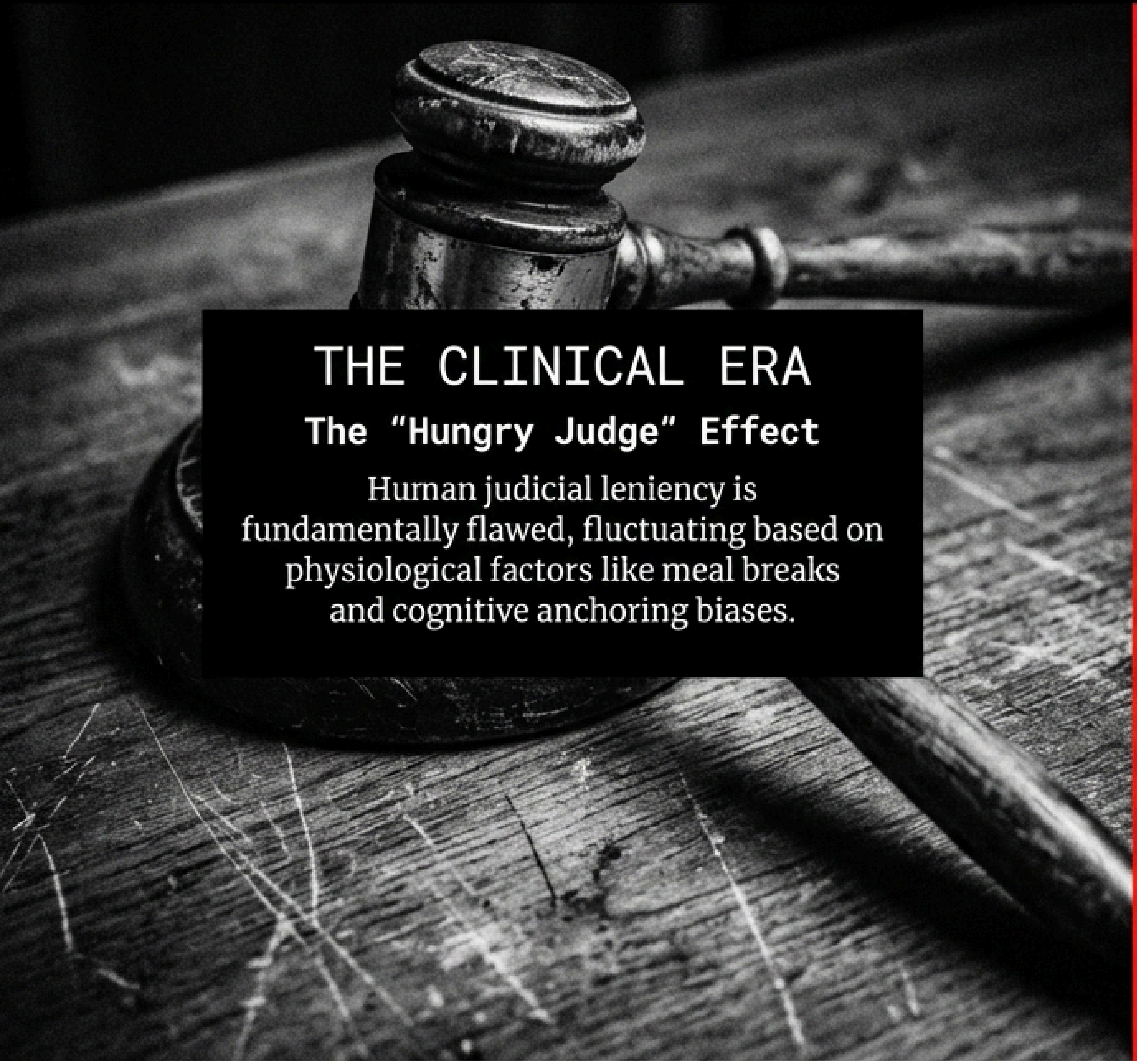
The Algorithmic Panopticon

A Forensic Analysis of the COMPAS Recidivism Case and the Fragmentation of Societal Fairness Norms.

