Sentam Research & Publishing

Vol. 2, Iss. 1 (2026)
Pages: 33-46

Open Journal of Interdisciplinary Studies

Volume 2, Issue 1, 2026

Journal homepage: https://sentam.org/journals/ojis



When Fairness Metrics Conflict: A Cross-Domain Analysis of Algorithmic Bias and the Regulatory Specification Gap

Tariq O. Mensah[1], Priya Subramanian[2] & Dario Escobar Llanos[3]*

[1]Bolgatanga Technical University, Ghana
[2]Manipur International University, India
[3]Universidad de Oriente, Venezuela

ABSTRACT

Algorithmic systems now govern consequential decisions in criminal justice, healthcare, and employment, yet persistent evidence demonstrates that these systems reproduce and amplify the very inequities they were designed to transcend. This paper advances the Fairness Metric Misalignment Framework (FMMF) as a structured analytical lens for understanding why algorithmic accountability failures persist even when individual actors comply with their respective obligations. Drawing on published audit studies, peer-reviewed empirical research, and regulatory documents across three high-stakes domains, the paper demonstrates that the root cause of accountability failure is not technical malfunction but metric selection incoherence: different stakeholders implicitly adopt incompatible fairness definitions, audit studies measure incommensurable outcomes, and regulatory frameworks mandate compliance with standards whose mathematical preconditions are mutually exclusive under real-world conditions. The FMMF organizes this diagnosis around three components: a metric taxonomy mapping the normative assumptions embedded in five major fairness definitions; a cross-domain mapping revealing systematic misalignment between the fairness metrics used in audit studies, system design objectives, and regulatory mandates; and a regulatory translation gap analysis demonstrating that major instruments including the European Union Artificial Intelligence Act (2024), the United States Equal Employment Opportunity Commission Technical Assistance (2023), and New York City Local Law 144 (2021) deploy fairness language without specifying the mathematical criteria that would render compliance verifiable. The paper concludes that adequate algorithmic governance requires regulators to specify fairness metrics explicitly, acknowledge the trade-offs their chosen metrics foreclose, and establish cross-domain coordination mechanisms that prevent governance fragmentation from insulating discriminatory systems from accountability. These findings carry direct implications for policymakers, system auditors, and institutional deployers navigating the emerging regulatory landscape for high-risk artificial intelligence.

Keywords:
algorithmic bias; fairness metrics; algorithmic accountability; AI governance; disparate impact; COMPAS; healthcare algorithms; hiring algorithms

INTRODUCTION

Artificial intelligence systems embedded in high-stakes decision-making contexts occupy an unusual position in contemporary governance discourse. They are simultaneously celebrated as instruments of objectivity and condemned as engines of discrimination. The apparent contradiction is not accidental; it reflects a deeper

* Corresponding Author: Bolgatanga Technical University, Ghana; Email: tariq.mensah@gmail.com

Published: June 11, 2026

This is an open access article licensed under CC BY license (http://creativecommons.org/licenses/by/4.0/).

structural problem that existing legal, regulatory, and technical frameworks have failed to resolve. Algorithms trained on historical data encode the distributional consequences of past discrimination, and when deployed to predict future behavior, allocate healthcare resources, or screen employment candidates, they reproduce those consequences at scale and with the procedural authority of computational neutrality (Angwin et al., 2016; Obermeyer et al., 2019; Raghavan et al., 2020).

The scholarly and policy literature has responded to this problem with considerable energy. Audit studies have documented racial disparities in recidivism prediction instruments, health risk scoring algorithms, and pre-employment assessment tools. Computer scientists have formalized competing definitions of fairness and proved that several cannot be satisfied simultaneously. Legal scholars have mapped existing anti-discrimination doctrine onto algorithmic contexts. Regulators in the European Union, the United States, and New York City have enacted or issued instruments that classify certain algorithmic systems as high-risk and impose accountability obligations on their developers and deployers. Yet the discrimination persists, and the governance literature has not provided a satisfying account of why.

This paper argues that the persistence of algorithmic discrimination is not primarily explained by inadequate regulation, insufficient technical sophistication, or bad faith on the part of system developers. It is explained by metric selection incoherence: the systematic failure to align the fairness criteria used in audit studies, system design, and regulatory mandates with each other and with the mathematical constraints that real-world conditions impose. When ProPublica documented racial bias in the COMPAS recidivism prediction instrument and Northpointe contested that finding in 2016, both organizations were correct under their respective fairness definitions (Angwin et al., 2016; Dieterich et al., 2016). That both could be simultaneously correct without contradiction is not a failure of analysis; it is a structural feature of the fairness impossibility results subsequently formalized by Chouldechova (2017) and Kleinberg et al. (2016). Yet neither the original controversy nor the subsequent regulatory responses have systematically confronted this structural feature or drawn the governance implications that follow from it.

To address this gap, the paper advances the Fairness Metric Misalignment Framework (FMMF), an original analytical instrument that organizes the diagnosis of algorithmic accountability failure around three components: a metric taxonomy that makes explicit the normative assumptions embedded in competing fairness definitions; a cross-domain empirical mapping that traces metric misalignment across criminal justice, healthcare, and employment; and a regulatory translation gap analysis that evaluates whether major regulatory instruments specify fairness criteria with sufficient precision to render accountability meaningful. The FMMF does not propose a solution to the impossibility results; those results are mathematical theorems, not policy failures. It proposes instead a governance architecture that acknowledges the trade-offs embedded in metric choice, requires regulators to specify those choices explicitly, and creates accountability structures capable of holding system deployers to the standards actually chosen rather than standards expressed in fairness language too vague to enforce.

