EXPLAINABLE ARTIFICIAL INTELLIGENCE AND ALGORITHMIC FAIRNESS: ADDRESSING HUMAN RIGHTS CHALLENGES IN DATA SCIENCE APPLICATIONS

Abdiyeva Dilfuza Ikhtiyorovna
National University of Uzbekistan, Faculty of Social Sciences, Department of Jurisprudence
abdiyevakookowa@gmail.com | Tel: +998900706550

Abstract: *Automated decision-making systems now govern access to employment, credit, healthcare, and criminal justice across dozens of jurisdictions. This article advances the thesis that algorithmic opacity and algorithmic discrimination are not two separate problems but expressions of a single structural failure: the Compounding Opacity Problem (COP), in which each successive layer of technical and legal accountability generates its own epistemic barrier, rendering oversight formally possible but substantively meaningless. We demonstrate that post-hoc explainability methods are legally inadequate, that fairness metrics are mathematically incompatible with one another, and that current regulatory frameworks lack the technical specificity to detect either failure. We further advance an original jurisprudential argument: the selection of a fairness metric is not a technical act but a normative one with the distributional consequences of primary legislation, and its current delegation to developers without statutory constraint constitutes an unconstitutional non-delegation of legislative authority. Against this diagnosis, we propose Pre-Deployment Adversarial Auditing (PDAA), an integrated framework combining adversarial pipeline probing, subgroup stress-testing under distributional shift, counterfactual recourse certification, and continuous post-deployment monitoring, supported by National Algorithmic Audit Authorities with independent technical access powers.*

Keywords: *explainable AI, algorithmic fairness, jurisprudence, human rights, PDAA, GDPR, AI Act, non-delegation doctrine, counterfactual recourse, distributional shift.*

INTRODUCTION

On 23 May 2023, Italy's data protection authority, the Garante, suspended a predictive policing algorithm after finding it had concentrated patrol resources in immigrant neighbourhoods at rates irreconcilable with crime data. The investigation collapsed not at the legal analysis stage but at the prior obstacle: the vendor invoked trade secrecy, leaving no technically adequate basis to establish what the model was doing [1]. This pattern recurs across criminal justice [2], healthcare [3], credit allocation [4], and hiring [5]: accountability mechanisms fail not because law is unenforced but because the technical preconditions for enforcement are absent. Existing scholarship treats opacity and bias as separate problems, addressed by explainability requirements and fairness standards respectively. This framing is itself the source of the field's persistent inadequacy. The Compounding Opacity Problem (COP) thesis holds that each layer of oversight introduces its own epistemic barrier: post-hoc explanation methods can be strategically fooled [13]; fairness metrics are mathematically incompatible with one another [16,17]; and legal disclosure obligations generate outputs that cannot be evaluated without technical tools

that are themselves compromised. The result is a recursive structure in which compliance becomes formally possible and substantively meaningless. This article diagnoses that structure in data science terms (Section 2), draws out its jurisprudential implications including an original constitutional argument (Section 3), surveys the regulatory landscape (Section 4), proposes the PDAA framework (Section 5), addresses objections (Section 6), and concludes with binding recommendations (Section 7).

The Compounding Opacity Problem: Data Science Dimensions

Data Science Foundations of Algorithmic Accountability. Supervised machine learning models learn to map input features to output predictions by optimising parameters against a labelled training dataset. Accountability failures do not appear suddenly in deployed systems; they accumulate across every technical decision from data collection to model evaluation. The training dataset is the foundational layer. When historical records reflect past discriminatory practices, as is routine in criminal justice, credit, and hiring datasets, the model learns to replicate the distributional patterns of a biased history rather than predict an objective outcome. This is what the literature terms historical bias, and it is structurally invisible to any evaluation metric computed on data drawn from the same distribution. Feature engineering compounds this problem. Composite features, normalisation transformations, and categorical encodings can concentrate the predictive signal carried by protected attributes across multiple variables simultaneously, so that no single engineering step appears discriminatory in isolation while the aggregate effect is substantial. Standard model evaluation uses aggregate metrics computed on a held-out test set: accuracy, AUC, F1 score. These are insensitive to subgroup performance disparities. A model achieving 85% overall accuracy while producing false positive rates of 40% for one demographic group and 15% for another passes standard evaluation without triggering any alert. Disaggregated evaluation reporting metrics separately for each protected subgroup is technically straightforward but required by no binding regulatory standard. Bias propagation ensures that errors introduced at any upstream stage amplify through subsequent pipeline stages and, in long-running deployments, compress historical discrimination into an increasingly dense mathematical structure. The output of this process, a numerical risk score governing bail, credit, or healthcare eligibility, carries an appearance of precision that the underlying methodology does not support. Closing the gap between technical reality and institutional perception is the core justification for legally mandated data science accountability standards.

