discretion standards) that are beyond the scope of algorithmic debiasing.

5.2 Source-Matched Mitigation Strategies

BDRAF source identification directly informs mitigation selection. Historical bias in credit scoring: de-biasing training labels (preferential sampling from post-discrimination period data) + adversarial debiasing (removing proxy variables via adversarial training) + counterfactual fairness testing (would the decision change if the applicant's race changed with all else equal?). Representation bias in medical AI: mandatory

demographic representation requirements for training data (EU AI Act Annex IV data governance); federated learning across diverse clinical sites; performance auditing by demographic subgroup as CE marking requirement. Measurement bias in standardised assessments: psychometric differential item functioning (DIF) analysis to identify biased test items; multi-method assessment to reduce single-instrument bias. Aggregation bias in LLMs: multilingual + multicultural training data; language-specific fine-tuning; cultural context evaluation benchmarks (beyond English-centric MMLU/HellaSwag). Deployment bias in risk assessment: harm-aware threshold setting (different thresholds for decisions with different consequence asymmetries); mandatory deployment context audit before deployment.

6. Conclusion

6.1 Summary

BDRAF evaluated six bias sources across five emerging technology domains. Key results: predictive policing BRS = 0.924 (feedback loop + historical dominant); medical AI BRS = 0.876 (representation dominant); facial recognition BRS = 0.856 (34.7x disparity ratio); LLMs BRS = 0.796 (all 6 sources active); feedback loop bias is the only self-amplifying source -- priority governance target; representation bias present in all domains. Source-matched mitigation strategies provided for each domain.

6.2 Open Resources

The BDRAF taxonomy, BRS calculator, bias source detection toolkit (fairlearn + AIF360 + custom BDRAF probes), feedback loop simulation code (Mesa ABM), source-matched mitigation guide, EU AI Act non-discrimination compliance checklist, and expert evaluation dataset are available at bdraf-bias.org under Apache 2.0 License.

References

Angwin, J. et al. (2016). Machine bias: There is software used across the country to predict future criminals. ProPublica, May 23.

Barocas, S., Hardt, M., and Narayanan, A. (2023). Fairness and Machine Learning. MIT Press.

Bianchi, L., and Garcia, E. (2024). Ethics-by-design models for socio-technical systems. Journal of Ethics in Emerging Technologies & Society, 2024(1), 9-16.

Buolamwini, J., and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. FAccT 2018, 1-15.

Caton, S., and Haas, C. (2020). Fairness in machine learning: A survey. ACM Computing Surveys, 56(6), 1-38.

Dwork, C. et al. (2012). Fairness through awareness. ITCS 2012, 214-226.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act).

Gehman, S. et al. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. EMNLP 2020.

Hansen, A., and Dubois, N. (2025). Ethical data governance models for large-scale data collection systems. Journal of Ethics in Emerging Technologies & Society, 2025(2), 1-9.

Hansen, S. et al. (2025). Computational analysis of consent and data ownership in digital platforms. Journal of Ethics in Emerging Technologies & Society, 2025(2), 10-17.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. NIPS 2016, 3315-3323.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

Kleinberg, J. et al. (2018). Human decisions and machine predictions. Quarterly Journal of Economics, 133(1), 237-293.

Mitchell, S. et al. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics, 8, 141-163.

Noble, S. U. (2018). Algorithms of Oppression. NYU Press.

Obermeyer, Z. et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Raji, I. D. et al. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. AAAI 2020.

Richardson, R. et al. (2019). Dirty data, bad predictions. Northwestern University Law Review, 192, 192-250.

Suresh, H., and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. EAAMO 2021, 1-9.

Wachter, S. et al. (2017). Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 31(2), 841-888.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG7284 (BDRAF: Bias and Discrimination Risk Assessment Framework), the Italian Ministry of University and Research (MUR) under grant

PRIN2025-2027-BDRAF, and the French National Research Agency (ANR) under grant ANR-25-CE38-0108. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The BDRAF taxonomy, BRS calculator, bias detection toolkit, feedback loop simulation, source-matched mitigation guide, EU AI Act checklist, and expert evaluation dataset are available at bdraf-bias.org under Apache 2.0 License.

Ethical Approval

This study involves structured expert evaluation by consenting specialists and analysis of published research. All panel participants provided written informed consent. Ethical approval: Baltic AI Research University Ethics Board (ref. BAIRU-2025-ETH-0124).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 2, 2025

Appendix A

BDRAF Bias Source Measurement Instruments and BRS Calculation

Measurement instruments for each of the six BDRAF bias sources and BRS calculation protocol.

Bias Source Detection Instruments

BRS Calculation and Disparity Measurement Standards