THEORETICAL BACKGROUND

Fairness, Mathematics, and the Impossibility Results

The concept of fairness in algorithmic systems is not a single coherent idea but a family of formally distinct criteria whose relationships to each other are the subject of active mathematical investigation. Understanding why algorithmic accountability is structurally difficult requires a precise account of those criteria and of the conditions under which they conflict. Five definitions have dominated the technical and applied literature, and each encodes a different answer to the normative question of what equal treatment requires.

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

Demographic parity, the most intuitive definition, requires that a classifier produce positive outcomes at equal rates across protected groups. An employment screening tool satisfies demographic parity if it advances Black and white candidates to the next hiring stage in equal proportions. This definition is sometimes called statistical parity and corresponds most directly to the legal concept of adverse impact as operationalized in the United States federal four-fifths rule (Equal Employment Opportunity Commission., 2023). Its normative assumption is that the relevant comparison is between group-level selection rates; it requires nothing about whether the selected individuals within each group have equal predicted or actual qualifications.

Equalized odds, formalized by Hardt et al. (2016), requires both equal true positive rates and equal false positive rates across groups. A recidivism prediction instrument satisfies equalized odds if it correctly identifies those who will reoffend at the same rate across racial groups, and if it incorrectly flags those who will not reoffend at the same rate across racial groups. This definition directly addresses the concern identified in the ProPublica analysis of COMPAS: that Black defendants who did not reoffend were flagged as high-risk at nearly twice the rate of white defendants in the same situation (Angwin et al., 2016). Its normative assumption is that the costs of false positives and false negatives should be distributed equally across groups.

Predictive parity, also called calibration, requires that a risk score carry the same meaning across groups: among all individuals assigned a given risk score, the actual rate of the predicted outcome should be equal regardless of group membership. The Northpointe response to ProPublica's COMPAS analysis invoked precisely this criterion, demonstrating that among all defendants assigned a high-risk score by COMPAS, the actual recidivism rate was approximately equal for Black and white defendants (Dieterich et al., 2016). Its normative assumption is that the score's predictive accuracy is the relevant fairness criterion, not the distribution of prediction errors across groups.

Individual fairness, proposed by Dwork et al. (2012), requires that similar individuals be treated similarly, where similarity is defined with respect to the task at hand. This definition avoids the group-level framing of the preceding three criteria and addresses concerns about the arbitrary assignment of group membership. Its normative assumption is that the unit of moral concern is the individual rather than the group, and its practical limitation is that defining a task-appropriate similarity metric requires prior agreement on which attributes are task-relevant.

Counterfactual fairness, proposed by Kusner et al. (2017), requires that an individual's predicted outcome not change in a counterfactual world in which their protected attribute was different while all other causally upstream variables were held constant. This definition engages most directly with the causal structure of discrimination and is the most technically demanding to operationalize, requiring a causal model of the data-generating process that is rarely available in practice.

The impossibility results established by Chouldechova (2017) and Kleinberg et al. (2016) demonstrate formally that demographic parity, equalized odds, and predictive parity cannot all be simultaneously satisfied by any classifier when the base rate of the predicted outcome differs across groups, except in the degenerate case of a perfect classifier. This result is not a statement about bad algorithm design; it is a mathematical theorem about the logical structure of these criteria under realistic conditions. Since base rates differ across racially defined groups for virtually every outcome that high-stakes algorithms are deployed to predict, including recidivism, healthcare utilization, and creditworthiness, the result applies to virtually every consequential deployment context. The practical implication is that choosing a fairness definition is not a technical decision but an inherently normative one: it determines which group bears the cost of the trade-off that cannot be avoided.

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

The existing literature has documented these impossibility results extensively within individual domains, most thoroughly in criminal justice. What has not been done is a systematic cross-domain analysis that maps which fairness metrics different stakeholders have implicitly or explicitly adopted across domains, traces the regulatory consequences of those choices, and evaluates whether current governance instruments are structured to manage the trade-offs the impossibility results impose. That is the contribution the FMMF is designed to make.

METHODOLOGY

This paper employs secondary data analysis as its primary methodological approach. Secondary data analysis involves the extraction, reanalysis, and reframing of data generated by prior studies to address research questions not posed or fully answered in the original research. It is distinct from literature review in that it treats quantitative findings from published studies as data points subject to comparative analysis, rather than treating those studies as arguments to be synthesized qualitatively. The methodological contribution of this paper is the organization of existing findings through a common analytical framework, the FMMF, that renders them comparable across domains in ways the original studies do not attempt.

Domain selection followed three criteria. First, the domain had to have produced published audit studies containing extractable quantitative findings on algorithmic bias, specifically studies reporting performance metrics disaggregated by race or other protected characteristics. Second, the domain had to be subject to an identifiable regulatory or governance response that could be evaluated for fairness metric specificity. Third, the domain had to represent a distinct institutional context with its own data practices, outcome variables, and accountability structures, so that cross-domain comparison would be analytically informative rather than redundant. Criminal justice, healthcare, and employment each satisfy all three criteria. Credit scoring was considered but excluded because the most instructive audit data in that domain is proprietary and not publicly available in a form extractable for secondary analysis.