Machine learning models do not process social reality directly; they process engineered representations of it. Feature selection and transformation are the operations through which historical inequalities become mathematical structure. A credit model trained on repayment history, account age, and credit utilisation encodes no protected attribute explicitly, yet audit studies demonstrate that it produces systematically different score distributions across racial groups that cannot be explained by genuine differences in creditworthiness [6,7]. The mechanism is historical feedback: data drawn from the redlining era encodes the predictive relationship between postcodes and default without encoding that the causal path runs through discriminatory exclusion. Suresh and Guttag's taxonomy identifies three analytically distinct bias types that can coexist in a single system:

historical bias (unjust training data), measurement bias (discriminated proxy labels), and aggregation bias (one model applied across heterogeneous populations) [8]. The COMPAS recidivism instrument exhibits all three simultaneously [2,9], and this cumulative structure is the norm rather than the exception in high-stakes deployment. LIME [10] and SHAP [11] generate feature attribution scores estimating each variable's contribution to a specific prediction. Three empirical findings establish their legal inadequacy. First, LIME explanations are locally unstable: minor input perturbations produce qualitatively different attributions for unchanged predictions [12], incompatible with the procedural consistency rule-of-law reasoning demands. Second, and more critically, Slack et al. (2020) proved that a model can be engineered to produce compliant explanations on the audit distribution while discriminating on the deployment distribution [13]. The attack requires a single conditional branch in the prediction pipeline. An institution with incentives to conceal discrimination can do so systematically while satisfying every current explanation requirement. Third, Wachter et al. (2017) established that counterfactual explanations of the form 'the decision would have differed had X taken value Y' are both the legally most relevant explanation type and structurally outside what feature attribution methods produce [14]. Feature attribution is not actionable recourse. Chouldechova (2017) and Kleinberg et al. (2017) proved independently that when outcome base rates differ across demographic groups, no classifier can simultaneously satisfy equalised false positive rates, equalised false negative rates, and calibration [16,17]. Every deployed high-stakes model is therefore, by mathematical necessity, unfair according to at least one formally defensible fairness criterion. Choosing which criterion to satisfy entails choosing whose errors matter more, a normative determination currently made implicitly by data scientists selecting evaluation metrics without democratic deliberation or legal constraint. Hardt et al. (2016) showed that post-processing for equalised odds paradoxically harms intended beneficiary groups by reducing prediction quality for them [18]. Distributional shift compounds this: Obermeyer et al. (2019) documented a healthcare algorithm that used past expenditure as a proxy for medical need, systematically underestimating need in Black patients because historical discrimination had produced lower expenditure for equivalent conditions [3]. No post-hoc explanation method would have revealed this failure, because it resided in the labelling choice, not in the model's subsequent reasoning.

Jurisprudential Dimensions: Three Original Arguments

Administrative law across constitutional systems prohibits the transfer of legislative power to private actors without an intelligible statutory principle constraining its exercise. The non-delegation doctrine in US constitutional law, German Basic Law Article 80, and EU proportionality jurisprudence all reflect the principle that decisions with fundamental distributional consequences for citizens require democratic legitimation. When a developer selects demographic parity over equalised odds as the governing fairness criterion, they are making a choice with the distributional consequences of primary legislation: determining, for the affected population, which group's errors are tolerable and which are not. This choice is currently made in model evaluation spreadsheets without statutory constraint, without public justification, and without the possibility of legal challenge. Article 22 GDPR [20] and the EU AI Act [21] neither specify which fairness criterion governs high-risk

decisions nor require developers to justify their choice. The constitutional defect is not in the absence of AI regulation generally but in the systematic absence of democratic constraint over the most consequential normative choice in system design. The right to an effective remedy under Article 8 UDHR and Article 13 ECHR [25,26] requires not merely formal access to a tribunal but a realistic prospect of vindication. The COP creates an informational asymmetry that constitutes an independent rights violation: a complainant cannot establish discrimination without access to training data held by the deploying institution; that institution can satisfy disclosure obligations with strategically manipulated explanations; and reviewing tribunals lack the technical capacity to evaluate outputs even if provided. This is the legal equivalent of a right of appeal in a language the court does not speak. The right to an effective remedy is structurally unavailable for the majority of individuals harmed by algorithmic decisions under current frameworks, and this unavailability is not a procedural inconvenience but a substantive rights failure. The GDPR's requirement of 'meaningful information about the logic involved' and the AI Act's requirement of 'sufficient transparency' are legally indeterminate without reference to data science methodology. Deriving the content of a legally adequate transparency obligation from administrative law's duty to give reasons, we argue it must specify at minimum: training data sources and preprocessing operations; the fairness criterion selected and alternatives considered; counterfactual recourse pathways for the affected individual; and distributional shift monitoring results disaggregated by protected group. No binding instrument currently requires any of these four elements. The gap is not incidental; it reflects the failure of legal frameworks to engage with the technical architecture of the systems they purport to regulate.

