INTERNATIONAL MULTI DISCIPLINARY JOURNAL FOR RESEARCH & DEVELOPMENT
eISSN 2394-6334 Impact factor: 8.854 Volume 13, issue 04 (2026)

THE BLACK BOX ON THE BENCH: CONSTITUTIONAL FRICTION BETWEEN ALGORITHMIC CERTAINTY AND DUE PROCESS IN CRIMINAL SENTENCING

Author: **Kurbanaliyev Sardor Alidjanovich**
Institution: Tashkent International University
Field of Study: Jurisprudence (Law), 2nd-year student
Phone: +99890-451-53-05 Email: sardorkurbon@gmail.com

Abstract. This article examines the constitutional viability of algorithmic risk assessment tools in criminal sentencing, contrasting the US judiciary's reliance on proprietary "trade secrets" with the European Union's newly enforced transparency mandates under the AI Act (2024/2026). By analyzing the post-Loomis legal landscape and the 2025 surge in "Generative Pre-Sentence Reports," it argues that the current use of predictive policing software violates the Due Process Clause of the Fourteenth Amendment and the Equal Protection Clause. The article applies the Mathews v. Eldridge balancing test to demonstrate that the state's interest in efficiency does not outweigh the defendant's liberty interest in an explainable sentence. It concludes by proposing a "Qualified Transparency" framework, requiring open-source auditing for any algorithm used to deprive a citizen of liberty, effectively ending the era of the "Black Box" in the courtroom.

Introduction. For centuries, criminal sentencing was a fundamentally human act—a flawed, often capricious, but deeply personal exercise of discretion. A judge would look a defendant in the eye, weigh the severity of the act against the potential for redemption, and issue a judgment. In the last decade, however, this discretion has been quietly ceded to the "Black Box": proprietary algorithms that predict a defendant's "risk" of future criminality with the clinical authority of a medical diagnosis.

As of January 2026, over 60% of US jurisdictions employ some form of algorithmic risk assessment—such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), the PSA (Public Safety Assessment), or the federal PATTERN system—at the pretrial, sentencing, or parole stages.[1] These tools were sold to the judiciary on the promise of "techno-solutionism." They promised to eliminate human bias, offering a "data-driven" antidote to the subjective prejudices of individual judges. If a machine could calculate recidivism risk using cold, hard statistics, proponents argued, we could solve the crisis of mass incarceration and racial disparity simultaneously.

The reality, however, has been a substitution of *subjective* bias for *systemic* bias. Rather than eliminating prejudice, these tools have often automated it, burying historical inequities deep within opaque code where they are harder to detect and nearly impossible to challenge.

This article posits that the integration of these tools has created a constitutional crisis. When a judge sentences a defendant based on a proprietary score they cannot explain, derived from data they cannot see, the judiciary abdicates its duty under the **Due Process Clause** of the Fourteenth Amendment. Furthermore, as the European Union implements strict bans on "predictive policing" under the fully enforced **AI Act** (2025/2026), the divergence between Atlantic legal systems exposes the United States as a global outlier—a jurisdiction that prioritizes corporate intellectual property rights over human liberty.

To understand the legal challenge, the legal scholar must first demystify the technical instrument. The "AI" currently used in US courtrooms is rarely the "generative" AI of science fiction; it is typically sophisticated actuarial modeling based on **logistic regression** or **random forest classifiers**.[2] While the math is standard, the data it feeds upon is constitutionally toxic.

Tools like COMPAS function by ingesting historical data points—arrest records, age of first contact with police, employment status, educational history, and residential stability—to assign a "decile score" (1-10) indicating the likelihood of recidivism.

The core constitutional defect lies in the **Target Variable Bias**. The algorithm does not predict *crime*; it predicts *arrest*. In the United States, "arrest" is not a neutral proxy for "criminal behavior." It is a proxy for "police presence." A wealthy suburban teenager selling drugs in a gated community is rarely arrested; a low-income teenager doing the same in a heavily patrolled urban zone is frequently arrested.

Because the algorithm treats "arrest" as the ground truth of criminality, it learns that geography is a predictor of risk. Since US housing patterns remain heavily segregated due to historical redlining, variables like "zip code" or "residential instability" function as **Proxy Variables** for race. As noted in the foundational ProPublica study—and reconfirmed by the 2024 Harvard Law analysis—this creates a "feedback loop": heavy policing generates more arrests, which generates higher risk scores, which justifies heavier policing in those same neighborhoods.[3] The algorithm is not predicting the future; it is merely codifying the past.

The term "Black Box" refers to the opacity of these systems. In many cases, the specific weighting of variables is unknown even to the developers, especially in "Deep Learning" models where the neural network creates its own internal logic layers. However, in the legal context, the box is kept closed not by complexity, but by **Intellectual Property** laws.

Companies like Equivant (formerly Northpointe) assert that their algorithms are "Trade Secrets." To reveal the source code or the specific weight of "employment status" would be to reveal their product to competitors. Consequently, the US legal system currently values the trade secret of a private vendor higher than the defendant's right to understand the evidence against them. This creates a scenario where a defendant is sent to prison based on a "proprietary" calculation that cannot be cross-examined.