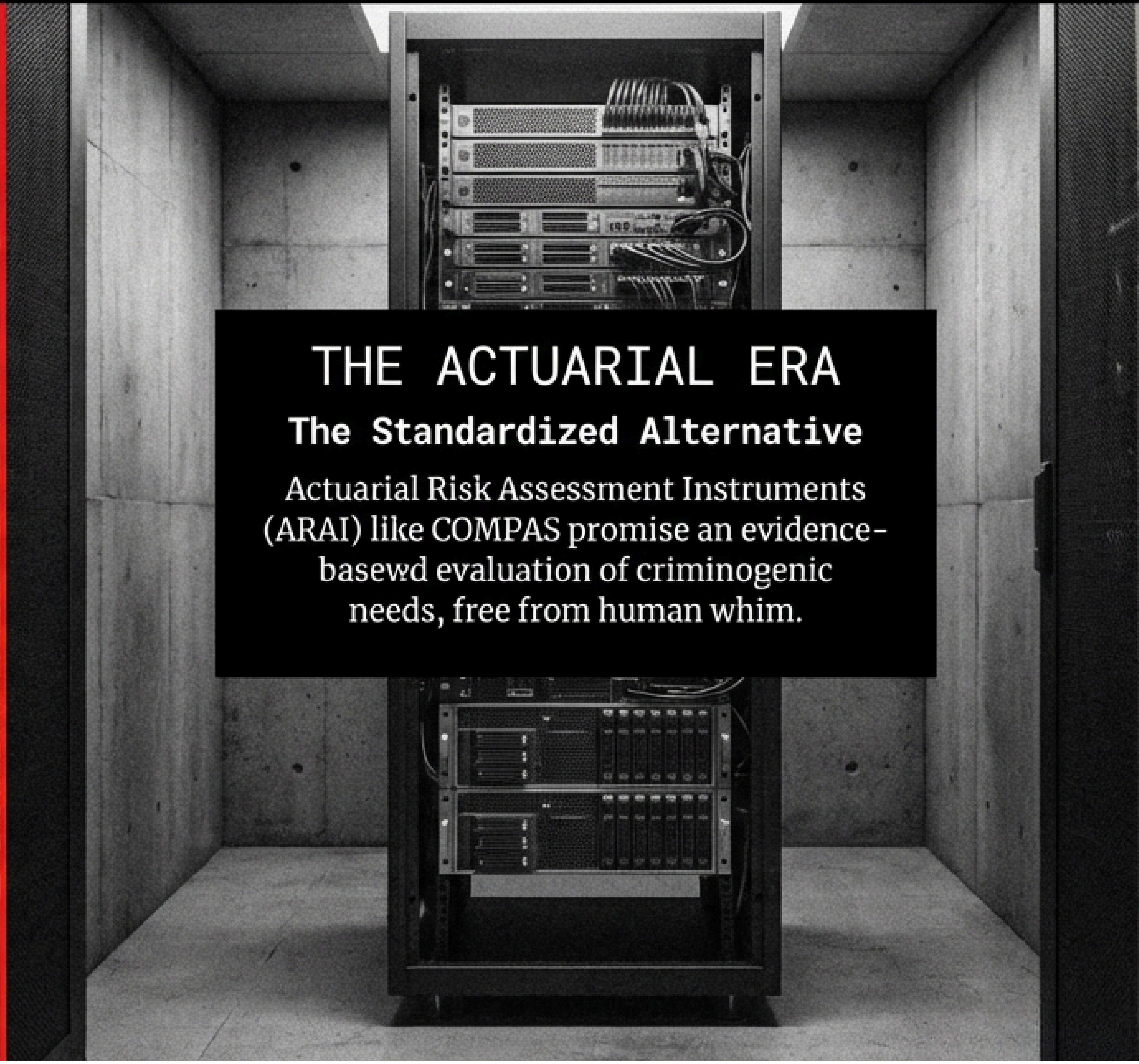
DELEGATING JUSTICE TO THE MACHINE



THE CLINICAL ERA

The “Hungry Judge” Effect

Human judicial leniency is fundamentally flawed, fluctuating based on physiological factors like meal breaks and cognitive anchoring biases.



THE ACTUARIAL ERA

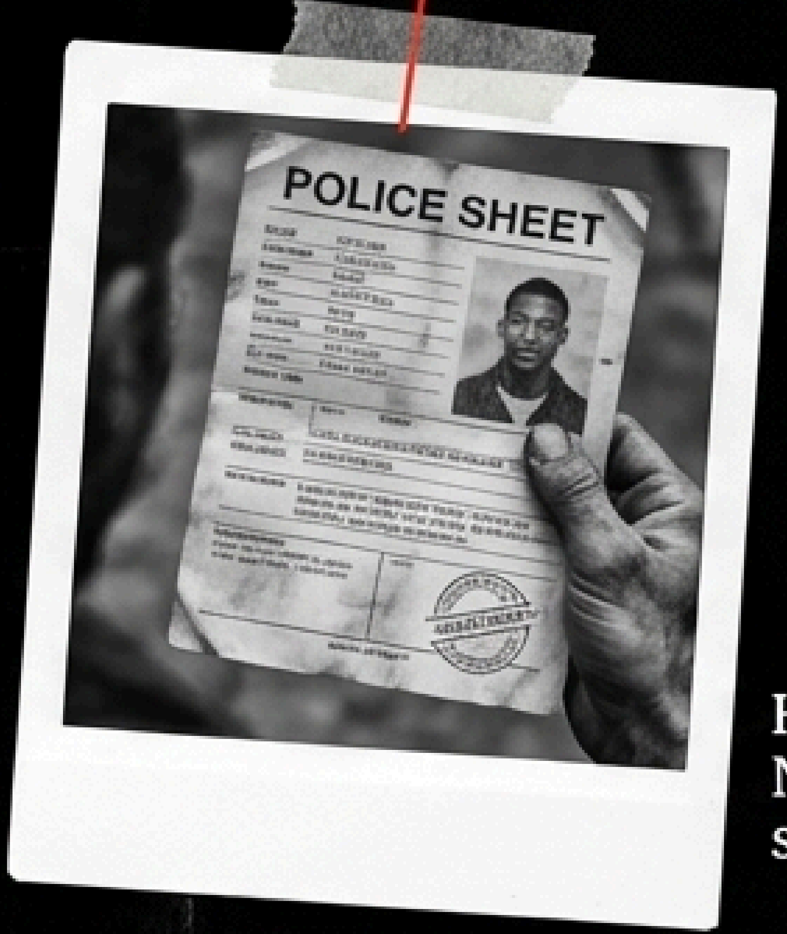
The Standardized Alternative

Actuarial Risk Assessment Instruments (ARAI) like COMPAS promise an evidence-basewd evaluation of criminogenic needs, free from human whim.

INSIDE THE BLACK BOX OF RISK PREDICTION

$$V = w_1 \cdot A + w_2 \cdot A_{first} + w_3 \cdot H_v + w_4 \cdot E_v + w_5 \cdot N_c$$

Current Age & Age at First Arrest. Heavily weighted, statistically correlated with violence, but heavily influenced by systemic neighborhood policing.

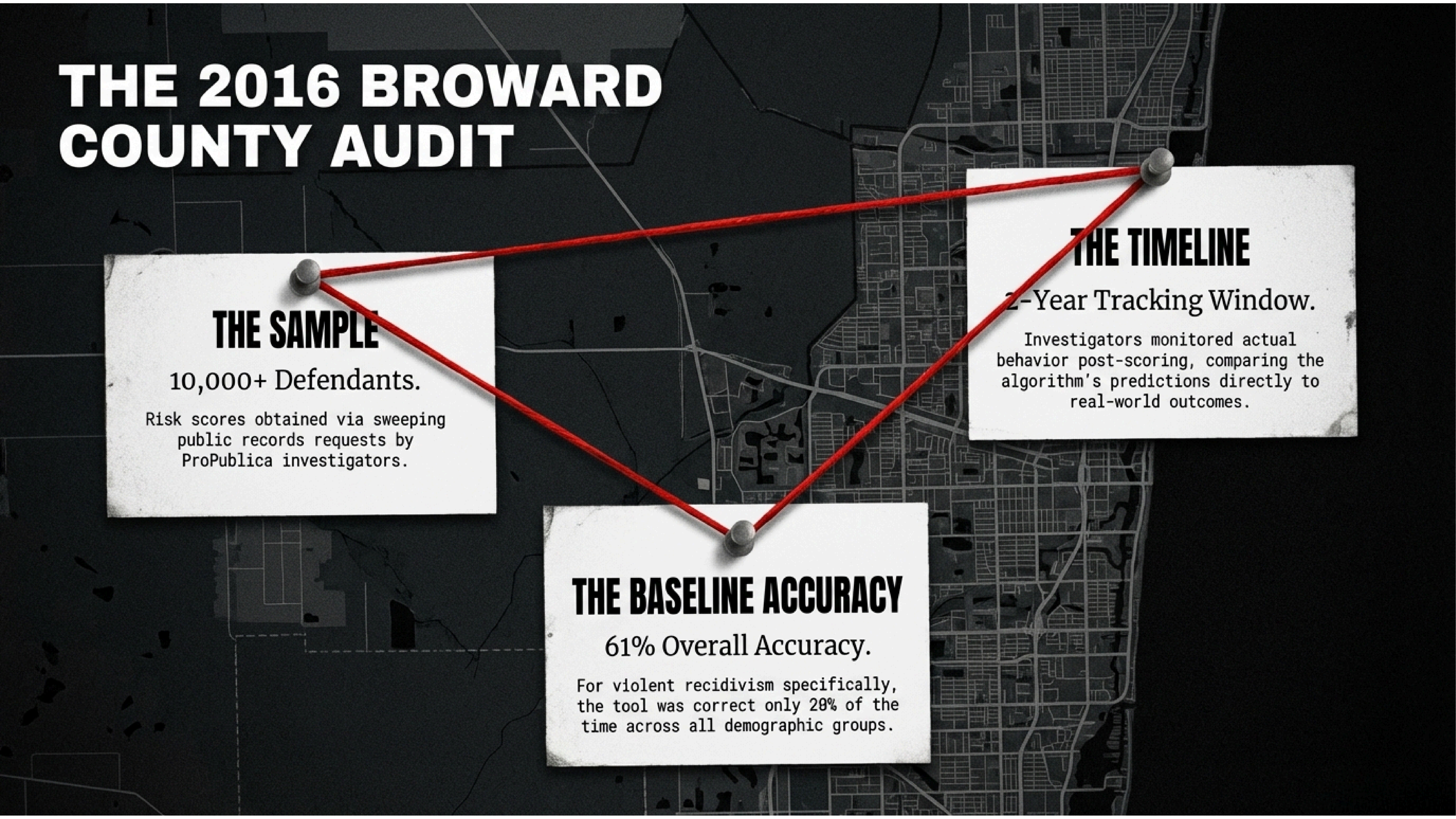


History of Violence & Noncompliance. The static criminal record.

