Discover Artificial Intelligence

RESEARCH Open Access

Dual erosion of equality and remedy in algorithmic decision systems

Bhavya Johari[1*]

*Correspondence:
Bhavya Johari
bhavya.johari@jgu.edu.in
[1]Jindal Global Law School, O. P. Jindal Global University, Sonipat, India

Abstract
This article develops a dual erosion framework demonstrating how Artificial Intelligence systems simultaneously undermine equality rights through discriminatory outcomes and obstruct effective remedies through opacity. Through comparative case synthesis spanning criminal justice algorithms in the United States and Canada, welfare eligibility systems in the European Union (EU) and India, credit scoring in Germany, employment screening, and content moderation, the analysis reveals that equality risk and remedy risk operate interdependently rather than in parallel. Biased algorithmic outputs become unreviewable due to opacity, while opacity enables bias to persist undetected. The article examines regulatory responses under the EU Artificial Intelligence Act, the Digital Services Act, and the revised Santa Clara Principles, proposing concrete implementation frameworks that include discovery protocols reconciling trade secrets with due process, burden-shifting standards adapted from employment discrimination law, risk-tiered due diligence requirements, independent audit mechanisms, and participatory governance models. These findings advance algorithmic accountability scholarship by providing adjudicable standards for courts and regulators, while demonstrating that comprehensive governance requires integrated approaches that address both equality and remedy dimensions.

Keywords Algorithmic accountability, Artificial intelligence governance, Discrimination law, Due process rights, Effective remedy, Algorithmic transparency, Digital rights regulation

1 Introduction

Artificial intelligence (AI) systems now determine access to employment, credit, housing, social benefits, and freedom itself through bail and sentencing algorithms. These systems make consequential decisions affecting millions of individuals, yet they operate through processes fundamentally different from traditional human decision-making. This article examines whether AI-driven decision-making presents unique challenges to fundamental rights frameworks and, if so, what legal reforms might be necessary to preserve human dignity in algorithmic governance.

Existing scholarship largely analyses algorithmic harms through either an equality lens, focusing on discriminatory outcomes [8, pp. 677–680], [15, pp. 153–154], or a

procedural lens, emphasising due process deficits [16, pp. 1278–1280], [62, pp. 1346–1348]. Contemporary AI fairness research demonstrates that this separation inadequately captures how algorithmic systems operate in practice. Comprehensive technical reviews document that algorithmic bias arises from interdependent sources, including training data reflecting historical discrimination, non-representative datasets producing differential accuracy across populations, and feature engineering creating proxies for protected characteristics [35, pp. 4–7], [43, pp. 5–9]. More fundamentally, empirical evaluation of bias mitigation techniques reveals structural limitations. Testing 17 bias mitigation methods across multiple scenarios shows that technical interventions improve fairness in only 46% of cases, while simultaneously degrading both fairness and performance in 25% of cases [14, pp. 3, 17–19]. These limitations stem from mathematical impossibility results. When base rates differ across demographic groups, algorithms cannot simultaneously achieve calibration, equal false-positive and false-negative rates, and equal false-positive and false-negative rates [15, pp. 155–157], [37, pp. 1–2, 4–5].

Satisfying one fairness criterion mathematically precludes satisfying others, requiring normative choices about which definition of fairness to prioritize. Nielsen [47] demonstrates that the AI fairness movement has uncritically adopted U.S. antidiscrimination law as both a normative foundation and a technical framework, failing to address how categorical approaches exclude intersectional harms and non-categorical forms of disadvantage. This technical literature supports what comparative legal analysis reveals. AI systems simultaneously entrench discrimination through data dependencies while frustrating remedies through opacity. These two dimensions interact dynamically rather than operating independently. Biased algorithmic outputs become unreviewable because of opacity, while opacity enables bias to persist undetected and uncorrected.

This article formalises this observation into an analytical framework, distinguishing equality risk from remedy risk. The equality risk dimension examines how training data bias, non-representative datasets, and discriminatory deployment contexts systematically disadvantage protected groups. The remedy risk dimension analyses how algorithmic opacity, evidential asymmetries, and inadequate procedural safeguards obstruct effective redress. Applying this framework to high-stakes decision systems across multiple jurisdictions reveals patterns invisible to single-axis analysis and generates concrete prescriptions for reform.

The research question is whether this dual erosion framework, when applied to algorithmic decision-making across criminal justice, welfare administration, employment, and content moderation, reveals systemic inadequacies in current legal frameworks and, if so, what concrete reforms might address them. The analysis proceeds through four stages. First, it establishes a methodological approach that combines doctrinal legal analysis with comparative case synthesis. Second, it formalises the dual erosion framework specifying the dimensions and interactions between equality risk and remedy risk. Third, it applies this framework systematically to case studies spanning the United States, Canada, the European Union, India, and Germany. Fourth, it translates findings into implementable legal instruments, including model statutory language, discovery protocols, burden-shifting frameworks, and risk-tiered due diligence standards.

This approach contributes to the field of algorithmic accountability scholarship in three ways. First, it demonstrates that equality and remedy are not parallel but interdependent concerns that require an integrated analysis. Second, it extends beyond

predominantly United States (U.S.)-focused literature by examining welfare algorithms in the Netherlands and India, Denmark's fraud detection systems, and Germany's credit scoring jurisprudence, revealing how different legal traditions confront similar challenges. Third, it moves from diagnosis to prescription by providing concrete, adjudicable standards that courts and regulators can apply, addressing critiques that algorithmic accountability scholarship remains abstract [46, pp. 780–781].

This analysis employs key terms from both legal frameworks and AI fairness scholarship. Legal definitions establish the regulatory obligations governing algorithmic decision-making, while technical definitions from AI fairness scholarship provide precision about how these systems generate discriminatory outcomes that trigger those obligations.

AI systems are machine-based systems designed to operate with varying levels of autonomy, inferring how to generate outputs such as predictions, recommendations, or decisions, as defined in Article 3(1) of the European Union (EU) AI Act (2024). **High-risk systems** are those deployed in domains where errors produce severe consequences for fundamental rights, including criminal justice, welfare eligibility, employment screening, credit access, and content moderation affecting speech rights, as identified in Annex III of the EU AI Act (2024). **Effective remedy** denotes legally enforceable mechanisms enabling affected individuals to understand, challenge, and potentially reverse adverse determinations, incorporating both procedural access and substantive review as required under Article 2(3) of the International Covenant on Civil and Political Rights (1966) and interpreted by the Human Rights Committee [32 paras. 15–17].

Transparency requires disclosure of sufficient information about system logic, data sources, and decision-making processes to enable meaningful scrutiny, as mandated by General Data Protection Regulations (GDPR) [51 Articles 13(2)(f), 14(2)(g), and 15(1)(h) for automated decision-making producing legal effects or similarly significant consequences under Article 22(1). Lastly, **Contestability** refers to the practical ability of affected individuals to dispute algorithmic determinations through accessible processes that provide genuine human review and reconsideration, as required by the GDPR (2016), Article 22(3), and Article 14 of the EU AI Act (2024).

These legal requirements operate against the backdrop of technical concepts from AI fairness scholarship that define how algorithmic systems produce discriminatory outcomes and why regulatory intervention becomes necessary. **Algorithmic harm** refers to adverse impacts on individuals or groups resulting from algorithmic decision-making. The AI fairness literature distinguishes allocative harms, which occur when algorithms deny opportunities or resources, from representational harms, which occur when algorithms reinforce demeaning stereotypes or fail to recognize group identities [59, p. 2]. **Algorithmic equality** distinguishes between disparate treatment, which occurs when decision-making processes classify individuals according to protected characteristics and differentiate treatment on that basis (equality of treatment), and disparate impact, which arises when facially neutral algorithmic practices produce disproportionately adverse outcomes for protected groups (equality of outcome) [8, pp. 694–695, 701–702]. An additional dimension, equality of opportunity, encompasses competing normative visions: treating currently similar individuals similarly,enabling individuals of similar ability and ambition to achieve similar outcomes despite historical disadvantage; or discounting current dissimilarity that results from past injustice [9, pp. 81–85]. Each

conception generates distinct technical requirements for algorithmic systems and distinct implications for remediation when those systems fail.

Algorithmic bias refers to the systematic, repeatable errors that produce unfair outcomes across demographic groups. Bias arises from training data that incorporate historical discrimination patterns, non-representative datasets that yield differential accuracy across populations, or feature selection that creates proxies for protected characteristics [43, pp. 3–4, 7–9].

Fairness metrics in machine learning formalize normative criteria through mathematical constraints. **Demographic parity** requires equal selection rates across groups regardless of base rates (the actual frequency of the outcome within each group). **Equalized odds** require equal true-positive rates (correctly identifying qualified individuals) and false-positive rates (incorrectly identifying unqualified individuals as qualified) across protected groups, imposing stricter requirements than demographic parity. **Calibration** requires that risk scores correspond to empirically observed outcome rates: individuals assigned the same score should, on average, experience the outcome at that rate, regardless of group membership [15, pp. 154–155].

These metrics prove mathematically incompatible when base rates differ across groups, necessitating value judgments about which fairness criterion to prioritize [37, pp. 4–5, 17]. This incompatibility demonstrates that algorithmic fairness involves normative choices that technical optimization cannot resolve.

2 Methodology

This article employs doctrinal legal analysis combined with comparative case synthesis. Doctrinal analysis examines legal rules, principles, and concepts through systematic interpretation of statutes, case law, and regulatory instruments [33, pp. 84–85]. The method synthesises facts, legal thoughts, and principles to construct coherent frameworks for understanding how law operates [11, pp. 148–152]. This approach is appropriate for algorithmic accountability research because it clarifies how existing legal categories apply to novel technologies and identifies doctrinal gaps that require reform.

The comparative dimension employs a case synthesis methodology, which combines the depth of case studies with the breadth of comparison [10, p. 7]. Case selection criteria prioritised algorithmic systems deployed in high-stakes domains (criminal justice, welfare, employment, and credit) where decisions produce legal effects or similarly significant consequences, as defined in Article 22 (1) of the GDPR (2016). Geographic diversity was maximised deliberately to test whether the dual erosion framework transcends specific legal traditions. Cases were selected based on three factors. First, the availability of authoritative documentation through court decisions, regulatory findings, or peer-reviewed investigations. Second, representation of distinct algorithmic applications spanning risk assessment, eligibility determination, and automated content moderation. Third, variation in regulatory response enables comparison of different governance approaches.