The EU AI Act establishes a risk-tiered framework subjecting high-risk AI applications in criminal justice, employment, credit, healthcare, and public administration to conformity assessment, transparency obligations, and human oversight requirements [21]. These represent genuine advances. The Act's critical weakness is its reliance on provider self-assessment for most high-risk applications. In pharmaceuticals, aviation, and financial services, self-assessment has systematically produced underdisclosure of risks; there is no principled basis for expecting different behaviour from AI developers, particularly given the strategic deception dynamics established by Slack et al. [13]. Technical standards are delegated to harmonised bodies operating on timelines unsynchronised with the Act's application schedule, creating a period of regulatory operation without enforceable technical specification during which compliance is formally possible and substantively unverifiable. In the United States, Title VII, the Equal Credit Opportunity Act, and the Fair Housing Act prohibit disparate impact, and enforcement agencies have issued guidance applying them to algorithmic systems [27], but disparate impact claims require plaintiffs to first quantify the disparity without access to the model's inputs or outputs. The 2023 Executive Order on AI [28] created interagency coordination without a private right of action or mandatory third-party auditing. The protection these frameworks offer is real in principle and largely inaccessible in practice.

Pre-Deployment Adversarial Auditing: An Original Framework

The COP cannot be resolved by improving any single oversight element in isolation. Stronger explanations do not address strategic deception or distributional shift. Fairer training procedures do not dissolve the impossibility of simultaneous metric satisfaction. More detailed disclosure obligations do not help if what is disclosed cannot be meaningfully evaluated. An adequate response must be architecturally integrated and adversarially designed, assuming that providers will exploit any compliance pathway that permits gaming, and closing those pathways before they are opened.

Component 1: Adversarial Pipeline Probing. Independent auditors with controlled access to the full training pipeline, including raw data, preprocessing code, and feature engineering logic, conduct systematic probing to identify proxy discrimination pathways. Rather than reviewing documentation, auditors test whether substituting or ablating features correlated with protected attributes changes subgroup predictions, drawing on concept erasure techniques from mechanistic interpretability research [29]. Audits are designed on the adversarial assumption that the pipeline may have been structured to pass standard documentation review.

Component 2: Subgroup Stress-Testing Under Distributional Shift. Fairness metrics are evaluated across a structured battery of distributional perturbations varying demographic composition, historical period, and geographic region of deployment. False positive rates, false negative rates, and calibration are reported separately for each protected subgroup under each scenario. A model that performs adequately in aggregate but degrades sharply for a protected subgroup under plausible deployment conditions is flagged as high-risk before deployment, directly addressing the Obermeyer et al. finding [3].

Component 3: Counterfactual Recourse Certification. Deployers must demonstrate that a feasible modification to an individual's input features would change an adverse prediction to a favourable one, and that the required modification is not systematically more demanding for protected groups. This operationalises the right to explanation as actionable recourse, drawing on Wachter et al. [14] and subsequent algorithmic recourse literature [30]. A system that cannot provide equitable recourse pathways across demographic groups fails certification.

Component 4: Continuous Post-Deployment Monitoring. Deployers implement real-time monitoring of prediction distributions disaggregated by protected group, with automated alerts when disparity metrics exceed pre-specified thresholds. Results are reported to the relevant authority quarterly and published publicly annually, creating an external accountability mechanism independent of individual complaints.

Component 5: Mandatory Metric Justification. Addressing the constitutional argument in Section 3.1, deployers must publicly disclose the fairness criterion selected, the alternatives considered, the trade-offs involved, and the normative justification expressed in rights-compatible terms. This transfers the choice from implicit technical default to explicit public record, where it is subject to democratic scrutiny, legal challenge, and regulatory review.

PDAA requires National Algorithmic Audit Authorities (NAAAs) with three features distinguishing them from existing data protection bodies: multidisciplinary technical capacity spanning data science, statistics, and law; independent access powers permitting

compelled inspection of training pipelines and data without developer consent, analogous to pharmaceutical and financial regulatory powers; and self-transparency obligations requiring publication of audit methodologies and significant decisions. The legal framework must include a pre-deployment clearance requirement for all high-risk applications, a private right of action enabling access to PDAA audit results in anti-discrimination proceedings with legal aid provision, and administrative penalties set as a percentage of global turnover to serve deterrence rather than mere compliance.