Vocational/Educational Scale. Captures socioeconomic stability, punishing poverty.

COMPAS does not generate a ‘guilt’ score. It labels defendants as Low, Medium, or High risk across dimensions of pretrial misconduct, general recidivism, and **violent recidivism** based on these weighted inputs.

THE 2016 BROWARD COUNTY AUDIT



THE SAMPLE

10,000+ Defendants.

Risk scores obtained via sweeping public records requests by ProPublica investigators.

THE TIMELINE

2-Year Tracking Window.

Investigators monitored actual behavior post-scoring, comparing the algorithm's predictions directly to real-world outcomes.

THE BASELINE ACCURACY

61% Overall Accuracy.

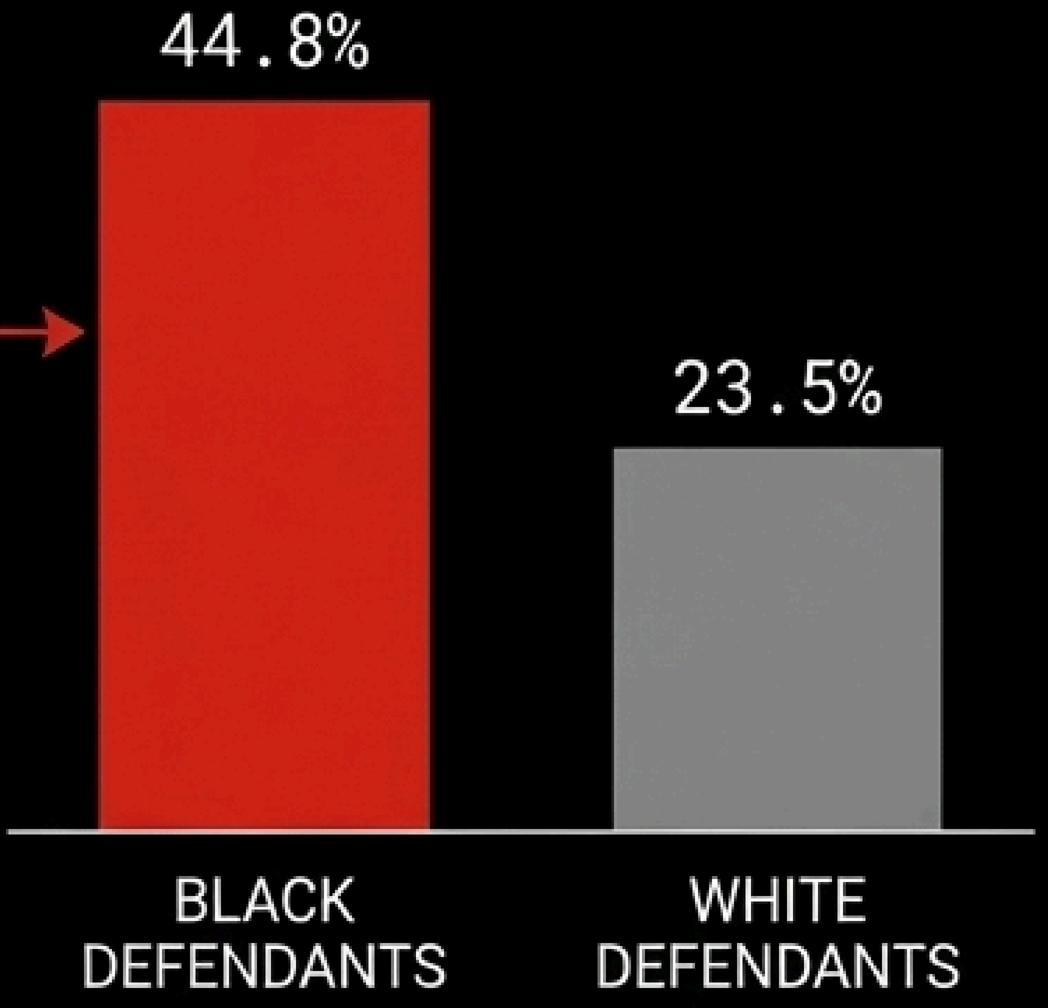
For violent recidivism specifically, the tool was correct only 20% of the time across all demographic groups.

THE DISPARATE BURDEN OF ALGORITHMIC ERROR

THE FALSE POSITIVE RATE

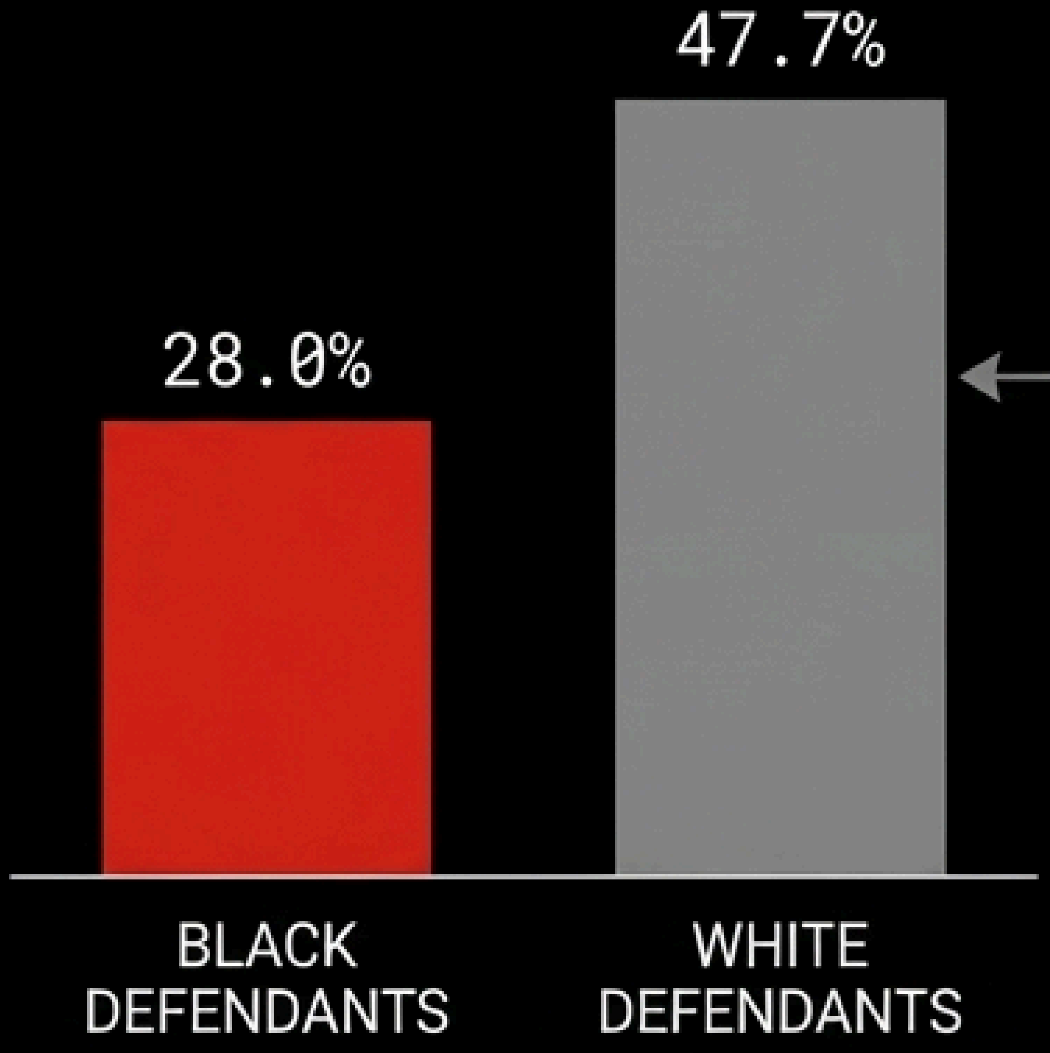
(Mislabeled High Risk)