The analytical framework operationalises assessment through four evaluation benchmarks derived from international human rights instruments and algorithmic fairness literature. The non-discrimination benchmark assesses whether systems produce disparate impacts across protected characteristics, measured through statistical analysis when available or a qualitative assessment of design choices when quantitative data are

unavailable. The explainability benchmark assesses whether affected individuals can obtain meaningful information about the decision logic, as required by GDPR Articles 13(2)(f), 14(2)(g), and 15(1)(h) (2016). The oversight benchmark assesses human review mechanisms and their substantive adequacy, as specified in Article 14 of the EU AI Act (2024). The access to evidence benchmark examines whether parties can obtain information necessary to challenge decisions, measured against disclosure standards established in relevant litigation.

This methodology acknowledges inherent limitations. Doctrinal analysis cannot definitively resolve normative questions about algorithmic fairness where reasonable people disagree on fundamental values. Comparative synthesis risks overlooking context-specific factors that make direct transplantation of legal solutions across jurisdictions problematic. Case selection bias may favour systems where harms became visible through litigation or investigation, while missing equally problematic systems that remain opaque. These limitations are addressed through triangulation across multiple jurisdictions, explicit acknowledgement of normative commitments, and limiting prescriptions to reforms demonstrably feasible within existing legal frameworks.

3 The dual erosion framework

The dual erosion framework formalises the observation that AI systems threaten rights through two interdependent dimensions. The framework is represented as a two-axis analytical structure where the equality risk axis captures how systems generate discriminatory outcomes. In contrast, the remedy risk axis captures how systems obstruct effective redress. Understanding both dimensions and their interaction is necessary for comprehensive analysis and effective reform.

3.1 The equality risk dimension

The equality risk dimension encompasses three primary mechanisms through which algorithmic systems produce discriminatory outcomes. First, training data bias occurs when datasets used to develop machine learning models contain historical patterns of discrimination. As Barocas and Selbst [8, pp. 680–681] demonstrate, if training data reflects past discriminatory practices, algorithms learn to replicate those patterns. For instance, if an employment algorithm is trained on hiring decisions from a period when women were systematically excluded from certain roles, the algorithm will likely perpetuate that exclusion. This problem is technically inherent rather than incidental. Machine learning systems optimise for patterns in training data, making historical bias a feature rather than a bug.

Second, non-representative datasets systematically exclude or underrepresent particular groups, resulting in poorer algorithmic performance for those populations. The Canadian Supreme Court's decision in Ewert v Canada provides a doctrinal foundation for this concern. The Court held that risk assessment tools developed and validated using predominantly white North American samples could not be reliably applied to Indigenous offenders without specific cultural validation [25, paras. 72–73]. The Court concluded that using such tools without validation constitutes a breach of procedural fairness because it creates systematic risk of inaccurate assessments (para. 82). This holding establishes that representativeness is not merely a statistical nicety but a legal requirement flowing from due process and equality principles.

Third, discriminatory deployment contexts occur when algorithmic systems are applied more intensively or punitively to already marginalised communities. The Netherlands' childcare benefits scandal exemplifies this mechanism. Dutch authorities deployed a fraud detection algorithm that utilised dual nationality and foreign-sounding names as risk indicators, then concentrated enforcement efforts in immigrant communities. In particular, tens of thousands of families, disproportionately from ethnic minority backgrounds, were falsely accused of fraud and forced to repay benefits [2, pp. 11–14]. Many faced financial ruin. This case demonstrates that even technically neutral algorithms become discriminatory instruments when deployed within contexts of structural inequality.

These three mechanisms often operate simultaneously. India's Telangana state Samagra Vedika system illustrates compounded equality risks. The entity resolution algorithms processed data on 30 million residents, cancelling 1.86 million food security cards between 2014 and 2019 [4, 48]. Research on India's algorithmic welfare systems reveals that matching algorithms generate false positives, which are concentrated among populations with common names, spelling variations across government databases, and incomplete records [13, pp. 891–893]. These patterns disproportionately affect the poorest and least educated populations, who lack the resources to correct data errors. The Telangana system exemplifies these dynamics through its deployment in contexts where vulnerable individuals had no alternative means of accessing subsistence.

AI fairness scholarship provides technical taxonomies that formalize these mechanisms and document the challenges of mitigation. Data quality frameworks distinguish between label bias, where human annotators' implicit biases affect ground-truth labels; selection bias, where datasets systematically exclude certain populations; and measurement bias, where proxy variables inadequately capture the constructs of interest [59, pp. 4–6]. These categories align with the training data bias, non-representative datasets, and discriminatory deployment contexts described above. Empirical analysis shows that preprocessing techniques that address training data bias improve fairness in only 24% to 59% of scenarios, depending on the metric used, with significant performance degradation in many cases [14, pp. 17–19]. These limitations arise because bias in training data interacts with model architecture choices, optimization objectives, and deployment contexts in ways that data preprocessing cannot address. Fairness interventions also prove brittle, with fairness-preserving algorithms demonstrating sensitivity to fluctuations in dataset composition and preprocessing methods [29, pp. 329–330, 336–338].

3.2 The remedy risk dimension

The remedy risk dimension captures how algorithmic systems obstruct effective redress through three interconnected mechanisms. First, algorithmic opacity prevents affected individuals from understanding why decisions were made. This opacity operates at multiple levels. Technical opacity stems from machine learning models that even their developers cannot fully explain, arising both from model complexity, where deep neural networks with millions of parameters resist comprehension, and from limitations in current explainability techniques [54, pp. 206–209]. Methods like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive exPlanations) provide approximations of model behavior but do not guarantee fidelity to actual decision logic. SHAP provides both local and global explanations by treating features as players in a

game-theoretic framework, while LIME approximates complex models through local surrogate models for specific instances. However, both methods are highly sensitive to model choice and feature collinearity, raising caution about over-trusting their outputs [55, pp. 2–5, 7].

The mathematical impossibility of simultaneously optimising for specific fairness criteria [37, p. 17] means that trade-offs remain embedded in systems with no transparent rationale. Legal opacity arises when vendors claim trade secret protection over algorithms, blocking disclosure. In State v. Loomis, the Wisconsin Supreme Court permitted COMPAS risk assessments in sentencing, despite acknowledging that proprietary protections prevented defendants from examining the algorithm's operation (2017, paras. 65–66). The Court's cautionary warnings could not overcome the fundamental barrier that defendants lacked the information necessary to meaningfully challenge their assessments.

Second, evidential asymmetries place insurmountable burdens on individuals seeking to prove discrimination. Establishing algorithmic bias requires technical expertise, access to training data, and statistical analysis of system outputs across populations. Plaintiffs typically possess none of these resources while defendants control all relevant information. The burden-shifting framework in age discrimination law, as codified in U.S. Code of Federal Regulations [19], illustrates one potential solution by requiring employers to prove that factors causing adverse impact are reasonable and properly validated. However, courts have not yet systematically applied comparable frameworks to algorithmic discrimination, leaving plaintiffs with practical barriers to establishing prima facie cases.

Third, inadequate procedural safeguards mean that existing review mechanisms fail to provide effective oversight. Denmark's automated welfare fraud detection system exemplifies these deficiencies. The system analyses over 60 algorithms processing millions of data points about Danish residents to flag potential fraud [3, pp. 18–19]. Individuals receive no notice of algorithmic monitoring, no explanation of why they were flagged, and no meaningful opportunity to contest determinations before enforcement actions begin. This procedural vacuum violates basic due process requirements that predate the advent of algorithmic systems,however, existing legal frameworks have failed to enforce them effectively against automated decision-making.

3.3 Interactions between dimensions

The critical insight of the dual erosion framework is that equality risk and remedy risk interact dynamically to produce compound harms greater than their sum. Opacity enables bias to persist undetected because affected individuals cannot determine whether discriminatory patterns exist. Without transparency, it becomes impossible to conduct a statistical analysis of disparate impacts. Conversely, bias makes opacity more harmful because it means hidden processes produce consequential discrimination rather than merely unexplained neutral outcomes.

This interaction distinguishes algorithmic discrimination from traditional forms of discrimination. Human decision-makers may harbour biases, but those biases can be examined through questioning, depositions, and pattern analysis of explicit reasoning. Algorithms harbour biases embedded in training data and model architecture, yet vendors successfully claim those biases are trade secrets. The Dutch childcare benefits

scandal became remediable only after political pressure forced disclosure of the algorithm's use of nationality and names as risk factors [49]. Without that disclosure, obtained through extraordinary circumstances rather than routine legal process, the systematic discrimination would have remained invisible and unreviewable.

4 Framework application to high-stakes decision systems

4.1 Criminal justice risk assessment

Risk assessment algorithms in criminal justice exemplify high equality risk combined with moderate remedy risk, though the remedy risk increases when proprietary claims limit access. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, widely used in U.S. jurisdictions, demonstrates measurable disparate impacts. ProPublica's investigation documented that the algorithm misclassified African Americans as high-risk at twice the rate it misclassified white defendants as high-risk (Larson, Mattu, Kirchner, & Angwin, 2016). African Americans were also misclassified as low-risk less frequently than white defendants. These disparities occurred despite comparable actual recidivism rates, suggesting that the algorithmic bias reflected the training data's incorporation of historical discrimination rather than a lack of predictive accuracy.

The equality risk arises from multiple sources. Training data necessarily incorporates historical policing and sentencing patterns that disproportionately criminalised communities of colour through mechanisms including disparate enforcement, selective prosecution, and harsher sentencing for equivalent conduct. When algorithms optimise for predicting arrest or conviction, they optimise for replicating these historical patterns. Technical research confirms that calibration across groups with different base rates requires accepting either differential false-positive or false-negative rates [15, pp. 157–159]. This mathematical impossibility means that algorithmic fairness necessarily involves value choices that current legal frameworks do not require vendors to articulate or justify.