Data extraction was conducted from three source categories. The first category comprised empirical audit studies published in peer-reviewed journals or as formally released investigative documents: Angwin et al. (2016) and Dieterich et al. (2016) for criminal justice; Obermeyer et al. (2019) and Sjoding et al. (2020) for healthcare; and Raghavan et al. (2020) for employment. From these sources, the paper extracted confusion matrix statistics, accuracy metrics, effect sizes, and descriptions of the outcome variables used to operationalize the prediction task. The second category comprised foundational theoretical papers establishing the impossibility results: Chouldechova (2017) and Kleinberg et al. (2016). From these, the paper extracted the formal conditions under which the impossibility results hold and the specific fairness criteria in conflict. The third category comprised regulatory and legal instruments: the European Union Artificial Intelligence Act (Regulation (EU) 2024/1689) (European Parliament and Council, 2024), the EEOC Technical Assistance of May 18, 2023 (Equal Employment Opportunity Commission., 2023), and New York City Local Law 144 of 2021 (Office of the New York State Comptroller, 2025). From these, the paper extracted the fairness language used, any metric specifications provided, the compliance obligations imposed, and the enforcement mechanisms established.

The Fairness Metric Misalignment Framework was constructed inductively from these materials through a three-stage process. In the first stage, the metric taxonomy was built by systematically coding each audit study and regulatory instrument for the fairness criterion it explicitly or implicitly invoked, producing a structured vocabulary for cross-domain comparison. In the second stage, the domain-metric mapping was produced by applying the taxonomy to each domain: for each domain, a mapping table was constructed identifying the fairness metric used in the audit study, the metric the deployed system appeared to optimize, and the metric invoked by the applicable

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

regulatory instrument. In the third stage, the regulatory translation gap was assessed by evaluating, for each regulatory instrument, whether the fairness language deployed was associated with a specified mathematical criterion, and whether that specification, where present, was coherent with what the impossibility results establish as technically achievable under the conditions present in that domain.

Limitations of this methodology warrant acknowledgment. Secondary data analysis is constrained by the quality and scope of the primary studies from which data is drawn. In the employment domain in particular, the most widely deployed hiring algorithms are proprietary systems whose internal workings and training data are not publicly disclosed; the available audit evidence therefore covers a sample of vendors and does not permit generalization to the full range of deployed systems. In the healthcare domain, the Obermeyer et al. (2019) study examined one algorithm at one health system; the extent to which its findings generalize to other systems is an open empirical question. These limitations do not undermine the core argument, which concerns the structural relationship between fairness metric choices and regulatory frameworks rather than the prevalence of bias in any specific system. They do suggest that the FMMF's domain-metric mapping should be understood as an illustration of the framework's diagnostic capacity rather than a comprehensive audit of algorithmic bias across these sectors.

DOMAIN ANALYSIS

Criminal Justice

The criminal justice domain provides the most extensively documented and methodologically contested case of algorithmic bias, and for that reason it is the most illuminating site for demonstrating how metric misalignment operates. The COMPAS recidivism prediction instrument, developed by Northpointe (now Equivant) and used in pretrial risk assessments across multiple United States jurisdictions, became the subject of a nationally prominent dispute in 2016 when ProPublica published an investigative analysis documenting racial disparities in its predictions (Dieterich et al., 2016). The subsequent controversy, which generated over two hundred academic citations within eighteen months of the original publication, is the canonical empirical case for the fairness impossibility results.

ProPublica's analysis of COMPAS predictions for a sample of defendants in Broward County, Florida documented that Black defendants who did not reoffend within two years were flagged as high-risk at a rate of 44.9%, compared to 23.5% for white defendants in the same situation. Conversely, white defendants who did reoffend were flagged as low-risk at a rate of 47.7%, compared to 28.0% for Black defendants (Angwin et al., 2016). These disparities in false positive and false negative rates constitute a violation of the equalized odds criterion: the costs of prediction error were not distributed equally across racial groups. Black defendants bore a substantially higher risk of being incorrectly designated as dangerous, with concrete consequences for bail decisions, sentencing recommendations, and parole outcomes.

Northpointe's response, published in July 2016, did not contest the accuracy of ProPublica's statistical calculations. It instead argued that ProPublica had applied the wrong fairness criterion. Analyzing the same Broward County data, Dieterich et al. (2016) demonstrated that COMPAS satisfied predictive parity: among all defendants assigned a given risk score, the actual two-year recidivism rate was approximately equal for Black and white defendants. A high-risk score assigned by COMPAS carried the same predictive meaning regardless of the defendant's race. The system was calibrated, in the technical sense, across racial groups.

Chouldechova (2017) subsequently provided the mathematical proof that explains why both analyses are simultaneously correct. When the base rate of the predicted outcome differs across groups, which is true of recidivism as measured in the Broward County data, a classifier that satisfies predictive parity will necessarily violate equalized odds unless it is a perfect classifier. Since no

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

recidivism prediction instrument achieves perfect accuracy, the trade-off is unavoidable: choosing predictive parity forecloses equalized odds, and vice versa. The COMPAS controversy was not, therefore, a dispute about statistical facts; it was a dispute about which fairness criterion should govern recidivism prediction, a normative question that neither party to the dispute explicitly named as such.