Black non-recidivists were nearly twice as likely to be misclassified as dangerous.



THE FALSE NEGATIVE RATE

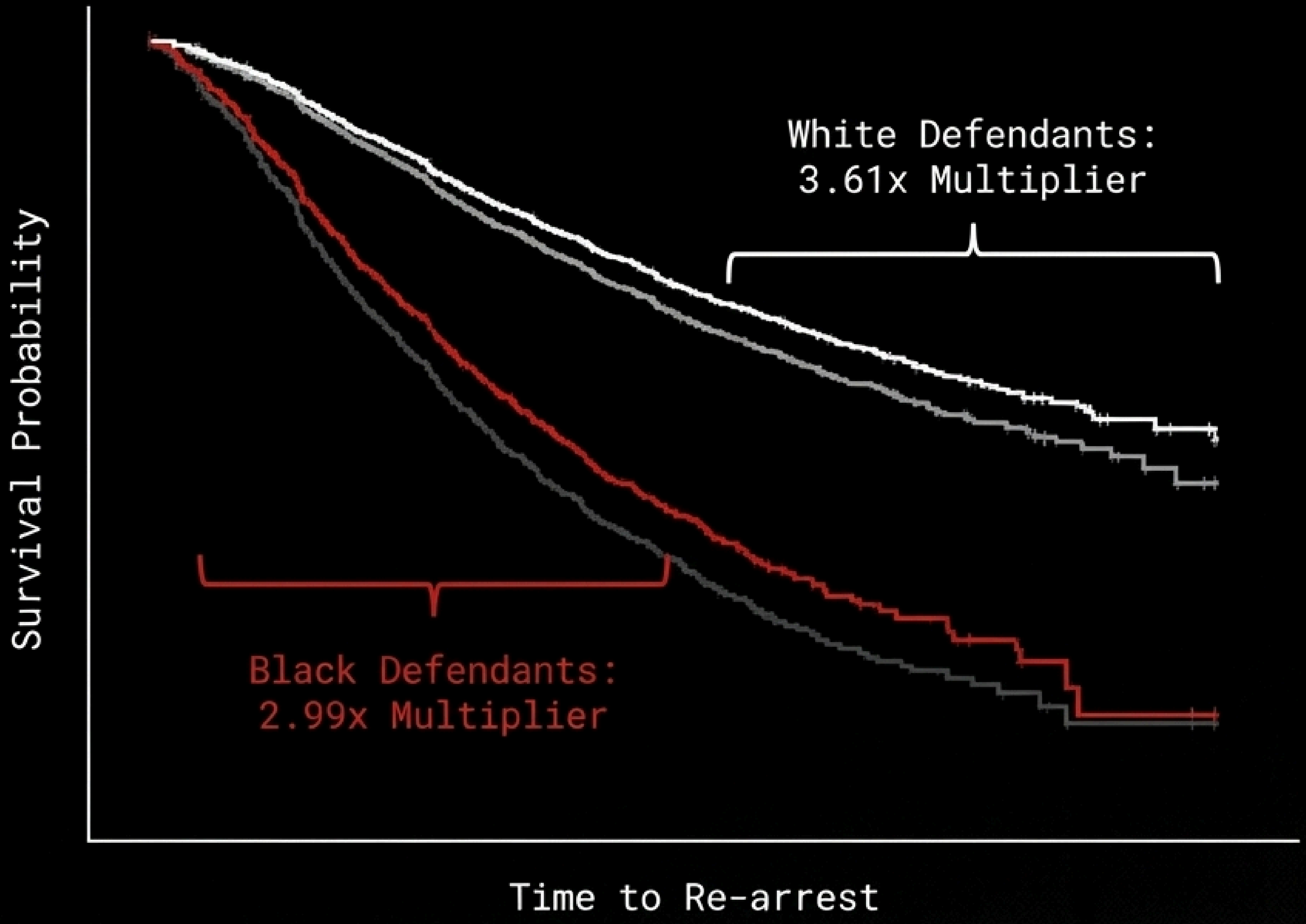
(Mislabeled Low Risk)



White recidivists were almost twice as likely to be misclassified as safe.

Takeaway: Even after controlling for prior crimes, age, and gender, Black defendants were 45% more likely to receive higher risk scores for general recidivism and 77% more likely for violent recidivism.

THE HAZARD RATIO DISPARITY



THE WEIGHT OF A LABEL

A “High Risk” label carries a fundamentally different statistical weight depending on the defendant’s race.

Because the multiplier is drastically different, the algorithm violates the foundational norm of consistent measurement across groups.

To the machine, risk is not absolute; it is relative to demographic identity.

THREE CONFLICTING DEFINITIONS OF FAIRNESS

01. CALIBRATION

Predictive Parity

A score of 7 means a 60% chance of recidivism for anyone, regardless of race.

Beneficiary: The Institution. (COMPAS passed this test. This was the corporate defense.)

02. EQUALIZED ODDS

Error Rate Balance

False positive and false negative rates must be equal across all demographic groups.

Beneficiary: The Defendant. (COMPAS failed this massively. This was the civil rights critique.)

