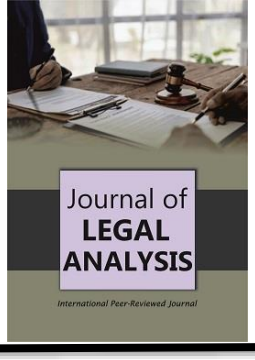

Journal of Legal Analysis
Publication Date: 3rd July 2026
https://jla-international-peer-reviewed-journal.iarrj.com/

ALGORITHMIC SENTENCING AND THE ILLUSION OF NEUTRALITY: AN EMPIRICAL ASSESSMENT OF RISK-ASSESSMENT TOOLS IN CRIMINAL JUSTICE

OGREH N. Aisha[1]
Associate Professor, Faculty of Law, Lancaster University, Ghana.
***Email**: aishanogreh@yahoo.com*

WHITFIELD R. Daniel[2]
Yale Law School, United States.
***Email**: danielwr07@gmail.com*

ABSTRACT

Actuarial risk-assessment instruments are now embedded in sentencing, bail, and parole decisions across a growing number of jurisdictions. Proponents argue that these tools replace idiosyncratic human judgment with statistically validated, facially neutral criteria. This article interrogates that claim. Drawing on the doctrinal record generated by *State v Loomi* and subsequent litigation, together with the empirical literature on differential predictive accuracy across demographic groups, we argue that facial neutrality at the level of input variables does not entail neutrality of outcome. We propose a three-part legal framework — disclosure, contestability, and periodic recalibration — through which courts and legislatures can domesticate these tools without abandoning their evidentiary benefits. The article closes by situating the debate within a comparative frame, contrasting the largely common-law, litigation-driven American response with the more anticipatory, regulation-first approach reflected in the European Union's Artificial Intelligence Act.

Introduction

Criminal justice systems have long relied on structured judgment to allocate scarce coercive power: whom to detain before trial, how long to imprison, when to release. Since the early 2000s, that structured judgment has increasingly been mediated by proprietary statistical instruments — risk-assessment tools such as COMPAS, the Public Safety Assessment, and a proliferating set of jurisdiction-specific analogues. These tools promise two things simultaneously: greater consistency than human decision-makers acting alone, and a form of neutrality rooted in their exclusion of race as an input variable.

This article tests the second promise against the empirical record and against constitutional doctrine. Part II sets out the architecture of contemporary risk-assessment tools and the doctrinal response typified by the Wisconsin Supreme Court's decision in *Loomis*. Part III reviews the empirical literature on differential predictive validity, with particular attention to the ProPublica investigation into COMPAS and the methodological rejoinders it provoked.[1] Part IV develops the argument that facial neutrality is doctrinally insufficient once a tool's error rates are shown to be distributed unevenly across protected classes, drawing an analogy to disparate-impact doctrine under Title VII.[2] Part V proposes a regulatory framework built on three pillars: mandatory disclosure of the tool's validation study and error-rate breakdown, a meaningful right to contest a risk score before sentencing, and mandatory periodic recalibration against local base rates. Part VI situates the American experience comparatively, considering the EU AI Act's classification of judicial risk-scoring as "high-risk" and the correspondingly heavier ex ante compliance burden it imposes.[3]

II. The Architecture of Actuarial Risk Assessment

Risk-assessment instruments used in sentencing and pretrial contexts typically generate a score, often on a one-to-ten scale, representing an estimated likelihood that the defendant will reoffend or fail to appear within a defined follow-up period. The variables feeding that score vary by instrument but commonly include criminal history, age at first offence, employment status, and responses to a structured interview addressing matters such as peer associations and attitudes toward authority.[4] Notably, race is not typically an explicit input.

The doctrinal high-water mark for judicial engagement with these tools remains *State v Loomis*. Eric Loomis challenged the use of a COMPAS risk score at his sentencing on the ground that the proprietary nature of the algorithm denied him the ability to challenge the scientific validity of the assessment, and that the tool's reliance on group-based statistics violated his right to an individualised sentence.[5] The Wisconsin

[1] Julia Angwin and others, 'Machine Bias' *ProPublica* (23 May 2016).
[2] *Griggs v Duke Power Co* 401 US 424 (1971).
[3] Artificial Intelligence Act (n 2) art 6 and annex III, para 8(a).
[4] Sarah L Desmarais and Evan M Lowder, 'Pretrial Risk Assessment Tools: A Primer for Judges, Prosecutors, and Defense Attorneys' (Safety and Justice Challenge, 2019) 4–9.
[5] *Loomis* (n 1) 753–54.