Empirical research confirms these theoretical limitations. Comparative evaluation shows that COMPAS achieves no greater accuracy than predictions made by individuals with no criminal justice expertise, with both achieving approximately 65% accuracy. Moreover, although COMPAS processes up to 137 features, comparable accuracy can be achieved using simple linear classifiers with only two features, suggesting that algorithmic complexity provides minimal predictive benefit [22, pp. 3–5]. Comprehensive fairness testing across multiple jurisdictions reveals that achieving one fairness criterion necessarily compromises others due to mathematical incompatibility when base rates differ by race [61, pp. 5–7]. These findings demonstrate that recidivism prediction faces fundamental technical constraints that legal frameworks alone cannot resolve.

Remedying risk in criminal justice algorithms primarily stems from trade secrecy claims that block examination. In State v. Loomis, the defendant challenged his sentence, arguing that the proprietary COMPAS algorithm violated due process because he could not examine its operation or verify its accuracy. The Wisconsin Supreme Court rejected this claim, holding that COMPAS could be used in sentencing provided courts considered other factors and did not rely on COMPAS scores as determinative (2017, paras. 65–66). This resolution left defendants unable to challenge algorithmic assessments substantively while theoretically limiting their weight in sentencing. The practical effect

is that algorithmic assessments influence judicial decision-making through anchoring effects and implicit deference to technical analysis, while defendants lack effective means to contest their validity.

Recent scholarship proposes protective order frameworks adapting civil discovery mechanisms to algorithmic contexts [62, pp. 1410–1413]. Under this approach, courts would order the disclosure of algorithmic details to qualified experts, subject to protective orders that limit further dissemination. The Federal Circuit's [26] decision in Royal Brush Manufacturing v United States provides legal foundation, holding that constitutional due process requirements override statutory trade secret protections when necessary for meaningful adversarial testing (2023, slip op. at 12–14). This precedent establishes that trade secrecy cannot categorically bar disclosure when disclosure is necessary to ensure procedural fairness.

4.2 Welfare eligibility and fraud detection systems

AI governance scholarship on welfare automation documents systematic patterns where algorithmic decision-making systematically disadvantages vulnerable populations. Automated eligibility systems transfer costs and burdens to beneficiaries who must navigate opaque appeals processes, correct data errors they did not create, and prove eligibility against presumptions of fraud [31, pp. 214–216]. Research across multiple jurisdictions shows that these systems lack meaningful human oversight, with review processes deferring to automated determinations rather than substantively evaluating individual circumstances [23]. The combination of vulnerable affected populations, pervasive opacity, and weak procedural protections creates compounded discrimination where both equality and remedy risks operate at maximum intensity. Three case studies demonstrate these patterns across different regulatory contexts.

Welfare algorithms pose an extreme risk of equality violations, compounded by a near-absolute risk of remedy failures stemming from vulnerable affected populations, pervasive opacity, and weak procedural protections. The Dutch System Risk Indication (SyRI) system integrated data from tax authorities, employment agencies, education institutions, and the police to generate risk scores, flagging individuals for potential fraud investigation. In *NCJM* et al*. and FNV v The State of the Netherlands*, the Hague District Court struck down SyRI in 2020, holding that the system violated Article 8(2) of the European Convention on Human Rights because it lacked sufficient transparency and safeguards (2020, paras. 6.76–6.79). The Court found that affected individuals received no information about how risk scores were calculated, what data contributed to assessments, or why they were targeted, making meaningful challenge impossible.

Subsequent investigation revealed that SyRI incorporated discriminatory risk factors, including nationality, migration background, and neighbourhood characteristics, as predictors of fraud [2, pp. 24–25]. The system targeted low-income neighbourhoods with higher immigrant populations, creating compounded discrimination. Individuals from targeted communities faced both disproportionate suspicion and practical inability to contest determinations. This pattern illustrates how welfare algorithms combine data-driven biases with procedural deficits, resulting in systematic exclusion from essential services.

India's Telangana Samagra Vedika system, examined earlier for its equality risks through entity resolution algorithms, demonstrates how remedy deficits compound

algorithmic discrimination beyond legal redress. The remedy risk operates through three mechanisms that interact with the false positive patterns already established. First, algorithmic opacity prevents affected individuals from understanding the benefits of cancellations. The system does not provide generic notices of "duplicate records" without identifying which data triggered the flags or how to correct the errors. Second, the system operates in a regulatory vacuum, lacking independent oversight, validation studies, or impact assessments [4]. Third, procedural barriers render appeal rights practically meaningless. Individuals discover cancellations only when denied rations at distribution points, with administrative review presuming algorithmic accuracy and placing the burden on individuals to prove system errors.

The interaction between equality risk and remedy risk produces compound harms. False-positive rates concentrated among vulnerable populations become unreviewable because opacity prevents the detection of systematic patterns. Individual cases receive no meaningful review because procedural barriers prevent affected persons from accessing remedies. Field documentation reveals instances in which families have exhausted their savings while attempting to restore benefits, creating perverse outcomes in which fraud-prevention measures inadvertently impoverish legitimate beneficiaries [48]. This dynamic exemplifies how opacity enables bias to persist undetected while bias makes opacity particularly harmful for populations surviving at subsistence levels.

Denmark's welfare fraud detection system demonstrates that remedy risks persist even within robust data protection regimes, revealing the limitations of general legal frameworks in the absence of specific algorithmic accountability mechanisms. Operating under GDPR Article 22 (2016) protections, the Udbetaling Danmark (UDK) system nonetheless provides minimal substantive transparency. Amnesty International's investigation documented that UDK's algorithms process data on millions of Danish residents to flag potential fraud. Nevertheless, individuals receive no notice of algorithmic monitoring, no explanation of why specific risk scores were assigned, and no practical opportunity to contest determinations before enforcement actions commence (2024a, pp. 39–40).

This pattern contrasts with India's complete regulatory vacuum yet produces comparable remedy deficits through different mechanisms. Where Telangana operates without legal constraints, Denmark operates within extensive legal protections that prove insufficient in practice. The GDPR (2016) provides Article 15 rights to explanation and Article 22 protections against automated decision-making; however, these provisions often fail to provide meaningful recourse when automated risk scoring triggers invasive investigations. The gap between formal legal rights and the availability of practical remedies demonstrates that general data protection principles, while necessary, remain insufficient without implementation mechanisms that specify what constitutes an adequate explanation for algorithmic risk assessments, what procedural safeguards must precede enforcement actions, and what oversight bodies can verify compliance.

4.3 Credit scoring and financial services

Credit scoring algorithms present a moderate risk of equality issues, with evolving remedy risks, as recent European jurisprudence establishes stronger procedural protections. In OQ v. Land Hessen (SCHUFA decision), the Court of Justice of the European Union [17] made a landmark ruling against Germany's SCHUFA credit scoring system,

establishing that automated credit score generation constitutes automated decision-making under GDPR Article 22(1) (2016) when third parties assign a determining role to scores (2023, paras. 71–73). This holding expanded Article 22 protections beyond previously narrow interpretations, requiring individuals to receive meaningful information about the logic behind their credit scores and to have the right to contest determinations.

The equality risk in credit algorithms stems from training data incorporating historical lending discrimination. Research documents persistent disparities in credit access based on race, gender, and socioeconomic status that algorithms trained on historical data necessarily replicate [30, pp. 42–43]. Alternative data sources, including mobile phone usage patterns, app installations, and social connections, can potentially expand credit access to underserved populations but simultaneously create new risks of discrimination through proxies for protected characteristics. Indian fintech lending exemplifies these tensions, with alternative credit scoring facilitating financial inclusion for individuals with credit invisibility, while raising concerns about privacy violations and discriminatory patterns in automated underwriting [41, pp. 8–9].

The SCHUFA decision addresses remedy risk by requiring a clear and meaningful explanation of the scoring logic. The Court held that data subjects must receive information enabling them to understand decision-making rationale and challenge outcomes effectively (2023, paras. 56–57, 66). This standard exceeds generic explanations of scoring methodologies, requiring personalised information about factors affecting individual assessments. Implementation challenges remain substantial. Credit scoring involves complex machine learning models where the importance of factors varies by individual. Providing genuinely meaningful explanations requires technical sophistication that regulatory guidance has not yet specified adequately. Nonetheless, the SCHUFA decision establishes judicial commitment to substantive rather than formal transparency rights.

4.4 Employment screening and hiring algorithms

Algorithmic employment screening poses a significant risk of equality issues, with emerging but incomplete frameworks for remedy. The pending litigation in Mobley v Workday, Inc., exemplifies both the risks and the evolving legal responses. Plaintiffs allege that Workday's AI-driven recruitment platform systematically excludes candidates over 40 from protected groups through biased training data and design choices (2023, pp. 1–2). The District Court granted preliminary certification for nationwide collective action under the Age Discrimination in Employment Act in May 2025, finding sufficient commonality to proceed as a class. While not the first AI discrimination lawsuit, Mobley represents the first private class action against an AI vendor rather than an employer, and it has survived a motion to dismiss and received conditional certification for potentially hundreds of millions of class members.

The equality risk in hiring algorithms stems from training data incorporating historical employment discrimination. AI fairness research demonstrates that employment screening systems perpetuate organizational homogeneity by learning patterns in which success correlates with the demographic majority. Analysis of commercial hiring platforms shows that algorithms trained on data from successful employees replicate the workforce composition rather than identifying talent [50, pp. 472–474]. If an algorithm learns from past hiring decisions in industries where women or minorities were systematically excluded, it will replicate those patterns.

Additionally, algorithms often use proxy variables correlated with protected characteristics. Resume screening algorithms systematically disadvantage candidates with non-traditional career paths, employment gaps, or educational credentials that do not align with historical patterns, creating barriers that are particularly acute for women returning from parental leave, candidates from under-resourced educational institutions, and career changers [34, pp. 800–803]. For instance, requiring specific zip codes, university attendance, or extracurricular activities may serve as proxies for race and socioeconomic status. Research demonstrates that seemingly neutral factors, such as names, addresses, and employment gaps, correlate with protected characteristics, allowing for discrimination without explicit consideration of race or gender. Empirical evaluation of language-model-based résumé screening shows that AI hiring systems favor names associated with White males, while resumes with Black male names are rarely [often never] ranked first across tested configurations, demonstrating how advanced language models perpetuate historical hiring inequalities when trained on unbalanced data [63, pp. 1584–1586], [1, pp. 155–157].