03. STATISTICAL PARITY

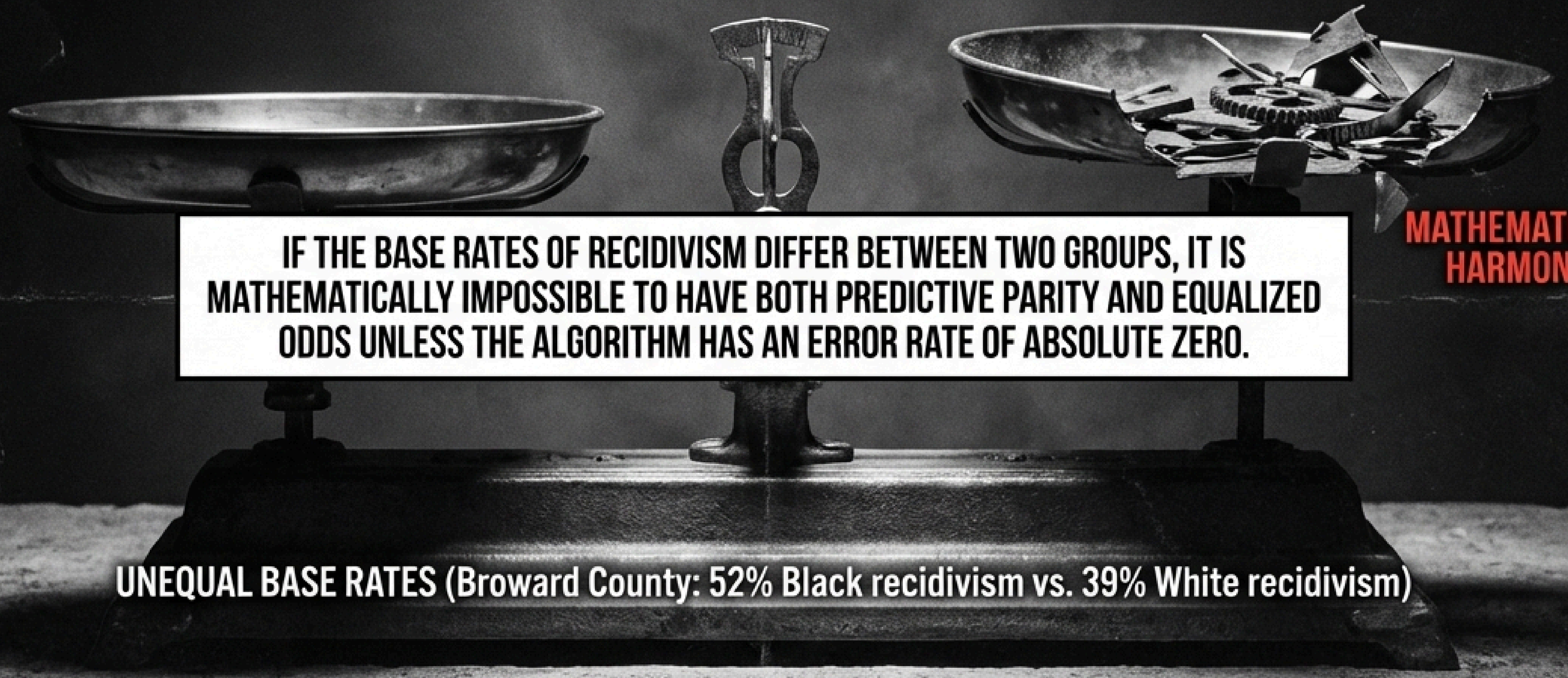
Equal Outcomes

The exact same proportion of each group is classified as high risk.

Reality: Rarely achieved without direct, manual intervention due to differing real-world arrest rates.

THE MATHEMATICAL IMPOSSIBILITY THEOREM

$$P(Y=1|\hat{Y}=1, G=B) = P(Y=1|\hat{Y}=1, G=W) \neq P(\hat{Y}=1|Y=0, G=B) = P(\hat{Y}=1|Y=0, G=W)$$



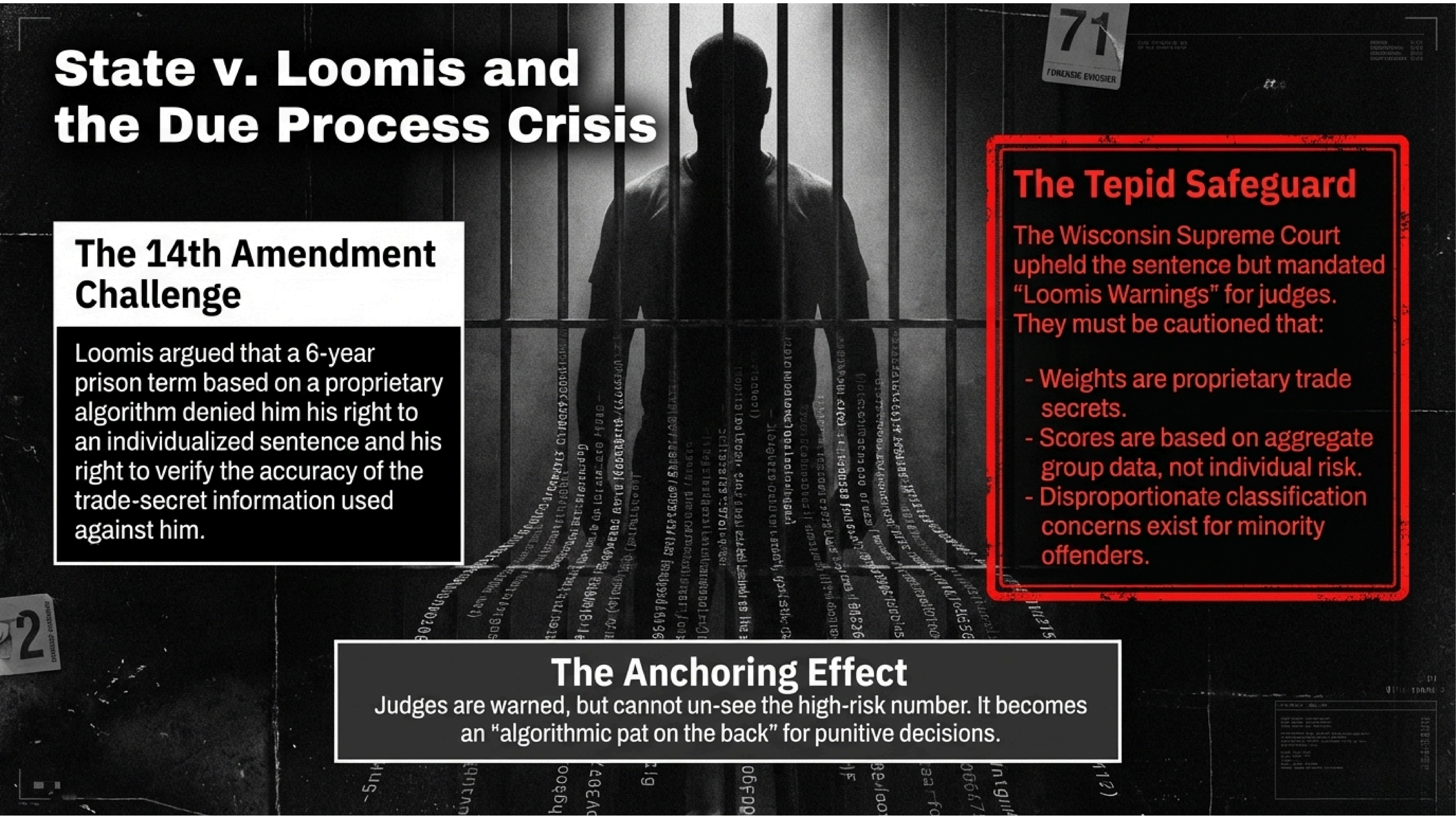
IF THE BASE RATES OF RECIDIVISM DIFFER BETWEEN TWO GROUPS, IT IS MATHEMATICALLY IMPOSSIBLE TO HAVE BOTH PREDICTIVE PARITY AND EQUALIZED ODDS UNLESS THE ALGORITHM HAS AN ERROR RATE OF ABSOLUTE ZERO.