Dressel and Farid (2018) added a further dimension to this analysis by demonstrating that COMPAS predictions were no more accurate than predictions made by individuals recruited through Amazon Mechanical Turk with access to only two features: age and number of prior convictions. COMPAS's overall accuracy of 65.2% was not statistically distinguishable from the lay prediction accuracy of 67.0% when defendants' race was not disclosed to predictors. Despite collecting 137 features on each defendant, the algorithm's predictive performance did not exceed that of a simple two-variable model applied by untrained individuals. This finding does not directly implicate metric misalignment, but it raises a prior question: if the algorithm does not outperform simpler approaches in predictive accuracy, the normative justification for deploying it in consequential decisions is weakened regardless of which fairness criterion it satisfies.

From the perspective of the FMMF, the criminal justice domain establishes the canonical illustration of the metric taxonomy's first component and the domain-metric mapping's central diagnostic. The ProPublica-Northpointe dispute maps precisely onto the equalized odds versus predictive parity conflict that Chouldechova (2017) formalized. The regulatory response to this dispute has been fragmented and largely metric-unspecified. Most jurisdictions using COMPAS or similar instruments have not enacted legislation specifying which fairness criterion governs their use. The risk assessment instruments continue to be deployed, the fairness dispute continues to be unresolved, and the governance literature has not compelled a normative resolution of the trade-off.

Healthcare

The healthcare domain presents a structurally different form of metric misalignment from the criminal justice case. In criminal justice, the bias emerged from the interaction of a classifier's design with unequal base rates across groups, and the dispute was about which fairness criterion to apply to a classifiable prediction task. In healthcare, the most consequential documented case of algorithmic bias arose from a design-stage decision about the outcome variable itself, a decision that was not framed as a fairness choice at all and that therefore evaded the fairness critique entirely until an external audit exposed the consequences.

Obermeyer et al. (2019) examined a widely deployed commercial algorithm used to identify patients with complex health needs for enrolment in care management program. The algorithm was designed to predict healthcare utilization, measured as future medical costs, and used that prediction as a proxy for health needs. The implicit assumption was that patients who will consume more healthcare resources are, by definition, patients with greater health needs. The study found that this assumption introduced a systematic racial bias: at any given predicted risk score, Black patients had substantially more active chronic conditions than white patients assigned the same score. The algorithm falsely concluded that Black patients with a given level of illness were healthier than equally ill white patients, because the historical data reflected a reality in which Black patients with the same conditions received less healthcare. Less care received translated to lower historical costs, which translated to lower predicted future costs, which translated to lower risk scores, which translated to reduced access to care management resources.

The quantitative findings are precise. Among patients assigned the same risk score by the algorithm, Black patients had approximately 26.3% more active chronic conditions than white patients. The study estimated that this bias reduced the proportion of Black patients identified for high-risk care management from the rate that

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

would prevail if needs were measured directly rather than through cost proxies. After reformulating the algorithm to use health needs indicators directly rather than cost as a proxy, the proportion of Black patients identified for care management programs increased substantially, with the researchers estimating the racial composition of the enrolled population would shift markedly toward equity.

Sjoding et al. (2020) documented a related but distinct form of healthcare algorithmic bias that operates through a medical device rather than a predictive scoring system. Pulse oximetry, the standard clinical tool for non-invasive measurement of blood oxygen saturation, was found to produce systematically less accurate readings for Black patients than for white patients. In a cohort study analyzing pairs of pulse oximetry measurements and arterial blood gas measurements from patients at the University of Michigan and from intensive care units at 178 other hospitals, Black patients exhibited occult hypoxemia at approximately three times the rate of white patients: an arterial oxygen saturation below 88% despite a pulse oximeter reading in the 92% to 96% range was detected in 11.7% of measurements for Black patients compared to 3.6% for white patients. Since pulse oximetry readings trigger clinical decisions about supplemental oxygen therapy and, during the COVID-19 pandemic, about eligibility for treatment interventions, the accuracy disparity carried direct clinical consequences for patient outcomes.

The healthcare domain extends the FMMF's metric taxonomy in an important direction. The Obermeyer et al. (2019) case is not a case of metric misalignment in the technical sense established in the criminal justice analysis. The algorithm was never evaluated against a fairness criterion at all because the bias was embedded in the choice of outcome variable, not in the model's predictive structure. This represents a category of algorithmic accountability failure that precedes the fairness metric debate: design-stage outcome variable selection decisions that introduce bias upstream of any metric evaluation. The FMMF identifies this as a distinct failure mode, outcome variable misspecification, that existing regulatory frameworks are particularly ill-equipped to address because they typically require post-deployment auditing against specified fairness criteria rather than prospective evaluation of whether the outcome variable adequately captures the intended social good.

The pulse oximetry case extends the analysis further by demonstrating that algorithmic bias in healthcare is not limited to predictive scoring software but encompasses the measurement infrastructure on which clinical algorithms depend. If the inputs to an algorithm are racially biased, the algorithm's outputs will be racially biased regardless of how carefully the model is designed or which fairness criterion it optimizes. This upstream measurement bias problem is not addressed by any of the major regulatory frameworks examined in the regulatory translation gap analysis, creating a governance blind spot with direct implications for patient safety.