The performance objection holds that adversarial auditing will deter beneficial AI development. The premise is overstated: Rudin (2019) demonstrated that inherently interpretable models achieve comparable accuracy to black-box alternatives in criminal justice and healthcare prediction [33], suggesting the performance cost of transparency requirements is frequently negligible. More fundamentally, compliance costs do not veto rights-protective requirements. The EU imposes substantial costs on pharmaceutical developers to protect patients; there is no principled basis for treating algorithmic systems differently when stakes are comparably high. The measurement objection holds that auditing is futile given the impossibility results. We reject the inference. PDAA does not require a single metric to satisfy all criteria simultaneously; it requires transparent public justification of the metric selected and its implications for those it disadvantages. The impossibility results establish that every choice has distributional costs; they do not establish that the choice may be made without accountability. The jurisdictional objection holds that unilateral PDAA requirements will push development to less regulated environments. The EU AI Act's extraterritorial provisions and the demonstrated Brussels effect address this concern: large-market jurisdictions have repeatedly succeeded in exporting regulatory standards through market access conditions. The objection is in any case an argument for international coordination, not regulatory abstention.

Conclusion

This article has advanced three original contributions. First, the Compounding Opacity Problem thesis recharacterises the field's central challenge as a unified structural failure rather than two discrete technical and legal deficits, reorienting the theoretical basis of algorithmic accountability analysis. Second, the non-delegation argument establishes a constitutional dimension of the fairness metric problem that has not been articulated in the literature, opening a new avenue of challenge to algorithmic systems operating without statutory constraint over their fairness criteria. Third, PDAA provides the first integrated framework to address all dimensions of the COP simultaneously, combining adversarial methodology, distributional stress-testing, counterfactual certification, and continuous monitoring within a legally specified institutional architecture.

The following binding recommendations follow from this analysis: amend the EU AI Act's conformity assessment to require independent PDAA auditing for all high-risk applications without exception; develop binding technical specifications for each PDAA component with mandatory three-year review cycles; establish NAAAs with the technical capacity and access powers specified above; create a private right of action enabling affected individuals to compel disclosure of PDAA results in anti-discrimination proceedings; and enact statutory provisions specifying the permissible class of fairness

criteria for each high-risk domain, removing this normative choice from the domain of unaccountable technical discretion.

The aggregate effect of algorithmic systems on citizens' material and legal circumstances is more consequential than most primary legislation, yet these systems are currently governed with less democratic constraint than a municipal by-law.

The choice to continue tolerating that condition is itself a governance choice, one that those who have examined the evidence can no longer make in good faith.

REFERENCES:

1. Garante per la Protezione dei Dati Personali. (2023). Decision on Predictive Policing System PredPol, Provision of 23 May 2023. Italian Data Protection Authority.
2. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica.
3. 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.
4. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. Journal of Finance, 77(1), 5-47.
5. Sanchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to solve the problem of discrimination in hiring? Proceedings of the 2020 ACM FAccT Conference, 458-468.
6. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30-56.
7. Chen, E., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory? Advances in Neural Information Processing Systems, 31.
8. Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. ACM EAAMO, 1-9.
9. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.
10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD, 1135-1144.
11. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30, 4765-4774.
12. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. ICML Workshop on Human Interpretability in Machine Learning.
13. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. AAAI/ACM AIES, 180-186.
14. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.
15. Verma, S., & Rubin, J. (2018). Fairness definitions explained. IEEE/ACM FairWare Workshop, 1-7.

16. Chouldechova, A. (2017). Fair prediction with disparate impact. Big Data, 5(2), 153-163.

17. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. ITCS, 43, 1-23.

18. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. NeurIPS, 29, 3315-3323.

19. Zhang, J., & Neill, D. B. (2021). Identifying significant predictive bias in classifiers. ACM FAccT, 340-350.

20. European Parliament and Council. (2016). Regulation (EU) 2016/679 (GDPR). Official Journal of the European Union, L 119, 1-88.

21. European Parliament and Council. (2024). Regulation (EU) 2024/1689 (AI Act). Official Journal of the European Union.

22. Council of Europe. (2020). Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems.

23. United Nations. (1966). International Covenant on Civil and Political Rights, Art. 26. Treaty Series, 999, 171.

24. United Nations. (1965). International Convention on the Elimination of All Forms of Racial Discrimination. Treaty Series, 660, 195.

25. United Nations. (1948). Universal Declaration of Human Rights, Art. 8. UN General Assembly Resolution 217 A.

26. Council of Europe. (1950). European Convention on Human Rights, Art. 13. ETS No. 005.

27. Equal Employment Opportunity Commission. (2023). Guidance on AI and Algorithmic Decision-Making Tools Under Title VII. EEOC.

28. Executive Office of the President. (2023). Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of AI. Federal Register, 88, 75191.

29. Ravfogel, S. et al. (2020). Null it out: Iterative nullspace projection. ACL, 7237-7256.

30. Karimi, A. H. et al. (2021). Algorithmic recourse: From counterfactual explanations to interventions. ACM FAccT, 353-362.

31. Raji, I. D. et al. (2020). Closing the AI accountability gap. ACM FAccT, 33-44.

32. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

33. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions. Nature Machine Intelligence, 1(5), 206-215.