MATHEMATICAL HARMONY

UNEQUAL BASE RATES (Broward County: 52% Black recidivism vs. 39% White recidivism)

A system that is “fair” to the institution (accurate scores) is inherently “unfair” to the marginalized group (disproportionate false positives). It is a zero-sum game.

State v. Loomis and the Due Process Crisis



The 14th Amendment Challenge

Loomis argued that a 6-year prison term based on a proprietary algorithm denied him his right to an individualized sentence and his right to verify the accuracy of the trade-secret information used against him.

The Tepid Safeguard

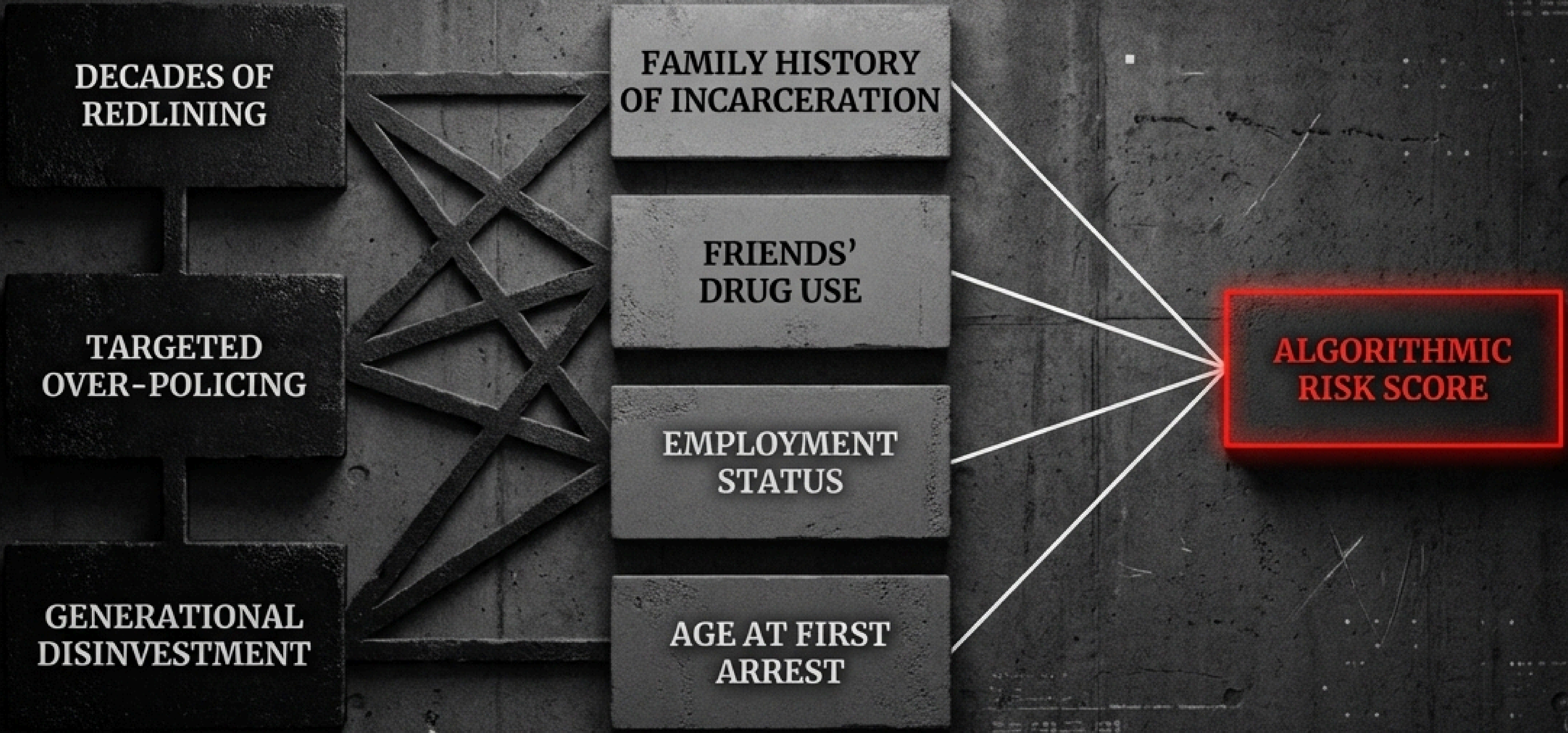
The Wisconsin Supreme Court upheld the sentence but mandated "Loomis Warnings" for judges. They must be cautioned that:

- Weights are proprietary trade secrets.
- Scores are based on aggregate group data, not individual risk.
- Disproportionate classification concerns exist for minority offenders.

The Anchoring Effect

Judges are warned, but cannot un-see the high-risk number. It becomes an "algorithmic pat on the back" for punitive decisions.