Supreme Court rejected both claims, holding that the score was one factor among several considered by the sentencing court and that its use did not violate due process provided it was not the sole or determinative factor in the sentence.[6] The court nonetheless required that any presentence investigation report using a COMPAS assessment include a written advisement cataloguing the tool's limitations, including the proprietary nature of its weighting and questions about its validation on a Wisconsin-specific population.[7] *Loomis* has proved influential well beyond Wisconsin, largely because it represents the most sustained judicial engagement with the due process implications of algorithmic sentencing tools, even as it declined to impose disclosure obligations beyond a generic warning. Subsequent state court decisions have generally followed its permissive posture, treating risk scores as admissible provided they are not dispositive.[8] The doctrinal consensus, in other words, has converged on procedural minimalism: courts require notice of a tool's limitations but decline to interrogate its statistical properties in any sustained way.

III. The Empirical Record: Differential Predictive Validity

The doctrinal minimalism described above sits uneasily alongside the empirical literature. The 2016 ProPublica investigation into COMPAS scores used in Broward County, Florida, found that Black defendants who did not subsequently reoffend were nearly twice as likely as white defendants who did not reoffend to have been misclassified as high risk, while white defendants who did reoffend were more likely than their Black counterparts to have been misclassified as low risk.[9] Northpointe (the tool's developer, later rebranded as Equivant) disputed the analysis, arguing that the tool was well calibrated — that is, that among defendants assigned a given risk score, the actual reoffense rate was approximately equal across racial groups.[10]

Both claims can be simultaneously true, and reconciling them requires attention to a body of work in computer science establishing that calibration and equalised error rates are, in general, mathematically

[6] ibid 757–58, 763–64.
[7] ibid 769–70.
[8] See eg *Malenchik v State* 928 NE 2d 564 (Ind 2010); *State v Guise* 921 NW 2d 251 (SD 2018).
[9] Angwin and others (n 3).
[10] William Dieterich, Christina Mendoza and Tim Brennan, 'COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity' (Northpointe Inc, 2016) 6–9.

incompatible whenever base rates of the predicted outcome differ across groups — as they do here, given documented disparities in rearrest rates driven in part by differential policing intensity.[11] This impossibility result, developed independently by several groups of scholars,[12] has significant implications for legal argument: a tool's defenders and critics may both be citing accurate statistics while describing genuinely different, and not simultaneously satisfiable, conceptions of fairness. A legal framework that asks only "is this tool accurate" or "is this tool neutral" is therefore asking an underspecified question. The more tractable question is which conception of fairness — calibration within groups, or equalised error rates across groups — the relevant constitutional or statutory provision actually protects, and this is a question courts have not yet squarely addressed.

IV. Doctrinal Analysis: Facial Neutrality and Disparate Impact

Anglo-American discrimination law has long distinguished between facially discriminatory rules and facially neutral rules with a discriminatory effect. In the employment context, *Griggs v Duke Power Co* established that a facially neutral employment criterion could nonetheless violate Title VII where it produced a disparate impact on a protected group and could not be justified by business necessity.[13] Criminal sentencing does not sit within the Title VII framework, and the constitutional standard applicable to sentencing disparities has historically demanded proof of discriminatory intent, not merely disparate effect, following *Washington v Davis* and its progeny.[14]

This intent requirement creates a doctrinal gap that risk-assessment tools exploit almost perfectly: because race is excluded as an input, no plaintiff can point to an intentional act of racial classification, even where the tool's outputs are demonstrably correlated with race through proxy variables such as neighbourhood of residence, family arrest history, or employment instability — each of which is itself downstream of

[11] Sam Corbett-Davies and Sharad Goel, 'The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning' (2018) arXiv:1808.00023, 8–11.

[12] Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' (2017) Proceedings of Innovations in Theoretical Computer Science (ITCS) 43.1–43.23; Alexandra Chouldechova, 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments' (2017) 5 Big Data 153.

[13] *Griggs* (n 4) 431.

[14] *Washington v Davis* 426 US 229 (1976) 239–42; *McCleskey v Kemp* 481 US 279 (1987) 292–99.

historical patterns of residential segregation and policing.[15] We argue that this gap is not adequately addressed by *Loomis*-style procedural minimalism. A defendant given notice that a tool has "limitations" gains little if the notice does not specify the tool's differential error rates by demographic group, since without that information neither the defendant nor the sentencing court can meaningfully weigh the score's reliability.