Employment

The employment domain is the site of the most active recent regulatory development in algorithmic accountability, and it illustrates the regulatory translation gap most directly. The combination of a growing commercial market for pre-employment assessment tools, documented evidence of discriminatory outcomes, and a layered regulatory response at federal, state, and municipal levels makes employment the domain where the gap between fairness language and metric specification is most consequential and most documentable.

Raghavan et al. (2020) conducted an audit of the claims and practices of companies offering algorithmic pre-employment assessment tools, examining the publicly available documentation from eighteen vendors whose products are used to screen candidates. The study found that vendors used inconsistent and non-comparable definitions of fairness in their marketing and validation materials. Most vendors did not disclose the specific fairness metric their tools optimized.

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

Validation studies, where disclosed at all, typically reported overall predictive accuracy for the intended workforce task but did not disaggregate performance metrics by race, gender, or other protected characteristics. The study identified this opacity as a structural feature of the vendor market rather than an isolated lapse: commercial incentives favor fairness representations that are broad enough to satisfy diverse buyer requirements without committing to specific metric choices that would expose the trade-offs the impossibility results impose.

The Amazon hiring algorithm case, disclosed through Reuters investigative reporting in 2018 and subsequently confirmed in Amazon's public communications, illustrated how bias can emerge in a high-profile deployment even when the developing organization has explicit commitments to equitable outcomes. The algorithm, trained on ten years of Amazon's own hiring data, assigned lower scores to resumes containing the word "women's" and to graduates of historically women's colleges (Ahmad et al., 2026; Dastin, 2018). The bias was not a design intention; it was a consequence of training on historical data that reflected a pre-existing gender imbalance in Amazon's technical hiring. Amazon discontinued the tool before deploying it in consequential decisions, but the case illustrates the mechanism by which historical discrimination is encoded into algorithmic predictions trained to replicate past outcomes.

The Electronic Privacy Information Center filed a complaint against HireVue with the Federal Trade Commission in 2019, documenting concerns about the company's use of facial analysis in video interviews to assess candidate suitability. HireVue subsequently discontinued its facial analysis component in 2021, citing advances in the science and shifts in the understanding of the tool's capabilities (EPIC, 2019; Harwell, 2019). The complaint and the company's response together document the lifecycle of a specific assessment modality from deployment through regulatory pressure to discontinuation without any formal adjudication of whether the tool produced discriminatory outcomes, illustrating the limits of complaint-based enforcement as an accountability mechanism.

The regulatory response in the employment domain has produced the most specific metric requirements of any domain examined in this paper, but even the most specific instruments fall short of what the impossibility results require. The EEOC Technical Assistance issued on May 18, 2023, titled "Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII," operationalizes its fairness requirement through the adverse impact concept inherited from the Uniform Guidelines on Employee Selection Procedures. Adverse impact, as measured by the four-fifths rule, is a demographic parity-adjacent criterion: it is triggered when a selection tool advances candidates from a protected group at a rate less than 80% of the rate for the group with the highest selection rate (Equal Employment Opportunity Commission., 2023). The guidance recommends that employers ask vendors whether they relied on the four-fifths rule or other court-approved standards when auditing their products.

New York City Local Law 144 of 2021, which took effect on January 1, 2023 with enforcement commencing on July 5, 2023, is the most operationally specific regulatory instrument in the employment domain. It requires employers and employment agencies using automated employment decision tools in hiring or promotion decisions affecting New York City employees or candidates to commission an independent annual bias audit and to publicly post the results. The bias audit is defined as including assessment of the tool's disparate impact on protected categories as specified in federal equal employment opportunity reporting requirements. This effectively anchors the audit requirement to a demographic parity-adjacent criterion. The law does not require auditors to assess whether the tool satisfies equalized odds, predictive parity, or any other fairness criterion; it does not specify how auditors should handle the trade-offs the impossibility results impose; and it does not require vendors to disclose the fairness criterion the tool was designed to optimize.

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

From the FMMF's perspective, the employment domain illustrates the regulatory translation gap in its clearest form. The regulatory instruments have moved closer to metric specification than their counterparts in criminal justice, but they have converged on demographic parity or adverse impact as the operative criterion without acknowledging that this choice forecloses other fairness criteria under realistic conditions. An employer whose hiring algorithm satisfies the four-fifths rule is in compliance with both the EEOC guidance and NYC Local Law 144 even if the tool systematically produces higher false negative rates for protected group members, which means qualified candidates from protected groups are disproportionately screened out despite meeting the relevant criteria.

Cross-Domain Synthesis: Patterns of Metric Misalignment

Comparing the three domain analyses through the FMMF reveals several structural patterns that single-domain studies cannot identify. These patterns constitute the paper's primary empirical contribution: they demonstrate that metric misalignment is not a domain-specific anomaly but a systemic feature of how algorithmic accountability has been constructed across consequential deployment contexts.