THE PROXY WEB OF SYSTEMIC INEQUALITY



DECADES OF REDLINING

TARGETED OVER-POLICING

GENERATIONAL DISINVESTMENT

FAMILY HISTORY OF INCARCERATION

FRIENDS' DRUG USE

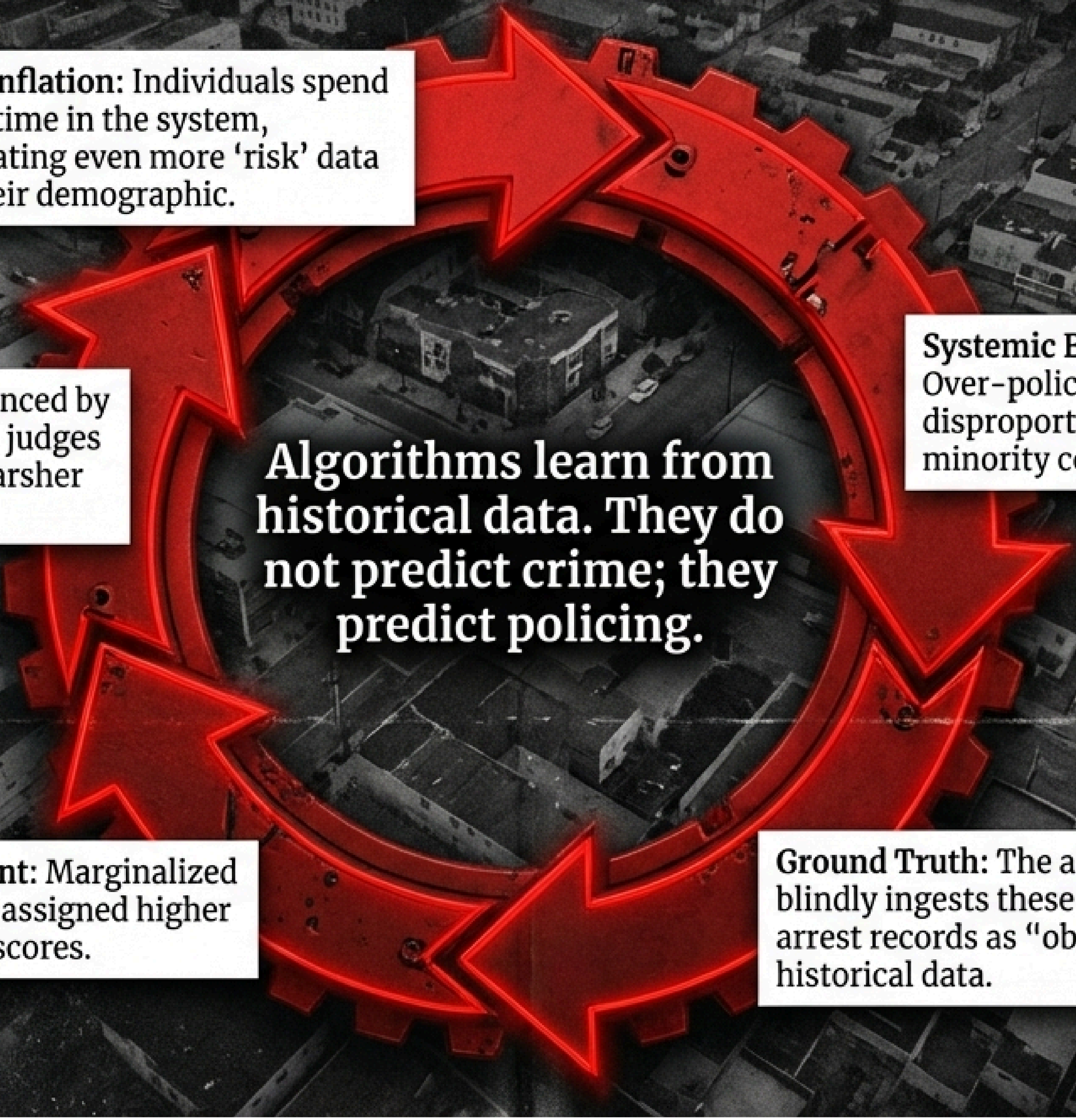
EMPLOYMENT STATUS

AGE AT FIRST ARREST

ALGORITHMIC RISK SCORE

GARBAGE IN, GARBAGE OUT. THE ALGORITHM NEVER EXPLICITLY ASKS FOR RACE. HOWEVER, IN THE AMERICAN CONTEXT, THESE SEEMINGLY NEUTRAL SOCIOECONOMIC SURVEY QUESTIONS SERVE AS HIGH-FIDELITY PROXIES FOR RACIAL IDENTITY DUE TO HISTORICAL STRUCTURAL BIAS.

The Runaway Feedback Loop



Algorithms learn from historical data. They do not predict crime; they predict policing.

Systemic Bias: Over-policing leads to disproportionate arrests in minority communities.

Ground Truth: The algorithm blindly ingests these skewed arrest records as “objective” historical data.

Risk Assignment: Marginalized individuals are assigned higher statistical risk scores.

Judicial Action: Influenced by the algorithmic score, judges deny bail or impose harsher sentences.

Data Inflation: Individuals spend more time in the system, generating even more ‘risk’ data for their demographic.

Divergent Paths in Risk Assessment

PATH A: CORPORATE EVOLUTION

COMPAS-R Core

- **Data Scope:** 83 questions (down from 125).
- **Socioeconomic Data:** Retains the 'Central 8' criminogenic factors (including proxies).
- **Transparency:** Point-additive 'Long Report' allows manual verification. Gender-neutral language adopted.

Critique: Solves the 'black box' opacity, but fundamentally ignores the base-rate and proxy variable problems.