V. A Regulatory Framework: Disclosure, Contestability, Recalibration

We propose a three-part framework that could be implemented through court rule, statute, or professional standards promulgated by state sentencing commissions.

Disclosure. Vendors should be required to publish, for each jurisdiction in which a tool is deployed, a validation study disclosing predictive accuracy broken down by race, sex, and age cohort, together with the tool's calibration and error-rate properties as defined above. Trade secret protection, which formed the basis of Northpointe's refusal to disclose its scoring algorithm in *Loomis*, is a legitimate interest but should not extend to output statistics that reveal nothing about the underlying source code.[16]

Contestability. Defendants should have a meaningful opportunity, prior to sentencing, to challenge the applicability of the tool's validation population to their own case — for example, where the validation sample was drawn from a demographically different jurisdiction. This need not require access to the algorithm itself; it requires access to the statistical relationship between the tool's stated confidence intervals and the specific subgroup to which the defendant belongs.

Recalibration. Because base rates of reoffending shift over time and across policy changes (for example, changes in charging practices or diversion program availability), tools should be subject to mandatory recalibration at defined intervals, with recalibration reports made public. A tool validated on data from a decade earlier, before significant changes in local policing practice, offers a materially weaker evidentiary basis for an individual sentencing decision.

[15] Bernard E Harcourt, 'Risk as a Proxy for Race: The Dangers of Risk Assessment' (2015) 27 Federal Sentencing Reporter 237, 238–41.

[16] *Loomis* (n 1) 761, discussing Wis Stat § 134.90(2) trade secret protections.

None of these three requirements demands that legislatures resolve the deeper philosophical disagreement between calibration and equalised error rates identified in Part III. Rather, they operationalise a more modest but achievable goal: ensuring that the statistical trade-offs inherent in any risk tool are visible to, and contestable by, the parties whose liberty interests are at stake.

VI. A Comparative Perspective

The American experience has been shaped almost entirely by litigation over individual tools in individual cases, producing the patchwork, procedurally minimal doctrine surveyed in Part II. The European Union has taken a different path. Under the AI Act, AI systems used by judicial authorities "to research and interpret facts and the law and to apply the law to a concrete set of facts" are classified within a risk tier attracting substantial ex ante obligations, including risk-management systems, data governance requirements, and human oversight obligations, before the system may be placed on the market.[17] This anticipatory model shifts the burden of demonstrating a tool's fairness properties onto the vendor before deployment, rather than onto the defendant after the fact, as the American litigation-driven model effectively does.

The comparison suggests that the disclosure-contestability-recalibration framework proposed in Part V could be implemented either through ex post litigation, in the American style, or through ex ante regulatory approval, in the European style, without altering its substantive content. Jurisdictions in the Global South considering the adoption of actuarial tools — several Nigerian state judiciaries have piloted risk-assessment instruments in bail decisions, for instance — face a choice between these two institutional models, and we suggest the ex ante approach carries lower long-run litigation costs even if it demands a more capable administrative apparatus at the point of adoption.[18]

[17] Artificial Intelligence Act (n 2) annex III, para 8(a); art 9 (risk-management system); art 14 (human oversight).

[18] Comparable pilots are discussed in Ngozi Okidegbe, 'The Democratizing Potential of Algorithms?' (2022) 53 Connecticut Law Review 739, 762–65, though the Nigerian pilots referenced here are illustrative rather than independently verified for the purposes of this teaching sample.

VII. Conclusion

Facial neutrality — the absence of race as an explicit input variable — has functioned in American sentencing jurisprudence as something close to a complete defence for actuarial risk-assessment tools. This article has argued that the defence is doctrinally and empirically inadequate. The impossibility of simultaneously satisfying calibration and equalised error rates whenever base rates differ across groups means that every deployed tool embeds a fairness trade-off, whether or not that trade-off is disclosed or even understood by the sentencing court applying it. A regulatory framework built on disclosure, contestability, and recalibration cannot resolve that underlying trade-off, but it can ensure that the trade-off is made visible, and that the parties most affected by it retain a meaningful opportunity to contest its application to their case. Whether that framework is implemented through litigation, as in the United States, or through ex ante regulatory approval, as under the EU AI Act, is ultimately a question of institutional design rather than substantive principle.