The first pattern is the consistent invisibility of metric choice in audit studies. With the partial exception of the ProPublica-Northpointe exchange, none of the major audit studies examined in this paper explicitly state the fairness criterion they are using as a normative standard or acknowledge that alternative criteria would produce different findings. Obermeyer et al. (2019) frame their analysis in terms of the gap between cost predictions and health needs, which implicitly invokes an individual fairness or outcome equity criterion, but do not engage with the formal fairness literature or acknowledge that their corrective proposal, reformulating the algorithm to use health needs indicators, would itself require a choice among the competing definitions. Sjoding et al. (2020) document a measurement accuracy disparity without mapping it onto the fairness taxonomy. Raghavan et al. (2020) note the inconsistency of vendor fairness claims but do not systematically evaluate vendor practices against the formal criteria. The consequence is that audit findings cannot be compared across domains or across studies within the same domain, because the criteria against which performance is measured are not standardized.

The second pattern is the systematic mismatch between the fairness criterion implicitly invoked in audit studies and the criterion operative in the relevant regulatory instrument. In criminal justice, ProPublica's analysis invoked equalized odds while no major regulatory instrument in that domain specifies equalized odds as a compliance requirement. In employment, the EEOC and NYC Local Law 144 invoke demographic parity or its adverse impact analogue, but the audit literature in that domain is more concerned with the predictive validity of assessment tools than with selection rate parity. In healthcare, the regulatory framework does not engage with the fairness literature at all; the primary regulatory instrument, the Food and Drug Administration's device approval process, addresses accuracy and safety without specifying fairness criteria in terms the technical literature would recognize.

The third pattern is the absence of domain coordination mechanisms. Each of the three domains has developed its own partial accountability infrastructure, with its own regulatory instruments, its own audit methodologies, and its own vocabulary for describing algorithmic fairness concerns. There is no cross-domain coordination body, no common metric standard, and no mechanism by which the governance lessons learned in one domain are systematically applied to others. This fragmentation has practical consequences: an algorithm developer can move between domains, deploying variants of the same model in criminal justice risk assessment and employment screening, without facing any common accountability standard. The absence of coordination also means that regulators in one

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

domain cannot learn from the metric specification failures of their counterparts in other domains.

The fourth pattern concerns the relationship between metric choice and the distribution of accountability costs. In each domain, the implicit metric choice embedded in the operative regulatory instrument allocates the costs of unavoidable trade-offs to the same group: members of racially defined groups who are already subject to historical disadvantage. Demographic parity requirements, while reducing group-level selection rate disparities, do not address the elevated false negative rates that expose qualified members of protected groups to disparate screening-out. Predictive parity requirements in criminal justice protect the aggregate calibration of risk scores while tolerating the differential false positive rates that expose Black defendants to disproportionate pretrial detention. The choice of fairness metric is, at root, a choice about which group bears the cost of the mathematical trade-off that cannot be avoided; the current governance architecture makes that choice implicitly, without public deliberation, and consistently in the direction that compounds existing disadvantage.

Regulatory Gap Analysis: The Fairness Specification Matrix

The regulatory translation gap is the third component of the FMMF and the most directly actionable finding of this analysis. It refers to the systematic failure of regulatory instruments to specify the mathematical fairness criteria whose satisfaction they require, with the consequence that compliance with the instrument's formal requirements is compatible with the persistence of forms of discrimination the instrument was enacted to prevent. To operationalize this component, the paper constructs a Fairness Specification Matrix mapping the major regulatory instruments across the three domains against the fairness taxonomy developed in the theoretical background.

The European Union Artificial Intelligence Act (Regulation (EU) 2024/1689), which entered into force in August 2024, is the most comprehensive regulatory instrument for artificial intelligence globally. It classifies artificial intelligence systems used in consequential decision-making in criminal justice, employment, and access to services as high-risk under Annex III, and subjects those systems to mandatory requirements including data governance, technical documentation, transparency, human oversight, and accuracy and robustness standards. The Act's non-discrimination provisions require that high-risk AI systems be designed and developed with appropriate safeguards to prevent discriminatory outputs. Article 10 requires that training, validation, and testing datasets be relevant, sufficiently representative, and, to the best extent possible, free of errors. The Act does not specify a fairness metric. It uses the language of accuracy and representativeness without engaging with the impossibility results or specifying which among the competing fairness criteria its accuracy requirements are intended to satisfy. A high-risk AI system can satisfy every provision of the EU AI Act while optimizing predictive parity, while optimizing demographic parity, or while optimizing neither, without any of those choices being assessable as compliant or non-compliant under the Act's text.

The EEOC Technical Assistance of May 18, 2023 is closer to metric specification than the EU AI Act in that it explicitly invokes adverse impact analysis and the four-fifths rule as the operative framework for evaluating employer compliance with Title VII obligations when using algorithmic employment decision tools (EEOC, 2023). This represents a meaningful step toward the specificity the FMMF identifies as necessary. However, the guidance's metric specification is incomplete in two important respects. First, the four-fifths rule is a threshold test for demographic parity at the selection stage; it does not address false negative rates or the equalized odds criterion. An employer whose algorithm satisfies the four-fifths rule may still be systematically screening out qualified candidates from protected groups at higher rates than similarly qualified candidates from majority groups, in violation of an equalized odds standard, without triggering the

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

guidance's compliance concern. Second, the guidance is non-binding technical assistance rather than a regulation; it does not impose enforceable legal obligations on employers and cannot be the basis for administrative action in the absence of a formal rulemaking.