Remedy risk in employment contexts stems from evidential asymmetries. Establishing algorithmic discrimination requires demonstrating that the system has produced disparate impacts, that alternative approaches exist with less discriminatory effects, and that the employer's justifications lack merit. Plaintiffs typically cannot access training data, model architecture, or validation studies necessary to make these showings. Defendants control all relevant information and can assert trade secrecy over algorithmic details. The Age Discrimination in Employment Act (ADEA) burden-shifting framework, codified in U.S. Code of Federal Regulations [19], partially addresses these asymmetries by requiring employers to prove that factors causing disparate impact are reasonable when viewed from a prudent employer's perspective. This framework should extend to algorithmic systems, requiring vendors and employers to demonstrate that algorithms were validated correctly, that alternatives were considered, and that disparate impacts were minimised, consistent with legitimate business needs.

4.5 Automated content moderation

Content moderation algorithms pose distinct challenges because they operate at scale, involve complex policy decisions that extend beyond technical classification, and directly implicate free speech rights. Meta's content moderation infrastructure demonstrates the technical imperatives driving large-scale automation. The Few-Shot Learner system, deployed in December 2021, processes content across more than 100 languages and offers multimodal capabilities, analysing both text and images [44]. The system employs few-shot learning approaches, enabling adaptation to new harmful content categories within weeks rather than the months required for traditional supervised learning model retraining. This architectural choice addresses a temporal lag problem in which novel policy violations evolve faster than conventional training cycles allow.

Combined with XLM-R cross-lingual classifiers, enabling multilingual content processing at scale [18, pp. 8440–8442], these systems enable proactive content moderation, making human review of all content technically infeasible due to volume constraints. Recent advances apply reinforcement learning to content moderation decisions, achieving 10–100 times higher data efficiency than supervised fine-tuning while introducing new challenges in specifying reward functions that encode normative policy judgments

[39]. This technical constraint of scale drives automation while simultaneously creating systematic risks of both over-removal through algorithmic conservatism and under-removal through adversarial evasion techniques.

The equality risk in content moderation stems from two sources. First, training data bias leads algorithms to misclassify content from marginalised communities more often. Research demonstrates that hate speech detection algorithms flag African American Vernacular English (AAVE) at significantly higher rates than standard American English for equivalent semantic content [57, pp. 1668–1670]. This reflects training data where annotators labelled AAVE as more offensive due to implicit bias, creating systematic over-removal of content from Black users. Second, policy enforcement decisions involve value choices about what constitutes acceptable speech that platforms make without adequate input from the affected communities. Platforms establish standards for nudity, political speech, religious expression, and other categories through internal policy processes that often lack democratic legitimacy and transparency.

Remedy for risk in content moderation often manifests through inadequate appeals processes and opacity regarding decision logic. The Santa Clara Principles on Transparency and Accountability, updated in 2021, establish minimum standards, including that users must receive notice identifying specific policy violations, must have access to human review, and must receive timely decisions on appeals. The revised Foundational Principles, in particular, Principle 5, specifically address automated content moderation, requiring platforms to explain when automation is used and provide meaningful human oversight. However, research documents widespread failures to meet these standards. Are and Briggs [6] found that appeals on Instagram and TikTok often fail to provide substantive review, with high rates of arbitrary denials and minimal explanation of reasoning (pp. 2009–2011).

The EU Digital Services Act (DSA) (2022) addresses some remedy deficits through Article 20 requirements for internal complaint-handling systems and Article 21 provisions for out-of-platform dispute resolution. The online platforms, therefore, must provide clear grounds for content decisions, enable appeals, and respond within reasonable timeframes. These provisions operationalise due process requirements for automated content moderation. Implementation challenges remain substantial, including determining what constitutes an adequate explanation for algorithmic decisions and ensuring that human review provides genuine oversight rather than rubber-stamping automated determinations.

4.6 Comparative case analysis

Table 1 synthesises the case studies examined above, enabling systematic comparison across equality risk mechanisms, remedy risk barriers, legal responses, and outcomes. This comparative structure reveals patterns across jurisdictions and domains that might otherwise remain obscured in narrative analysis.

The table reveals three systematic patterns. First, equality risks manifest consistently across domains through training data bias, non-representative datasets, and discriminatory deployment contexts. Second, remedy risks cluster around proprietary claims that block disclosure, evidential asymmetries that favour system operators, and inadequate procedural safeguards. Third, legal responses vary by jurisdiction and regulatory maturity, with European frameworks offering stronger procedural protections than

Table 1 Comparative analysis of algorithmic decision systems

Case/System	Jurisdiction	Domain	Equality risk	Remedy risk	Legal posture	Outcome
State v. Loomis [58]	Wisconsin, US	Criminal Justice (COMPAS)	Training data reflects historical sentencing disparities	Proprietary algorithm; disclosure blocked	Permitted with cautionary warnings; other factors required	Trade secrecy upheld; no effective challenge mechanism
Ewert v Canada [25]	Canada	Criminal Justice (correctional service risk tools)	Tools validated only on white populations; bias against Indigenous offenders	Operated without cultural validation	Supreme Court required cultural validation; procedural fairness breach found	Mandatory validation for diverse populations established
ProPublica [38]	US (national)	Criminal Justice (COMPAS)	African Americans misclassified as high-risk at 2 × rate	Proprietary methodology; no access to weighting	Journalistic investigation; no judicial remedy	Increased awareness; limited jurisdictional responses
Mobley v. Workday [45]	US (N.D. Cal.)	Employment (AI recruitment)	Training data incorporates historical hiring discrimination	No access to training data or model architecture	Conditional class certification granted under ADEA	Litigation pending; potential burden-shifting framework
SyRI/NJCM v. Netherlands (2020)	Netherlands	Welfare (fraud detection)	Nationality, migration background as risk factors; targeted immigrants	No transparency on methodology or scoring	Court struck down under ECHR Art. 8(2)	System discontinued; transparency requirement established
Telangana Samagra Vedika	India	Welfare (entity resolution)	False positives concentrated among vulnerable populations	Generic notices; no oversight; algorithmic accuracy presumed	No judicial challenge; regulatory vacuum	1.86 M cards cancelled; ongoing harm
UDK Denmark	Denmark	Welfare (fraud detection)	Over 60 algorithms targeting vulnerable groups	No monitoring notice; no explanation; limited contestation	GDPR applies but minimal enforcement	Invasive investigations; formal rights insufficient
SCHUFA/OQ v. Land Hessen (2023)	Germany/EU	Credit Scoring	Training data reflects historical lending discrimination	Credit bureaus-controlled methodology	CJEU expanded Article 22; meaningful explanation required	Stronger protections; implementation challenges remain
Platform Moderation (ongoing)	Global	Content Moderation (Meta Systems etc.)	AAVE flagged at higher rates; annotator bias in training data	Inadequate appeals; opacity in logic; limited human review	EU DSA (2022) + Santa Clara Principles [56]	Procedural improvements; substantive quality uncertain

U.S. approaches, which often defer to trade secrecy claims. These patterns demonstrate that the dual erosion framework captures dynamics transcending specific technological implementations or legal traditions.

5 Emerging regulatory frameworks

Recent regulatory developments respond to algorithmic harms through increasingly sophisticated approaches that recognise the dual erosion framework's core insights, even if they do not articulate them explicitly. Three regulatory instruments merit detailed analysis for their approaches to equality risk, remedy risk, and the interaction between them.

5.1 European Union artificial intelligence act

The EU AI Act (2024), which entered into force on 1 August 2024, establishes a comprehensive risk-based framework that operationalises the dual erosion logic. The Act categorises AI systems into four tiers based on risk levels. Prohibited systems present unacceptable risks and include social scoring by governments and real-time biometric identification in public spaces, absent narrow exceptions (Article 5). High-risk systems, defined as safety components or systems used in domains listed in Annex III, including employment, education, law enforcement, and access to essential services, trigger extensive requirements (Article 6). While limited risk systems face transparency obligations (Article 50), minimal risk systems remain largely unregulated.

For high-risk systems, the Act addresses equality risk through Article 10 data governance requirements. Training datasets must be relevant, representative, and error-free to the extent possible. Where datasets contain bias that cannot be eliminated, providers must detect, document, and mitigate such bias. Providers must establish data governance practices that ensure appropriate design choices, data collection, and processing operations (Article 10(2)). These requirements mandate affirmative attention to data quality rather than allowing providers to use convenience samples that systematically underrepresent particular populations.

The Act addresses remedy risk through overlapping mechanisms. Article 13 transparency requirements mandate that providers prepare instructions for deployers, including system capabilities, limitations, reasonably foreseeable misuse, and the degree of accuracy expected from the system. Article 14 human oversight requirements demand that deployers assign oversight to natural persons with competence, training, and authority to intervene or interrupt systems displaying anomalous outputs. Article 27 requires deployers in public sector contexts or certain high-risk domains to conduct fundamental rights impact assessments before deploying systems. Article 72 post-market monitoring obligations require providers to actively collect and document information about system performance in real-world conditions.

These provisions collectively reduce remedy risk by ensuring multiple checkpoints where affected individuals and regulatory authorities can examine system operation. However, significant limitations remain. The Act does not specify what constitutes an adequate explanation of high-risk system logic beyond generic instructions for deployers. It does not establish private rights of action enabling individuals to enforce requirements directly. Enforcement depends on designated authorities with uncertain capacity

and political will. Implementation will determine whether the Act's ambitious framework translates into adequate protection or remains aspirational.

5.2 European Union digital services act

The Digital Services Act (2022), which came into full effect on 17 February 2024, establishes platform accountability requirements, including those addressing content moderation algorithms. While not explicitly framed within the dual erosion framework, the DSA's provisions align with both dimensions. Article 27 transparency obligations for recommender systems require Very Large Online Platforms to provide users with at least one option for recommender systems not based on profiling. This addresses equality risk by reducing the algorithmic amplification of divisive or harmful content that is targeted based on inferred characteristics.