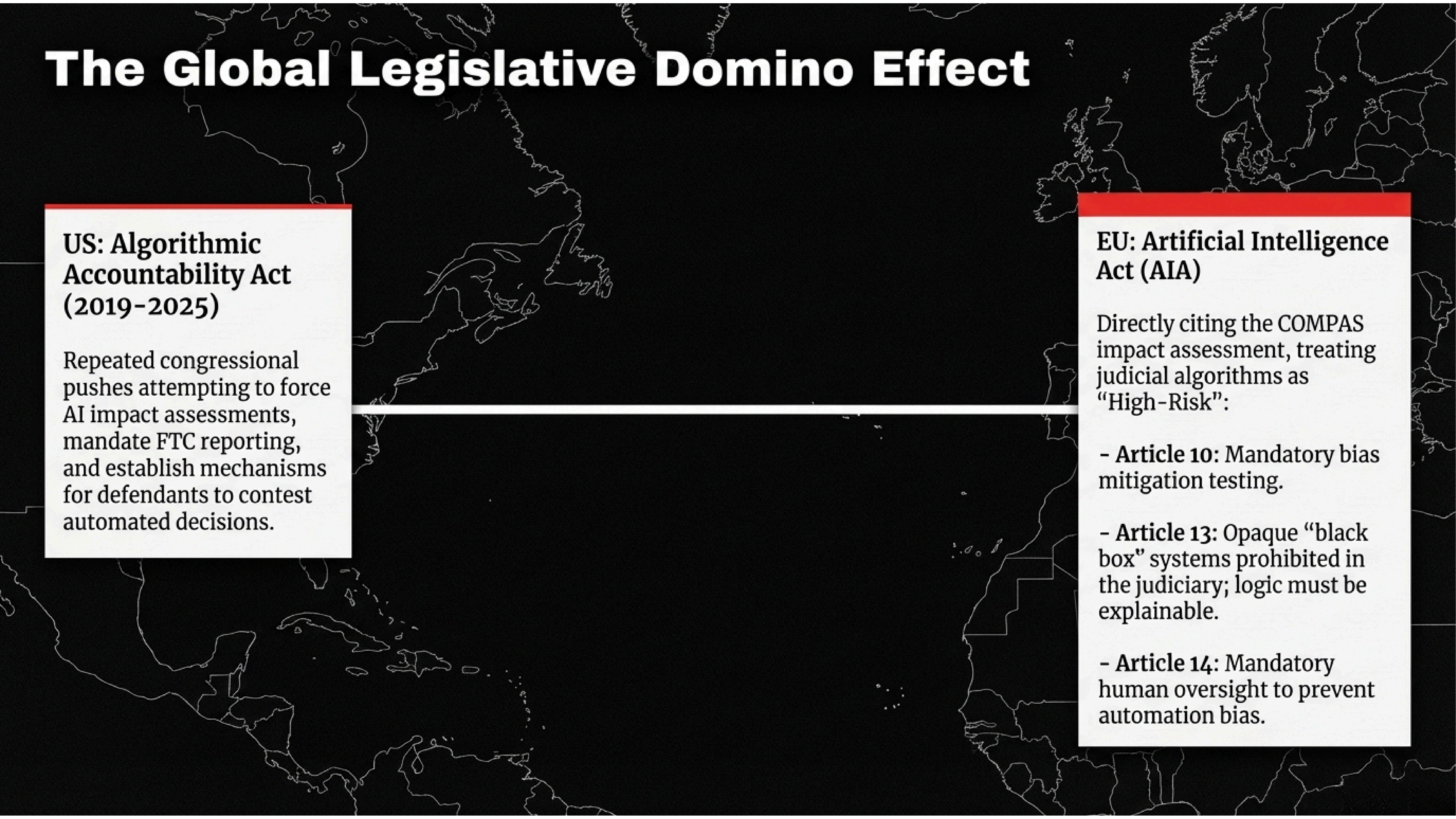
PATH B: TACTICAL RETREAT

New Jersey's Public Safety Assessment (PSA)

- **Data Scope:** Official criminal records only.
- **Socioeconomic Data:** Explicitly excluded to reduce systemic bias.
- **Transparency:** Publicly available factors and open-source mathematical weights.

Critique: Even with this 'bias-free' tool, NJ's Black incarcerated population remains identical at 54%—proving algorithms cannot fix skewed inputs.

The Global Legislative Domino Effect



US: Algorithmic Accountability Act (2019-2025)

Repeated congressional pushes attempting to force AI impact assessments, mandate FTC reporting, and establish mechanisms for defendants to contest automated decisions.

EU: Artificial Intelligence Act (AIA)

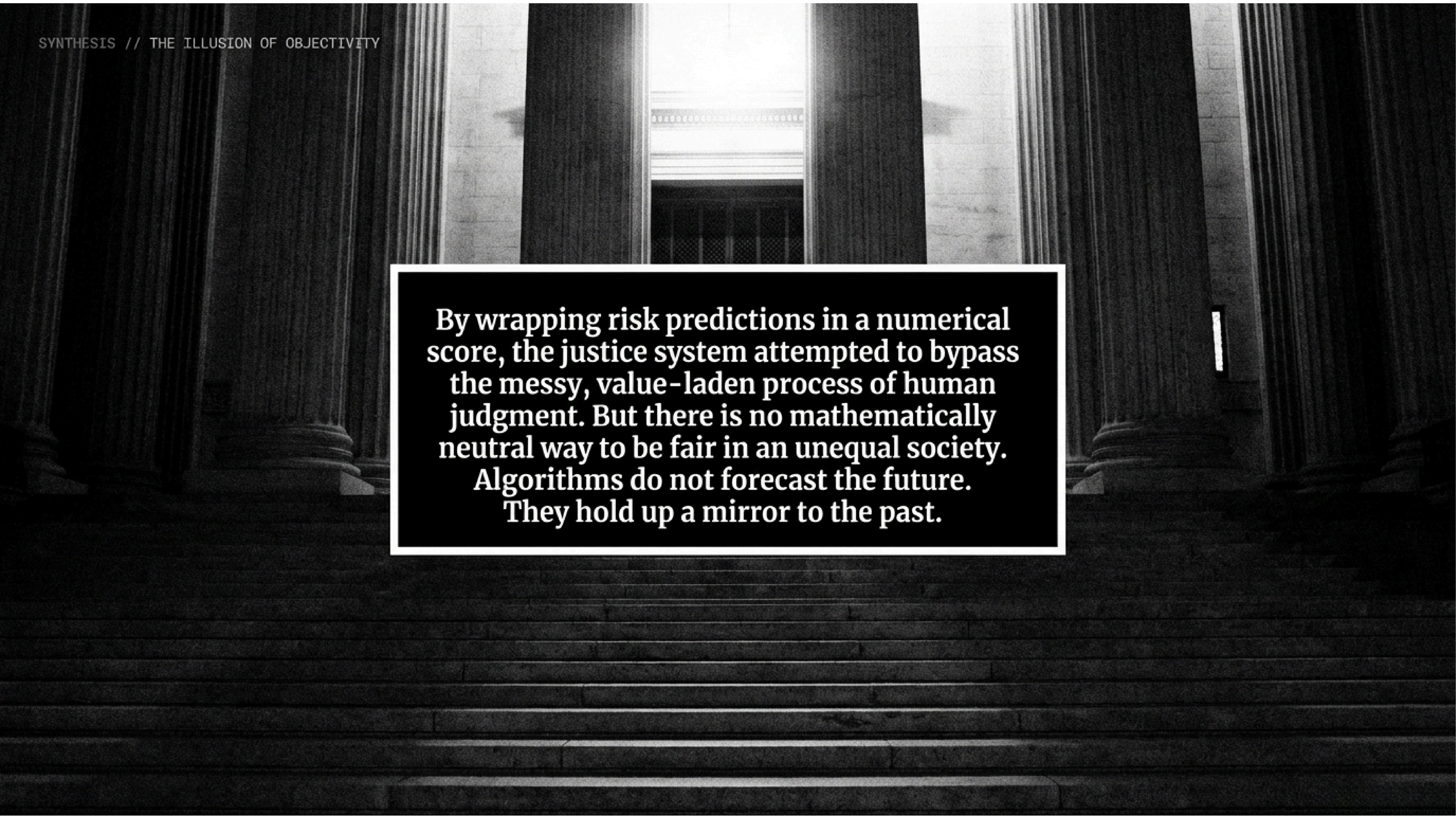
Directly citing the COMPAS impact assessment, treating judicial algorithms as "High-Risk":

- **Article 10**: Mandatory bias mitigation testing.

- **Article 13:** Opaque "black box" systems prohibited in the judiciary; logic must be explainable.

- **Article 14**: Mandatory human oversight to prevent automation bias.

SYNTHESIS // THE ILLUSION OF OBJECTIVITY



By wrapping risk predictions in a numerical score, the justice system attempted to bypass the messy, value-laden process of human judgment. But there is no mathematically neutral way to be fair in an unequal society. Algorithms do not forecast the future. They hold up a mirror to the past.

A Framework for Post-Actuarial Justice

I. RADICAL TRANSPARENCY

Abandoning trade secret protections for any proprietary tool that dictates or influences civil liberties. The right to be heard requires the right to be explained.

II. NORMATIVE ALIGNMENT

Acknowledging that fairness is a political and moral choice, not a scientific one. Jurisdictions must explicitly vote on which fairness metrics (Predictive Parity vs. Equalized Odds) they value most.

III. HUMAN DISCRETION

Designing robust mechanisms for judges to explicitly override risk scores that capture systemic socioeconomic hardship rather than individual, actionable danger.

The goal of a fair society is not merely to predict who will fail, but to interrupt the cycles of disadvantage that our algorithms so efficiently document.

Follow our work on

GlobalSouth.ai