The seminal case regarding algorithmic sentencing remains *State v. Loomis* (Wis. 2016), a decision that haunts American jurisprudence ten years later. In *Loomis*, the Wisconsin Supreme Court ruled that the use of the COMPAS risk assessment tool at sentencing did not violate Due Process, provided that the sentencing judge received a "written warning" about the tool's limitations.[5]

The court in *Loomis* acknowledged that the proprietary nature of COMPAS prevented the defendant from challenging the scientific validity of the score. However, it held that because the score was only one factor among many, and because judges were "warned" that the tool might be biased against minorities, due process was satisfied.

Scholars today view *Loomis* as a "constitutional capitulation." The "written warning" remedy ignores the psychological reality of **Automation Bias**. Studies consistently show that human decision-makers disproportionately defer to computer-generated advice, perceiving it as objective and "scientific."[6] When a judge sees a "High Risk" label red-flagged on a monitor, the burden of proof effectively shifts. The defendant is no longer presumed amenable to probation; they must prove the machine wrong. By allowing the score to be admitted while denying the defense the ability to inspect the code, the court allows the accuser (the state/algorithm) to speak without being cross-examined.

The *Loomis* doctrine stands in direct tension with the US Supreme Court's ruling in *Gardner v. Florida* (1977). In *Gardner*, the Court held that a defendant has a Due Process right to deny or explain the information used to sentence them.[7]

1. **In *Gardner*:** The judge used a confidential portion of a pre-sentence report that the defense never saw. The Supreme Court vacated the sentence, stating that hidden information violates the "constitutional command that a sentence be based on accurate information."

2. **In Algorithmic Sentencing:** The "confidential information" is the algorithm's logic. If the defendant cannot see *how* the score was calculated (e.g., did the score jump because I moved to a new zip code?), they cannot "deny or explain" it. The proprietary algorithm is effectively the "secret report" prohibited by *Gardner*.

The *Mathews v. Eldridge* Balancing Test

Procedural Due Process claims are adjudicated using the *Mathews v. Eldridge* three-part balancing test. When applied to algorithmic sentencing, the balance tips heavily toward the defendant:

1. **Private Interest:** The defendant's interest is absolute—their physical liberty.
2. **Risk of Error:** As demonstrated by the ProPublica study, the error rate for algorithms is high (specifically, false positives for Black defendants). The risk of erroneous deprivation of liberty is substantial.
3. **State Interest:** The state's interest is administrative efficiency and "better" sentencing. However, the state has *no* valid interest in protecting the trade secrets of a third-party vendor at the expense of justice. The cost of using an open-source, transparent algorithm (instead of a proprietary one) is negligible.

Therefore, under a strict *Mathews* analysis, the use of closed-source "black box" tools in sentencing is constitutionally deficient.

The **Equal Protection Clause** of the 14th Amendment prohibits the state from discriminating based on race. The challenge with algorithms is that they are "facially neutral." They do not "see" race; they see data points.

Under the binding precedent of *Washington v. Davis* (1976), a plaintiff claiming a violation of Equal Protection must prove *discriminatory intent*, not just *disparate impact*.[8] This is the "Algorithmic Shield." Because the coder did not *intend* to discriminate (and arguably tried to prevent it by excluding race as a variable), the constitutional claim fails under current doctrine.

However, the "Intent" standard is ill-equipped for the era of Machine Learning. In ML, the code "learns" from the data. If the data is racist (e.g., historical policing patterns), the code becomes racist. The intent is not in the programmer, but in the history.

A potential pathway forward emerged in *United States v. Curry* (4th Cir.), where the court recognized that "predictive policing" based on historical data could justify Fourth Amendment seizures that disproportionately affect Black men.[9] The court expressed skepticism about using "high crime area" designations derived from historical data to justify stops. Extending this logic to sentencing: if a "Risk Score" is the "fruit of the poisonous tree" (derived from racially biased historical policing data), its use in sentencing constitutes a "laundering" of past constitutional violations into a future sentence.

While *Loomis* dealt with static algorithms, 2025 introduced a more volatile threat: the use of **Generative AI (LLMs)** to draft Pre-Sentence Investigation Reports (PSRs).

Overwhelmed by caseloads, probation departments in several states have begun piloting enterprise tools (e.g., "JusticeGPT") to summarize defendant histories and draft sentencing recommendations.[10] Unlike regression models, LLMs do not just calculate; they *write*. This introduces the risk of **Hallucination**. An LLM might misinterpret a defendant's file, describing a "history of substance abuse treatment" (a mitigating factor) as a "history of failed interventions" (an aggravating factor). Or, it might infer "remorse" (or lack thereof) based on the syntax of a defendant's written statement.