Articles 34 and 35 require Very Large Online Platforms to conduct systemic risk assessments and implement mitigation measures. Risk assessment must include consideration of how systems affect fundamental rights, including freedom of expression, non-discrimination, and privacy (Article 34(1)(b)). This requirement acknowledges that content moderation involves value choices affecting protected interests that cannot be resolved through purely technical optimisation. Article 37 independent auditing requirements provide external verification of compliance, addressing information asymmetries that otherwise allow platforms to control narratives about algorithmic performance.

The DSA addresses remedy risk through Articles 17 and 20, which establish internal complaint systems, and Article 21, which provides for out-of-platform dispute resolution. Users must receive clear grounds for content decisions, the possibility to contest decisions, and timely resolution of complaints (Article 17(3)). Very large online platforms must provide access to certified out-of-platform dispute resolution bodies, enabling users to resolve disputes outside platform-controlled processes (Article 21). Article 40's data access provisions require Very Large Online Platforms to provide vetted researchers with access to platform data, addressing epistemic asymmetries that prevent external analysis of algorithmic effects.

The DSA's limitations include vague standards for what constitutes adequate explanations, limited enforcement mechanisms for individual users, and dependence on regulatory capacity. Nonetheless, the framework represents a significant advance over previous regimes by acknowledging that procedural protections require specific mechanisms rather than general principles.

5.3 Revised Santa Clara principles

The revised Santa Clara Principles on Transparency and Accountability in Content Moderation (2021) provide normative standards for content moderation governance that civil society organisations, academics, and some platforms have adopted. The updated principles expand from three operational principles (numbers, notice, appeal) to a comprehensive framework including foundational principles and operational requirements.

Foundational Principle 5 on integrity and explainability directly addresses algorithmic content moderation, requiring that platforms explain when and how automated systems are used, that automated systems are validated for accuracy and bias, and that meaningful human oversight exists. This principle acknowledges that automation creates distinctive risks requiring specific safeguards beyond general due process requirements.

Operational Principle 2 on notice requires that platforms explain whether content was detected or removed through automated means and provide meaningful information about how automation reached particular decisions. Operational Principle 3 on appeal requires that platforms provide access to human review and that appealed decisions receive substantive reconsideration rather than automated confirmation of initial determinations.

These principles remain voluntary rather than legally binding, limiting their practical force. However, they establish consensus standards that inform regulatory development and provide benchmarks for evaluating platform practices. The European Commission explicitly referenced the Santa Clara Principles when developing DSA (2022) complaint-handling requirements, demonstrating how civil society standards can inform the development of hard law [12, pp. 11–13].

6 Implementable reform framework

The dual erosion framework generates specific reform proposals addressing both equality risk and remedy risk through implementable legal instruments. Rather than offering aspirational principles, these proposals provide concrete standards that courts and regulators can apply immediately within existing legal frameworks.

6.1 Discovery protocols reconciling trade secrets with due process

Algorithmic systems pose discovery challenges because plaintiffs require access to training data, model architecture, and validation studies to establish claims of discrimination. However, defendants often assert trade secret protections that block disclosure. Royal Brush Manufacturing v United States establishes a constitutional foundation for resolving these tensions by holding that due process requirements override statutory trade secret protections when disclosure is necessary for meaningful adversarial testing (2023, slip op. at 12–14). Building on this precedent, courts should adopt a tiered disclosure protocol modelled after those used in complex commercial litigation.

Stage one discovery should require defendants to produce general information about algorithmic systems, including descriptions of intended purpose, categories of data used, general methodology, and validation procedures. This information aligns with the requirements for meaningful information about decision logic under GDPR Articles 13(2)(f) and 15(1)(h) (2016). Defendants cannot assert trade secrecy over general descriptions of system operation because meaningful adversarial testing requires an understanding of the basic system architecture.

Stage two discovery, triggered after plaintiffs establish a preliminary showing of discriminatory effects, should require the production of specific technical details to qualified experts under protective orders. Protective orders should permit disclosure to experts retained by plaintiffs but prohibit experts from sharing information beyond what is necessary to evaluate discrimination claims. This approach strikes a balance between defendants' legitimate interests in protecting commercially sensitive information and plaintiffs' rights to test claims effectively. The U.S. Federal Rules of Civil Procedure [27] already authorise protective orders limiting disclosure, and courts regularly employ them in patent litigation and trade secret cases. Algorithmic discrimination cases warrant no less protection for plaintiffs' procedural rights.

Stage three should permit limited in-camera review where parties dispute whether information is genuinely confidential or whether protective orders provide adequate protection. Courts retain inherent authority to examine materials in camera to determine discoverability. In *United States v. Zolin*, the Supreme Court held that U.S. Federal Rule of Evidence 104(a) [28] does not prohibit the use of in camera review when determining the existence of a privilege [60, pp. 565–570]. This authority extends to evaluating whether claimed trade secrets are sufficiently important to justify limiting plaintiffs' access to information needed for effective adjudication.

6.1.1 Model statutory provision: algorithmic disclosure requirements

In any proceeding where an algorithmic determination produces legal effects or similarly significant consequences for an individual, the deployer shall disclose to the affected individual and adjudicating authority, under protective order if necessary:

(a) All categories of personal data processed, including derived or inferred attributes; (b) A plain-language description of the decision-making logic, including factors considered and their relative importance; (c) The specific basis for the particular determination, including which data elements contributed to the outcome; (d) Validation studies demonstrating the system's accuracy and any identified limitations or biases.

Trade secret protections shall not preclude disclosure necessary to enable meaningful contestation of algorithmic determinations affecting fundamental rights. Where deployers assert trade secret protection, courts shall employ graduated disclosure under protective orders permitting access to qualified experts while safeguarding legitimate commercial confidentiality through procedures established in [jurisdiction's civil procedure rules].

This provision operationalises the graduated disclosure framework, providing adaptable language for incorporation into civil procedure rules. The framework strikes a balance between trade secret protection and due process, establishing that fundamental rights considerations take precedence over commercial confidentiality when necessary for meaningful contestation.

6.2 Reversed burden framework adapted from employment law

The ADEA regulatory framework, which addresses age discrimination through disparate impact analysis, is codified in the U.S. Code of Federal Regulations [19], providing a template for algorithmic discrimination claims. Under this framework, plaintiffs establish prima facie cases by identifying specific employment practices causing adverse impact and demonstrating statistical disparities. The burden then shifts entirely to employers to prove reasonable factors other than age as an affirmative defence. Critically, in *Meacham v. Knolls Atomic Power Laboratory* [42, pp. 91–92], the Court held that employers bear both the burden of production and the burden of persuasion regarding reasonableness.

Courts should adapt this framework to address algorithmic discrimination by requiring defendants to demonstrate reasonable factors other than protected characteristics once plaintiffs have shown a disparate impact. The reasonableness standard, codified in the U.S. Code of Federal Regulations [19], requires evaluating whether practices are objectively reasonable when viewed from the perspective of a prudent employer mindful of its responsibilities under anti-discrimination law. U.S. Code of Federal Regulations [19], specifies relevant considerations including the extent to which factors relate to

stated business purposes, the extent to which factors were defined and applied fairly, the extent to which employers limited subjective discretion, the extent to which employers assessed adverse impacts on protected groups, and the degree of harm and steps taken to reduce it.

Applying these factors to algorithmic systems would require defendants to demonstrate that training data were representative, that alternative approaches with less discriminatory effects were considered, that the system was validated using appropriate populations, that adverse impacts were measured and addressed, and that ongoing monitoring detects emerging bias. This framework is immediately implementable through the existing disparate impact doctrine without requiring new legislation. Courts possess the authority to adapt burden-shifting frameworks to novel contexts, and the ADEA model provides tested standards for evaluating the reasonableness of practices producing disparate impacts.

6.2.1 Worked example: ADEA-style algorithmic hiring discrimination case

A worked example illustrates how this framework would operate in practice. Consider a hypothetical case involving an AI-powered resume screening system deployed by a large employer.

Pleading and Prima Facie Case: Plaintiff, a 52-year-old applicant with fifteen years of relevant experience, alleges that the employer's algorithmic screening system rejected her application despite strong qualifications. To establish a prima facie case, plaintiff demonstrates: (1) membership in a protected class (age 40 +); (2) qualification for the position based on education and experience; (3) adverse employment action (rejection); and (4) statistical evidence showing the system rejects applicants over 40 at 1.8 times the rate of younger applicants, exceeding the four-fifths (80%) threshold for disparate impact. The court finds that a prima facie case has been established, shifting the burden entirely to the employer.

Discovery and Burden-Shifting: The employer must now articulate and prove reasonable factors that are not based solely on age. Initial discovery reveals the screening algorithm considers: years of experience, educational institution attended, employment history continuity, and skills assessments. The employer argues these factors serve legitimate business purposes. However, plaintiff's counsel, pursuant to a protective order, obtains detailed technical documentation from a qualified expert. Expert analysis reveals problematic design choices: (a) the system treats experience beyond fifteen years as a negative factor, penalizing rather than valuing longer careers; (b) institution prestige rankings disproportionately favour recently established programs, disadvantaging candidates who graduated decades ago; (c) employment gaps trigger automatic penalties without contextual assessment, disproportionately affecting workers over 40 who experienced 2008 recession layoffs. The employer cannot demonstrate that these specific algorithmic choices were necessary for identifying qualified candidates or were validated on age-diverse populations. The court finds the employer failed to satisfy its burden under [19].

Less Discriminatory Alternatives: Even if the employer had established reasonableness, the plaintiff demonstrates feasible alternatives with reduced discriminatory impact: capping experience consideration at fifteen years rather than penalising longer careers, providing opportunities to contextualise employment gaps, and removing institutional

prestige weights. The plaintiff's expert demonstrates through simulation that these modifications would reduce the age disparity from 1.8:1 to 1.1:1, while maintaining or improving the system's predictive validity for job performance. The employer's failure to adopt or even evaluate these apparent alternatives further undermines any reasonableness defence.