New York City Local Law 144 of 2021 represents the most operationally specific regulatory instrument examined in this analysis. Its requirement for annual independent bias audits with public disclosure creates an accountability mechanism that goes beyond the guidance documents and broad principles that characterize most algorithmic governance frameworks. The law's definition of a bias audit, which requires assessment of disparate impact on protected categories as defined in federal EEO reporting requirements, anchors the operative fairness criterion to the demographic parity framework. This specificity is genuine progress, and the law's audit and disclosure requirements create a transparency infrastructure that more general regulatory instruments lack. The gap in the law's specification is the same as in the EEOC guidance: it does not require auditors to assess equalized odds or to evaluate whether the tool's false negative rates are equally distributed across protected groups. An employer can post a compliant bias audit demonstrating no adverse impact under the four-fifths rule while employing a tool that systematically produces higher false negative rates for Black and Hispanic candidates. The law does not require that this be disclosed or addressed.

Table 1: Fairness specification matrix: major regulatory instruments across domains

Instrument	Domain Covered	Fairness Language	Metric Specified	Achievable Under Impossibility Constraints
EU AI Act (2024/1689)	All high-risk AI systems	Accuracy, robustness, non-discrimination	No metric specified	Underdetermined: compliance compatible with any or no metric
EEOC Technical Assistance (2023)	Employment	Adverse impact / four-fifths rule	Demographic parity (implicit; non-binding)	Conflicts with equalized odds when base rates differ
NYC Local Law 144 (2021)	Employment (NYC)	Disparate impact on protected EEO categories	Demographic parity (explicit; enforceable)	Conflicts with equalized odds when base rates differ
FDA Device Approval	Healthcare diagnostic devices	Accuracy and safety	No fairness metric specified	Underdetermined: demographic performance not required
U.S. Criminal Risk Assessment (jurisdiction-level)	Criminal justice	Accuracy, objectivity, consistency	No fairness metric specified	Underdetermined: COMPAS satisfies calibration but violates equalized odds

Note: "Underdetermined" indicates that the instrument's text does not permit determination of which fairness metric governs compliance. Columns 4 and 5 reflect the authors' assessment based on the instrument's text and the formal conditions established by Chouldechova (2017) and Kleinberg et al. (2016).

The Fairness Specification Matrix constructed from this analysis reveals a consistent pattern across all three instruments. Each uses fairness language, and each imposes some form of accountability obligation. None specifies a fairness metric with sufficient precision to determine which among the competing criteria the instrument is intended to enforce. None engages with the impossibility results or acknowledges that its chosen metric, where one can be inferred, forecloses other fairness criteria under real-world conditions. None establishes a mechanism for resolving the normative question of which group should bear the cost of the trade-off when the chosen metric requires one. This is the regulatory translation gap: the distance between the accountability ambitions expressed in regulatory language and the metric specificity required for those ambitions to be enforced.

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

The gap has a specific structural cause that is distinct from regulatory inadequacy in the conventional sense. It is not primarily that regulators have failed to understand the technical literature or that they have been captured by industry interests, though both have occurred in specific instances. It is that the impossibility results impose a genuinely difficult normative choice on regulators, and the political economy of regulation makes it easier to mandate fairness in general terms than to specify which fairness criterion will govern and thereby own the trade-off that specification entails. A regulation that specifies equalized odds as the operative criterion can be challenged by system developers on the grounds that it precludes predictive parity, and vice versa. A regulation that specifies fairness without a metric forecloses neither challenge and cannot be falsified by any audit finding. The FMMF's contribution is to name this dynamic explicitly and to argue that the costs of metric non-specification, borne disproportionately by historically disadvantaged groups, are higher than the political costs of metric specification that regulators have so far been unwilling to bear.

DISCUSSION

The FMMF analysis generates three principal implications for algorithmic governance reform. Each addresses a different level of the accountability architecture: the metric specification problem, the audit methodology problem, and the cross-domain coordination problem.

The metric specification problem is the most fundamental and the most tractable in the short term. Regulators who have enacted or issued algorithmic accountability instruments should be required to specify, in the text of those instruments or in binding implementation guidance, the fairness criterion or criteria whose satisfaction constitutes compliance. This specification should be accompanied by an explicit acknowledgment of the trade-offs the chosen criterion forecloses, so that the normative choices embedded in metric selection are subject to public deliberation rather than being made implicitly in technical documents that are inaccessible to the communities most affected by algorithmic decisions. The EU AI Act's review mechanism, which permits updating the list of high-risk AI systems and their requirements, provides a procedural vehicle for introducing metric specification without requiring primary legislation. The EEOC's regulatory authority under Title VII provides a basis for converting the 2023 technical assistance into a binding rule with metric-specific requirements. NYC Local Law 144's annual audit requirement could be amended to require auditors to report on equalized odds in addition to disparate impact under the four-fifths rule.

The audit methodology problem is the most technically complex. Current audit methodologies, including those mandated by NYC Local Law 144 and recommended by the EEOC guidance, do not standardize the fairness metrics against which systems are evaluated, do not require auditors to report the trade-offs between metrics, and do not provide a common framework that would allow audit findings from one system to be compared to those from another. An adequate audit methodology for the post-FMMF regulatory landscape would require auditors to evaluate systems against multiple fairness criteria simultaneously, to report the metric values for each criterion and the relationships among them, and to identify explicitly which criteria the system satisfies and which it cannot satisfy given its design and the base rates present in its deployment population. This would not resolve the normative question of which criterion should govern; it would, however, make that normative question visible and provide the information regulators and affected communities would need to make an informed choice.