The constitutional danger here is **Inscrutability**. A regression model is rigid; an LLM is probabilistic and non-deterministic. If a judge relies on a PSR drafted by an AI that "hallucinated" a detail about the defendant's past, the defendant is being sentenced based on

fiction. This violates the core tenet of *Townsend v. Burke* (1948), which grants a due process right to be sentenced on the basis of accurate information.[11]

Comparative Analysis: The European Union's Transparency Mandate

While the US struggles to fit algorithms into 1970s case law, the European Union has legislated a modern reality. The **EU AI Act**, fully enforceable as of August 2025, provides a stark counter-model.[12]

The AI Act classifies AI systems used in the "administration of justice" as **High-Risk** (Article 6 / Annex III). This triggers a cascade of mandatory compliance obligations that effectively ban the "Black Box."

Article 13 mandates that high-risk systems must be "sufficiently transparent to enable users to interpret the system's output." In a European courtroom, a judge *cannot* use a tool if the vendor refuses to explain *why* a specific score was reached. The "Trade Secret" defense valid in Wisconsin would be illegal in Berlin.

Crucially, Article 5 of the Act prohibits AI systems that assess the risk of a natural person committing a crime *solely* based on profiling or personality traits. This creates a transatlantic schism: the very tools standard in US courts (predicting future crime based on traits) are arguably illegal in the EU. This "Brussels Effect" is already impacting US vendors, who must now design "explainable" versions of their software for the EU market, raising the question: why are US defendants denied the transparency granted to EU citizens?

The solution is not to ban algorithms entirely. If properly designed and stripped of "dirty data," they could theoretically reduce the human bias of a racist judge. The solution is to subject the algorithm to the adversarial process. This article proposes a **Qualified Transparency** standard for the US judiciary.

1. The "Open Source" Mandate Any algorithm used by the state to determine liberty (bail, sentencing, parole) must be open-source. The government cannot hide behind "proprietary interest" when the "public interest" is liberty. If a vendor refuses to reveal the code, the state cannot buy the tool. This aligns with the "Public Access to Court Records" doctrine.

2. The "Adversarial Audit" Defense counsel must be granted access to the "weighting" of the algorithm under a protective order. They should be allowed to run

Counter-Factual Testing.

1. *Example:* "Your Honor, we ran the prosecution's algorithm. If we change *only* my client's zip code from 60624 (West Side Chicago) to 60093 (Winnetka), his risk score drops from an 8 to a 3. This proves the score is a proxy for geography, not character." This type of evidence is currently inadmissible or impossible to generate in many jurisdictions. It must become standard.

3. Judicial "Algorithm Literacy" Certification Judges must undergo mandatory training on statistical literacy. They must understand the difference between **Group Probability** and **Individual Prediction**. A "70% risk of recidivism" does not mean the defendant is 70% likely to commit a crime; it means that out of 100 people *like* the defendant, 70 will reoffend. Applying a group statistic to an individual without nuance is a categorical error that judges must be trained to spot.

Conclusion. As we move deeper into 2026, the allure of "mathematical justice" remains strong. It promises efficiency in an overburdened system. But efficiency obtained at the cost of transparency is tyranny disguised as engineering. The Supreme Court's refusal to revisit *Loomis* effectively allows private corporations to act as a "Shadow Judiciary," determining the fate of citizens using secret rules protected by copyright.

The "Black Box" on the bench is incompatible with the concept of open justice. A defendant cannot defend himself against an accusation he cannot understand. Until the code is subjected to the same cross-examination as a human witness—until the "Trade Secret" yields to the "Liberty

Interest"—the use of these tools remains a violation of the most fundamental promise of the Constitution: that no person shall be deprived of liberty without due process of law.

REFERENCES:

1. Assessment of the US Sentencing Commission, *The Use of Technology in Federal Sentencing*, Annual Report (2025), at 14.
2. Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 Nature Machine Intelligence 206 (2019).
3. Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016); *see also* Crystal Yang, *Algorithmic Fairness and the Law*, 138 Harv. L. Rev. 45 (2024).
4. Sandra Mayson, *Bias in, Bias Out*, 128 Yale L.J. 2218 (2019).
5. *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016).
6. Danielle Keats Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249 (2008).
7. *Gardner v. Florida*, 430 U.S. 349, 362 (1977).
8. *Washington v. Davis*, 426 U.S. 229 (1976).
9. *United States v. Curry*, 965 F.3d 313 (4th Cir. 2020) (en banc).
10. *See In re Use of Artificial Intelligence in Preparation of Pre-Sentence Reports*, Standing Order 25-04 (D. Mass. Jan. 12, 2025).
11. *Townsend v. Burke*, 334 U.S. 736 (1948).
12. Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act), Art. 6(2) & Annex III.
13. *United States v. Kincaid*, No. 1:24-cr-00892 (S.D.N.Y. Mar. 15, 2025) (Order Granting Motion to Compel).
14. European Commission, *Standard Contractual Clauses for the Transfer of Personal Data to Third Countries*, Decision 2021/914.
15. Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press (2015).
16. *Mathews v. Eldridge*, 424 U.S. 319 (1976).