Remedies and Monitoring: The court orders comprehensive relief: (1) compensatory damages for plaintiff's lost wages and emotional distress; (2) injunctive relief requiring the employer to modify the algorithmic system to eliminate identified age-correlated disparities within ninety days; (3) prospective monitoring through annual independent bias audits for three years; (4) conditional class certification enabling other affected applicants to seek relief. The employer must submit quarterly compliance reports to a court-appointed monitor, demonstrating that the modified algorithms no longer produce a disparate impact.

Impact Analysis: This outcome demonstrates balanced consideration of competing interests. The societal impact establishes that algorithmic employment decisions receive scrutiny equivalent to that of human decisions, thereby protecting workers from automated age discrimination. The commercial impact creates clear compliance incentives, such as proactive bias testing, transparent documentation, and consideration of less discriminatory alternatives, without imposing impossible standards. Employers who use certified independent audits and implement recommendations before litigation are more likely to receive qualified immunity, thereby encouraging voluntary compliance over defensive litigation. The procedural innovation demonstrates that burden-shifting frameworks function effectively in algorithmic contexts, with protective orders enabling meaningful discovery of technical details without compromising legitimate trade secrets. Courts can apply this framework using existing disparate impact doctrine and discovery rules, requiring no new legislation.

6.3 Risk-tiered due diligence standards

The EU AI Act's (2024) risk-based framework provides a model for risk-tiered due diligence obligations applicable beyond the European Union. Systems that produce legal effects or similarly significant effects should trigger heightened due diligence, including robust data governance, validation requirements, transparency obligations, and ongoing monitoring. The Act defines significant effects as circumstances where algorithmic decisions affect access to essential services, education, employment, law enforcement, migration, border management, or the administration of justice (2024, Annex III). These categories align with domains where algorithmic errors result in substantial harm to fundamental rights and freedoms.

For high-risk systems, providers should be required to establish and document risk management systems that identify foreseeable risks, implement mitigation measures, and test systems throughout their lifecycle (Article 9). Data governance obligations should mandate that training datasets are relevant, representative, and appropriately examined for bias (Article 10(2)). Technical documentation must describe the system design, data processing, human oversight mechanisms, and expected accuracy levels (Article 11). Post-market monitoring should require the systematic collection and analysis of data about system performance in operational contexts (Article 72).

Deployers of high-risk systems should conduct fundamental rights impact assessments before deployment, when systems will be used by public authorities or in domains where errors systematically disadvantage particular populations. These assessments should identify affected groups, analyse potential discriminatory effects, evaluate the adequacy of safeguards, and document consultation with potentially affected communities. The requirement for community consultation distinguishes this approach from purely technocratic risk assessment by acknowledging that affected populations possess expertise about how systems impact their lives that developers and regulators lack.

Implementation could occur through regulatory guidance that specifies due diligence expectations or through common law development, as courts recognise heightened duties of care when deploying high-stakes algorithmic systems. The advantage of regulatory implementation is comprehensive coverage and clear, standardised guidelines. The advantage of common law development is its flexibility in adapting to evolving technologies and contexts. Both approaches could operate simultaneously, with regulations establishing minimum floors and tort liability incentivising practices exceeding regulatory minima.

6.3.1 Independent audit requirements and safe harbour provisions

Independent algorithmic audits provide external verification of compliance with non-discrimination and transparency obligations while creating safe harbour incentives for proactive risk management.

To ensure audit quality and independence, auditors should meet minimum qualifications including: (1) demonstrated technical expertise in machine learning systems, statistical bias detection, and fairness metrics; (2) independence from the deployer, with no financial relationships within the preceding three years; and (3) professional liability insurance covering audit activities. The audit scope for high-risk systems should encompass the examination of training data for representativeness and bias, a review of model architecture and feature selection; validation of testing procedures across demographic groups, an assessment of deployment context and monitoring mechanisms, and measurement of disparate impact using established fairness metrics.

Auditors should produce two reports: a public summary that documents the methodology, key findings, and recommendations for affected communities and regulators; and a detailed technical report for deployers, identifying specific vulnerabilities and remediation measures. To incentivize voluntary auditing, deployers who commission independent audits and implement recommended remediation measures within ninety (90) days should receive qualified immunity from civil liability for issues identified in audit reports, provided that: (1) qualified independent auditors conducted the audit; (2) remediation substantially addresses identified risks; and (3) the deployer maintains ongoing monitoring to detect emerging issues.

This safe harbour encourages proactive compliance without creating liability shields for deployers who ignore audit findings or fail to implement remediation. High-risk systems should undergo annual independent audits, while medium-risk systems require biennial auditing. Audit reports should be submitted to relevant regulatory authorities, with public summaries published to enable accountability. This framework adapts established financial and environmental auditing practices to algorithmic contexts, providing tested mechanisms for independent oversight and assurance.

6.4 Operationalising meaningful information requirements

The GDPR (2016) requires that data subjects receive meaningful information about the logic involved in automated decision-making (arts. 13(2)(f), 14(2)(g), 15(1)(h)). The SCHUFA decision interprets this requirement broadly, holding that meaningful information must enable data subjects to understand decision-making rationale and contest outcomes effectively (2023, paras. 56–59). However, regulatory guidance has not specified concrete standards for determining whether explanations are meaningful, creating implementation challenges.

Courts and regulators should adopt tiered explanation standards based on the significance of the decision and the rights affected. For decisions that produce legal effects, explanations must include the categories of data used, the relative importance of different factors in the particular decision, the reasons why specific data elements led to adverse outcomes, and information about how similar cases are typically treated. This standard exceeds generic methodology descriptions by requiring a personalised explanation of individual decisions.

For complex machine learning systems where a complete explanation is technically infeasible, explanations must provide approximations enabling meaningful contestation. Counterfactual explanations, which describe how changes to the input would alter the outcomes, provide an alternative approach. For instance, credit denials could explain that approval would have required different debt-to-income ratios, a different employment history, or different payment patterns. Feature importance analysis, which identifies the factors contributing most to particular decisions, provides another approach. The key principle is that explanations must enable affected individuals to determine whether decisions reflect the appropriate application of legitimate factors or an inappropriate weight on problematic proxies or protected characteristics.

Regulatory authorities should develop adequacy guidelines for explanations through notice-and-comment procedures, incorporating input from affected communities, technical experts, and deployers. These guidelines should specify the minimum information requirements for different decision contexts, provide examples of adequate and inadequate explanations, and establish safe harbours for explanation approaches that meet regulatory standards. This approach strikes a balance between flexibility for innovation and certainty regarding compliance obligations.

6.5 Participatory governance mechanisms

Current algorithmic governance suffers from democratic deficits because technical experts and corporate actors make value choices that affect fundamental rights without meaningful input from the affected communities. The EU AI Act's (2024) risk assessment requirements improve this situation by mandating consideration of the impacts on fundamental rights, but they do not require direct participation by affected populations. Governance frameworks should incorporate participation mechanisms at multiple stages, including system design, deployment decisions, ongoing monitoring, and dispute resolution.

For public sector algorithmic systems, participation should occur through advisory bodies that include members from affected communities, who have the authority to review proposed systems, access technical documentation, and recommend modifications or rejection before deployment. These bodies should have sufficient technical

support to evaluate systems meaningfully, including access to independent experts and the ability to commission external audits. Their recommendations should be public, and agencies should be required to respond substantively to objections, creating accountability for deployment decisions.

For private sector systems in high-stakes domains, participation could occur through mandatory consultation requirements in fundamental rights impact assessments. The EU AI Act (2024) requires such assessments for certain deployers (Article 27) but does not specify consultation procedures. Regulatory guidance should require deployers to conduct structured consultations with representatives of affected communities, document community concerns, and explain how these concerns were addressed or why they were deemed not to warrant system modifications.

Content moderation governance provides a domain in which participatory approaches can operate immediately. Platforms should establish oversight boards with diverse membership that reflects affected communities, and these boards should have the authority to review policy decisions, examine enforcement patterns, access aggregated data about moderation outcomes, and issue binding decisions about specific cases. Meta's Oversight Board provides a partial model but falls short of meaningful participation because Meta selects board members, the board lacks access to training data and algorithmic details, and binding authority extends only to individual content decisions rather than policy choices or algorithmic design [21, pp. 567–568]. Genuine participatory governance would require community selection mechanisms, full access to information, and authority over systemic issues beyond individual cases.

6.5.1 Concrete oversight body framework

To operationalise participatory oversight, governance bodies should be structured to ensure affected communities have a meaningful voice while maintaining technical competence and operational viability. Composition should prioritise representation from the affected community, with mechanisms that ensure technical expertise and deployer input supplements rather than dominate community perspectives. One workable model allocates representation proportionally: affected community representatives (40–50%), independent technical experts (25–35%), and deployer representatives (20–25%). This composition ensures community priorities shape oversight while incorporating necessary technical and operational knowledge.

Selection mechanisms critically determine oversight legitimacy. Community representatives should be selected through civil society organisations with established relationships to affected populations, rather than appointed by deployers or government agencies. Rotating terms (e.g., two- to three-year staggered terms) prevent capture while building institutional knowledge. Technical experts should be selected based on demonstrated expertise in algorithmic systems, fairness metrics, and bias detection, with independence requirements excluding those with recent financial relationships to deployers. Oversight authority should include advisory powers over system design, deployment parameters, and monitoring practices for all systems, with escalated authority for high-risk deployments.

For systems affecting fundamental rights, oversight bodies should possess conditional veto authority over deployment in new domains until fundamental rights impact assessments are completed and community concerns are substantively addressed. Access

rights should include audit reports, technical documentation, and aggregated outcome data, with confidentiality agreements protecting legitimately sensitive information while preventing confidentiality claims from blocking meaningful oversight. Resource allocation determines whether oversight remains cosmetic or becomes effective. Deployers should fund oversight operations at levels that enable genuine scrutiny, scaled to the scope and impact of the deployment.

A proportional funding model (e.g., 0.3–0.5% of the system's operating budget) provides sufficient resources without incurring prohibitive costs. Funding should flow through independent intermediaries to prevent the deployer from exerting control over oversight activities. Transparency mechanisms ensure accountability. Oversight bodies should publish annual reports that document: the systems reviewed, the concerns identified, the recommendations made, the deployer responses, and the outcomes of disputed deployment decisions. This public reporting creates reputational incentives for deployers to address community concerns while enabling civil society, regulators, and researchers to track systemic patterns across deployments.