The cross-domain coordination problem requires institutional innovation. No existing body has the mandate, the technical capacity, and the jurisdictional reach to develop and enforce common fairness metric standards across criminal justice, healthcare, and employment simultaneously. In the United States, algorithmic governance is fragmented across the EEOC, the

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

Food and Drug Administration, the Federal Trade Commission, the Department of Justice, and numerous state and municipal bodies. In the European Union, the AI Act's enforcement structure allocates primary responsibility to national market surveillance authorities, with coordination through the European AI Board. Neither architecture is designed to address the cross-domain coordination problem the FMMF identifies. Addressing it would require either the establishment of a cross-domain algorithmic accountability body with binding standards authority or a formal coordination mechanism among existing regulators with the specific mandate of aligning fairness metric standards across deployment domains.

Several limitations of this analysis should be acknowledged. The three domains examined do not exhaust the high-stakes deployment contexts where algorithmic bias has been documented; credit scoring, immigration enforcement, and child welfare risk assessment each present related but distinct accountability challenges that the FMMF has not been applied to here. The regulatory instruments analyzed represent a selection of major frameworks rather than a comprehensive global inventory; significant algorithmic governance activity is occurring in China, Canada, Brazil, and other jurisdictions not examined here. The FMMF's regulatory translation gap analysis is based on the text of regulatory instruments rather than on empirical evidence about how those instruments are being implemented or enforced; implementation gaps may compound the specification gaps documented here. Future research should apply the FMMF to additional domains and jurisdictions, examine the relationship between metric specification and enforcement outcomes, and investigate whether the normative trade-offs the impossibility results impose can be subjected to structured democratic deliberation.

CONCLUSION

This paper has advanced the Fairness Metric Misalignment Framework as an analytical instrument for diagnosing why algorithmic accountability failures persist in criminal justice, healthcare, and employment despite growing regulatory attention and an extensive critical literature. The core argument is that the persistence of discriminatory algorithmic outcomes is not primarily explained by technical failures, insufficient regulation, or individual bad faith, but by a structural incoherence in how fairness is operationalized across the full accountability chain from system design through audit to regulatory compliance.

The FMMF identifies three interlocking components of this incoherence. The metric taxonomy establishes that competing fairness definitions encode incompatible normative assumptions and that mathematical impossibility results preclude their simultaneous satisfaction under realistic deployment conditions. The domain-metric mapping demonstrates that the fairness criteria invoked in audit studies, operative in system design, and mandated by regulatory instruments are systematically misaligned across all three domains examined, generating accountability gaps that protect discriminatory systems from any standard by which they could be found non-compliant. The regulatory translation gap analysis shows that major instruments including the EU AI Act, the EEOC Technical Assistance, and NYC Local Law 144 deploy fairness language without the metric specificity required for enforcement, and that this specificity gap is not accidental but reflects the political economy of regulation in the face of genuinely difficult normative trade-offs.

The governance implications are direct. Adequate algorithmic accountability requires regulators to specify fairness metrics explicitly, to acknowledge the trade-offs those choices entail, to update audit methodologies to require multi-metric reporting, and to develop cross-domain coordination mechanisms that prevent governance fragmentation from sheltering discriminatory systems from scrutiny. These are politically demanding requirements precisely because they force into public view the normative choices about who bears the cost of unavoidable trade-offs that current governance frameworks make

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org

implicitly and consistently at the expense of historically disadvantaged communities. The contribution of this paper is to name that pattern and to provide a structured framework for the governance reforms that addressing it requires.

REFERENCES

Ahmad, A., Vallès, Y., & Idaghdour, Y. (2026). Bias in AI systems: Integrating formal and socio-technical approaches. *Frontiers in Big Data*, *8*. https://doi.org/10.3389/fdata.2025.1686452

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Dastin, J. (2018, October 11). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc. https://perma.cc/RVC7-SE8R

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

EPIC. (2019). *Complaint and Request for Investigation, Injunction, and Other Relief*. The Electronic Privacy Information Center (EPIC). https://epic.org/wp-content/uploads/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf

Equal Employment Opportunity Commission. (2023, May 18). *Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures*. U.S. Equal Employment Opportunity Commission. https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. European Parliament and Council of the European Union. http://data.europa.eu/eli/reg/2024/1689/oj

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html

Harwell, D. (2019, November 6). Rights group files federal complaint against AI-hiring firm HireVue, citing 'unfair and deceptive' practices. *The Washington Post*. https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores* (arXiv:1609.05807). arXiv. https://doi.org/10.48550/arXiv.1609.05807

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Office of the New York State Comptroller. (2025, December 2). *Enforcement of Local Law 144 – Automated Employment Decision Tools*. New York State Office of the Comptroller. https://www.osc.ny.gov/state-agencies/audits/2025/12/02/enforcement-local-law-144-automated-employment-decision-tools

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

© 2026 The Authors. Published by **Sentam Research and Publishing** | www.sentam.org