6.5.2 Implementation challenges and mitigation strategies

Participatory governance models confront documented implementation challenges requiring systematic institutional design. Ensuring the legitimacy of civil society organizations requires verification beyond self-designation. The identification process should assess organizations against concrete criteria that demonstrate genuine community relationships. Appropriate indicators include governance structures that enable affected community members to participate in organizational decision-making, sustained engagement records rather than episodic consultation, funding transparency, and demonstrated capacity to articulate community concerns through direct engagement rather than presumed representation.

Even with legitimate organizational representation, conflicting interests among communities require structured deliberation beyond preference aggregation. Meta's Oversight Board illustrates the consequences of inadequate deliberative structures: 80% of the Board's decisions were overturned on appeal, as decision-making lacked effective representation and structured deliberation [5]. Multi-stage deliberation, where panels review independently before collective discussion, enables minority viewpoints to be heard before majority positions coalesce, recognizing that preferences form through deliberation rather than as fixed inputs. When conflicts persist after deliberation, decision-making should proceed through supermajority voting requirements that prevent narrow majorities from overriding significant community opposition, with dissenting perspectives documented in public decisions and provisions for minority protection measures addressing specific harms identified by objecting communities.

However, deliberative processes alone do not guarantee meaningful power redistribution. Critical scholarship applies Arnstein's [7] participation ladder, documenting that many approaches remain at informing or consulting rather than achieving citizen power. The proposed oversight body addresses this through binding authority to block deployments, not merely advisory capacity. However, implementation requires legislative frameworks that specify jurisdictional scope, procedural requirements, and judicial review standards, ensuring that community input receives substantive rather than procedural consideration.

7 Limitations

This analysis acknowledges several limitations affecting interpretation and generalizability. First, the dual erosion framework presents descriptive categories rather than a normative resolution of when algorithmic discrimination is justified or how to balance competing interests. A reasonable disagreement exists about the appropriate trade-offs between accuracy and fairness, privacy and transparency, and innovation and regulation. The framework clarifies these tensions without definitively resolving them. Different societies may legitimately reach different conclusions about acceptable algorithmic risks based on distinctive legal traditions, cultural values, and institutional capacities.

Second, case selection prioritised systems where harms became visible through litigation, regulatory investigation, or journalistic exposure. This selection bias means the analysis may overestimate the prevalence of problematic systems or underestimate the effectiveness of existing safeguards for systems that function adequately without generating visible controversies. Additionally, the cases examined predominantly involve governmental or high-profile commercial deployments where documentation is available. Proprietary systems in contexts lacking public accountability may exhibit distinct patterns.

Third, comparative synthesis across jurisdictions with different legal traditions, regulatory capacities, and cultural contexts risks overlooking factors that make direct transplantation of legal solutions problematic. The Netherlands' robust data protection regime and cultural commitment to privacy facilitate judicial review of algorithmic systems, which might face insurmountable political obstacles in jurisdictions with weaker privacy protections or greater deference to technological solutions. Similarly, India's particular challenges with automated public distribution systems, in the context of widespread poverty, limited digital literacy, and infrastructure constraints, may not generalise to other developing nations with different conditions.

Fourth, the analysis focuses on cases where algorithmic systems produced discriminatory outcomes or obstructed remedies, whereas scholarly literature documents contexts in which properly designed algorithms may reduce discrimination compared to human decision-makers. Kleinberg et al. [36, pp. 288–289] demonstrate that bail algorithms optimised purely for risk prediction can reduce racial disparities compared to judicial decision-making in their analysis of over 750,000 defendants in New York City, where judges systematically made errors distributed unevenly across racial groups. Ludwig and Mullainathan [40, pp. 85–86] emphasise that the relevant comparison is between a realistic algorithm and a realistic human in actual institutional contexts, rather than idealised versions of either. The dual erosion framework remains applicable to these beneficial cases by clarifying the conditions enabling success. Algorithms that reduce equality risk through debiased prediction require transparency and validation to verify that debiasing succeeded, addressing remedy risk.

Conversely, algorithms designed for contestability must still ensure training data does not encode historical discrimination, addressing equality risk. The framework thus identifies that beneficial algorithmic deployment requires integrated attention to both dimensions rather than optimising one at the expense of the other. Future research should extend this analysis to additional domains where algorithms demonstrate superior performance, identifying design choices, oversight mechanisms, and deployment contexts that distinguish beneficial from harmful implementations.

Fifth, the proposed reforms assume judicial and regulatory capacity to implement complex technical oversight that may exceed realistic institutional capabilities. Discovery protocols that require court-appointed technical experts, burden-shifting frameworks that demand sophisticated statistical analysis, and participatory governance that necessitate sustained community engagement all rely on resources and expertise that may not be readily available in practice. Implementation research examining how proposed frameworks operate in real institutional contexts would provide crucial evidence about feasibility and necessary adaptations.

Finally, the analysis focuses on legal frameworks rather than technical solutions, including fairness-aware machine learning, algorithmic auditing methodologies, or privacy-preserving techniques that might address discrimination and opacity through system design rather than external regulation. This choice reflects disciplinary focus rather than judgment that technical solutions are unimportant. A comprehensive approach to algorithmic accountability requires integrating legal frameworks with technical safeguards, organisational practices, and cultural norms that collectively shape system development and deployment.

8 Conclusion

AI systems present distinctive challenges to fundamental rights by simultaneously entrenching discrimination and frustrating redress. The dual erosion framework formalises this observation into an analytical structure, distinguishing equality risk from remedy risk while emphasising their dynamic interaction. Equality risk encompasses biased training data, non-representative datasets, and discriminatory deployment contexts. Remedy risk encompasses algorithmic opacity, evidential asymmetries, and inadequate procedural safeguards. These dimensions interact to produce compound harms: opacity enables bias to persist undetected, while bias makes opacity more consequential.

Comparative case analysis suggests this framework applies across specific legal traditions and technological contexts. From COMPAS risk assessments in Wisconsin to SyRI welfare monitoring in the Netherlands, from automated public distribution systems in India to SCHUFA credit scoring in Germany, similar patterns emerge: algorithmic systems producing discriminatory outcomes that individuals cannot effectively contest. These patterns reveal systematic inadequacies in current legal frameworks that treat algorithmic systems as variations on traditional decision-making rather than recognising their distinctive characteristics, which require adapted governance.

Recent regulatory developments, including the EU AI Act (2024), the Digital Services Act (2022), and revised Santa Clara content moderation principles (2021), demonstrate a growing recognition that algorithmic accountability requires specific mechanisms beyond general principles. The AI Act's risk-based approach operationalises the dual erosion logic by requiring data governance practices that address equality risk and transparency obligations that address remedy risk for high-stakes systems. The DSA (2022) establishes procedural protections for content moderation, enabling users to understand the decisions made, contest determinations, and access external review. These frameworks provide a foundation for comprehensive governance but leave significant implementation challenges that require judicial and regulatory elaboration.

This article proposes five concrete reforms that translate analysis into implementable instruments. Discovery protocols adapted from complex commercial litigation enable

plaintiffs to access the algorithmic details necessary to prove discrimination while protecting legitimate confidentiality interests through protective orders. Burden-shifting frameworks, adapted from employment discrimination law, address evidential asymmetries by requiring defendants to prove the reasonableness of systems that produce disparate impacts. Risk-tiered due diligence standards, based on the EU AI Act (2024), establish graduated obligations proportional to the significance of the decision. Operationalised explanation requirements specify what constitutes meaningful information, enabling effective contestation. Participatory governance mechanisms integrate affected communities into design, deployment, and oversight processes.

These proposals respond to critiques that algorithmic accountability scholarship remains abstract by providing actionable standards for courts and regulators, though implementation will require adaptation to specific jurisdictional contexts and institutional capacities. They acknowledge that implementing comprehensive algorithmic governance requires sustained attention to capacity building, resource allocation, and institutional design. They recognise that technical solutions, organisational practices, and legal frameworks must operate synergistically rather than as substitutes for one another. They acknowledge that reasonable disagreement exists regarding the appropriate trade-offs between competing values.

The fundamental insight is that algorithmic accountability cannot focus exclusively on either equality or remedy but must address both dimensions and their interaction. Systems that appear technically neutral may produce systematic discrimination when deployed in contexts of structural inequality. Systems that provide adequate explanations in isolation may fail to enable meaningful contestation when evidential asymmetries prevent affected individuals from testing accuracy. Comprehensive governance requires integrated approaches, recognising that equality and remedy are interdependent rather than parallel concerns.

Future research should extend the dual erosion framework to additional domains, including healthcare algorithms, insurance underwriting, educational assessment, and emerging applications in autonomous vehicles and robotics. It should examine how different legal traditions approach tensions between innovation and regulation, developing comparative accounts of governance strategies across diverse institutional contexts. It should investigate participatory governance models that genuinely empower affected communities rather than providing symbolic inclusion. It should integrate technical and legal perspectives to develop comprehensive frameworks spanning system design, organisational deployment practices, and external accountability mechanisms.

The stakes are substantial. Algorithmic systems increasingly determine life opportunities, access to essential services, and freedom itself. Without effective governance ensuring these systems respect equality and provide meaningful recourse, society risks entrenching systemic discrimination beyond remedy. The dual erosion framework provides an analytical structure for understanding these risks and proposes implementable tools for addressing them, though effective application will require iterative refinement based on implementation experience and ongoing technical developments. Implementation will determine whether fundamental rights frameworks adapt successfully to algorithmic governance or whether the promise of fair and accountable decision-making becomes an algorithmic illusion.

Acknowledgements
N/A.

Author contributions
Corresponding Author Name: Bhavya Johari (Sole Author).

Funding
None.

Data availability
The author confirms that all data analysed during this study are included in this article. Furthermore, all primary and secondary sources supporting these findings are available through the references cited herein.

Declarations

Ethics approval and consent to participate
This research employed doctrinal legal analysis and comparative case synthesis of publicly available legal decisions, regulatory documents, and peer-reviewed literature. The study did not involve human subjects research, experimental interventions, or primary data collection. Ethical approval was therefore not required. Not applicable. This research did not involve human participants.

Consent for publications
Not applicable. This research did not involve any individual persons' data that required consent for publication.

Competing interests
The author declares none.

Received: 4 December 2025 / Accepted: 12 February 2026
Published online: 18 February 2026

References

1. Adams-Prassl J, Binns R, Kelly-Lyth A. Directly discriminatory algorithms. Mod Law Rev. 2023;86(1):144–75. https://doi.org/10.1111/1468-2230.12759.
2. Amnesty International. (2021). Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal (Report EUR 35/4686/2021).
3. Amnesty International. (2024a). Coded injustice: Surveillance and discrimination in Denmark's automated welfare state (Report EUR18/8709/2024). https://www.amnesty.org/en/documents/eur18/8709/2024/en/
4. Amnesty International. (2024b). Use of entity resolution in India: Shining a light on how new forms of automation can deny people access to welfare. https://www.amnesty.org/en/latest/research/2024/04/entity-resolution-in-indias-welfare-digitalization/
5. Ang PH, Haristya S. The governance, legitimacy and efficacy of Facebook's oversight board: a model for global tech platforms? Emerging Media. 2024;2(2):169–80. https://doi.org/10.1177/27523543241266860.
6. Are C, Briggs R. "Dysfunctional" appeals and failures of algorithmic justice in Instagram and TikTok content moderation. Inf Commun Soc. 2024;27(10):1997–2014. https://doi.org/10.1080/1369118X.2024.2396621.
7. Arnstein SR. A ladder of citizen participation. J Am Inst Plann. 2007;35(4):216–24. https://doi.org/10.1080/01944366908977225.
8. Barocas S, Selbst AD. Big data's disparate impact. Calif Law Rev. 2016;104:671–732.
9. Barocas S, Hardt M, Narayanan A. Fairness and machine learning: Limitations and opportunities. MIT Press; 2023.
10 Bartlett L, Vavrus F. Rethinking case study research: A comparative approach. Routledge; 2017.
11. Bhat PI. Doctrinal legal research as a means of synthesizing facts, thoughts, and legal principles. In Idea and methods of legal research (pp. 143–168). Oxford University Press, 2020. https://doi.org/10.1093/oso/9780199493098.003.0005.
12 Blankertz A, Jaursch J. Responses to the European Commission's consultation on the Digital Services Act (DSA). Stiftung Neue Verantwortung; 2020.
13. Chaudhuri B. Programmed welfare: an ethnographic account of algorithmic practices in the public distribution system in India. New Media Soc. 2022;24(4):887–902. https://doi.org/10.1177/14614448221079034.
14. Chen Z, Zhang JM, Sarro F, Harman M. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. ACM Trans Softw Eng Methodol. 2023;32(4):106. https://doi.org/10.1145/3583561.
15. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data. 2017;5(2):153–63. https://doi.org/10.1089/big.2016.0047.
16. Citron DK. Technological due process. Wash U Law Rev. 2008;85(6):1249–313.
17. Court of Justice of the European Union. (2023). SCHUFA Holding AG (Scoring), Case C-634/21 (7 December 2023).
18. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451). Association for Computational Linguistics, 2020. https://aclanthology.org/2020.acl-main.747/.
19. Disparate impact and reasonable factors other than age under the Age Discrimination in Employment Act, 29 C.F.R. § 1625.7 (2012). https://www.ecfr.gov/current/title-29/section-1625.7
20. District Court of The Hague. (2020). NJCM et al. and FNV v. The State of the Netherlands, ECLI:NL:RBDHA:2020:865 (5 February 2020).
21. Douek E. Content moderation as systems thinking. Harv Law Rev. 2021;135:526–607.

22. Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Sci Adv. 2018;4(1):eaao5580. https://doi.org/10.1126/sciadv.aao5580.
23. Eubanks V. Automating inequality: How high-tech tools profile, police, and punish the poor. St: Martin's Press; 2018.
24. European Convention for the Protection of Human Rights and Fundamental Freedoms, Nov. 4, 1950, 213 U.N.T.S. 221.
25. Ewert v Canada. [2018] SCC 30, Supreme Court of Canada.
26. Federal Circuit. (2023). Royal Brush Manufacturing v. United States, No. 22–1226 (27 July 2023).
27. Federal Rules of Civil Procedure, 28 U.S.C. 26(c) (1938). https://www.uscourts.gov/rules-policies/current-rules-practice-procedure/federal-rules-civil-procedure
28. Federal Rules of Evidence, 28 U.S.C. 104(a) (1975). https://www.uscourts.gov/rules-policies/current-rules-practice-procedure/federal-rules-evidence
29. Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (pp. 329–338). Association for Computing Machinery, 2019. https://doi.org/10.1145/3287560.3287589.
30. Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A. Predictably unequal? The effects of machine learning on credit markets. J Finance. 2022;77(1):5–47.
31. Henman P. Improving public services using artificial intelligence: possibilities, pitfalls, governance. Asia Pac J Public Adm. 2020;42(4):209–21. https://doi.org/10.1080/23276665.2020.1816188.
32. Human Rights Committee. (2004, March 29). General Comment No. 31: The nature of the general legal obligation imposed on States Parties to the Covenant (UN Doc CCPR/C/21/Rev.1/Add.13). United Nations.
33. Hutchinson T, Duncan N. Defining and describing what we do: doctrinal legal research. Deakin Law Rev. 2013;17(1):83–119.
34. Köchling A, Wehner MC. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Bus Res. 2020;13(3):795–848. https://doi.org/10.1007/s40685-020-00134-w.
35. Kurumayya V. Towards fair AI: A review of bias and fairness in machine intelligence. J Comput Soc Sci. 2025;55:1–26. https://doi.org/10.1007/s42001-025-00386-8.
36. Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. Human decisions and machine predictions. Q J Econ. 2018;133(1):237–93. https://doi.org/10.1093/qje/qjx032.
37. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. Proceedings of Innovations in Theoretical Computer Science, 2017. arXiv:1609.05807. https://doi.org/10.48550/arXiv.1609.05807.
38. Larson J, Mattu S, Kirchner L, Angwin J. How we analyzed the COMPAS recidivism algorithm. ProPublica; 2016.
39. Liu R, Yang R, Jia C, Zhang G, Zhou D, Dai AM, Yang D, Vosoughi S. Scaling reinforcement learning for content moderation with large language models, 2025. arXiv preprint arXiv:2512.20061. https://doi.org/10.48550/arXiv.2512.20061.
40. Ludwig J, Mullainathan S. Fragile algorithms and fallible decision-makers: lessons from the justice system. J Econ Perspect. 2021;35(4):71–96. https://doi.org/10.1257/jep.35.4.71.
41. Marda V, Sinha A. FinTech lending in India: taking stock of implications for privacy and autonomy. Indian J Law Technol. 2022;18(1):6.
42. Meacham v. Knolls Atomic Power Laboratory, 554 U.S. 84 (2008).
43. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):115. https://doi.org/10.1145/3457607.
44. Meta AI. (2021, December 8). Harmful content can evolve quickly. Our new AI system adapts to tackle it. https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/
45. Mobley v. Workday, Inc., No. 23-cv-00770 (N.D. Cal. Feb. 21, 2023). Class Action Complaint.
46. Mulligan DK, Bamberger KA. Procurement as policy: administrative process for machine learning. Berkeley Technol Law J. 2019;34:773–852.
47. Nielsen A. AI's categorical fairness. Am J Law Equal. 2025;5:89–120. https://doi.org/10.1162/AJLE.a.5.
48. Nagaraj A. How an algorithm denied food to thousands of poor in India's Telangana. Al Jazeera, 2024. https://www.aljazeera.com/economy/2024/1/24/how-an-algorithm-denied-food-to-thousands-of-poor-in-indias-telangana.
49. Rafaela S, Semedo M, Herzberger-Fofana P, Ernst C, Incir E, González MS, Carême D, Van Sparrentak K, Rego S, Pineda M, Melchior K, Nemec M, Lagodinsky S, Chahim M, in 't Veld S, Marquardt E, Donáth AJ, Mebarek N, Andrieu E, Carvalhais I. The Dutch childcare benefit scandal, institutional racism and algorithms (Question O-000028/2022) [Question for oral answer, European Parliament], 2022. https://www.europarl.europa.eu/doceo/document/O-9-2022-000028_EN.html.
50. Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020 (pp. 469–481). ACM. https://doi.org/10.1145/3351095.3372828.
51. Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), OJ L 119/1 (2016).
52. Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act), OJ L 277/1 (2022).
53. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L 207/1 (2024).
54. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x.
55. Salih A, Elsayed A, Challita K, Pharaon M, Abou-Abbas L, Serhani MA. A perspective on explainable artificial intelligence methods: SHAP and LIME. Adv Intell Syst. 2024;6:2400304. https://doi.org/10.1002/aisy.202400304.
56. Santa Clara Principles on Transparency and Accountability in Content Moderation, Second Iteration. (2021). https://santaclaraprinciples.org/
57. Sap M, Card D, Gabriel S, Choi Y, Smith NA. The risk of racial bias in hate speech detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1668–1678, 2019. https://doi.org/10.18653/v1/P19-1163.
58. State v. Loomis, 881 N.W.2d 749 (Wis. 2017).

59. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Article 17, pp. 1–9), 2021. Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305.
60. United States v. Zolin, 491 U.S. 554 (1989).
61. Watamura E, Hu L, Jung J. A fairness-focused approach to recidivism prediction: Implications for accuracy, trust, and equity. AI & Soc Adv Online Publ. 2025. https://doi.org/10.1007/s00146-025-02452-1.
62. Wexler R. Life, liberty, and trade secrets: intellectual property in the criminal justice system. Stanf Law Rev. 2018;70:1343–429.
63. Wilson K, Caliskan A. Gender, race, and intersectional bias in resume screening via language model retrieval. In Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024) (pp. 1578–1590). Association for the Advancement of Artificial Intelligence, 2024. https://doi.org/10.1609/aies.v7i1.31748.

Publisher's note