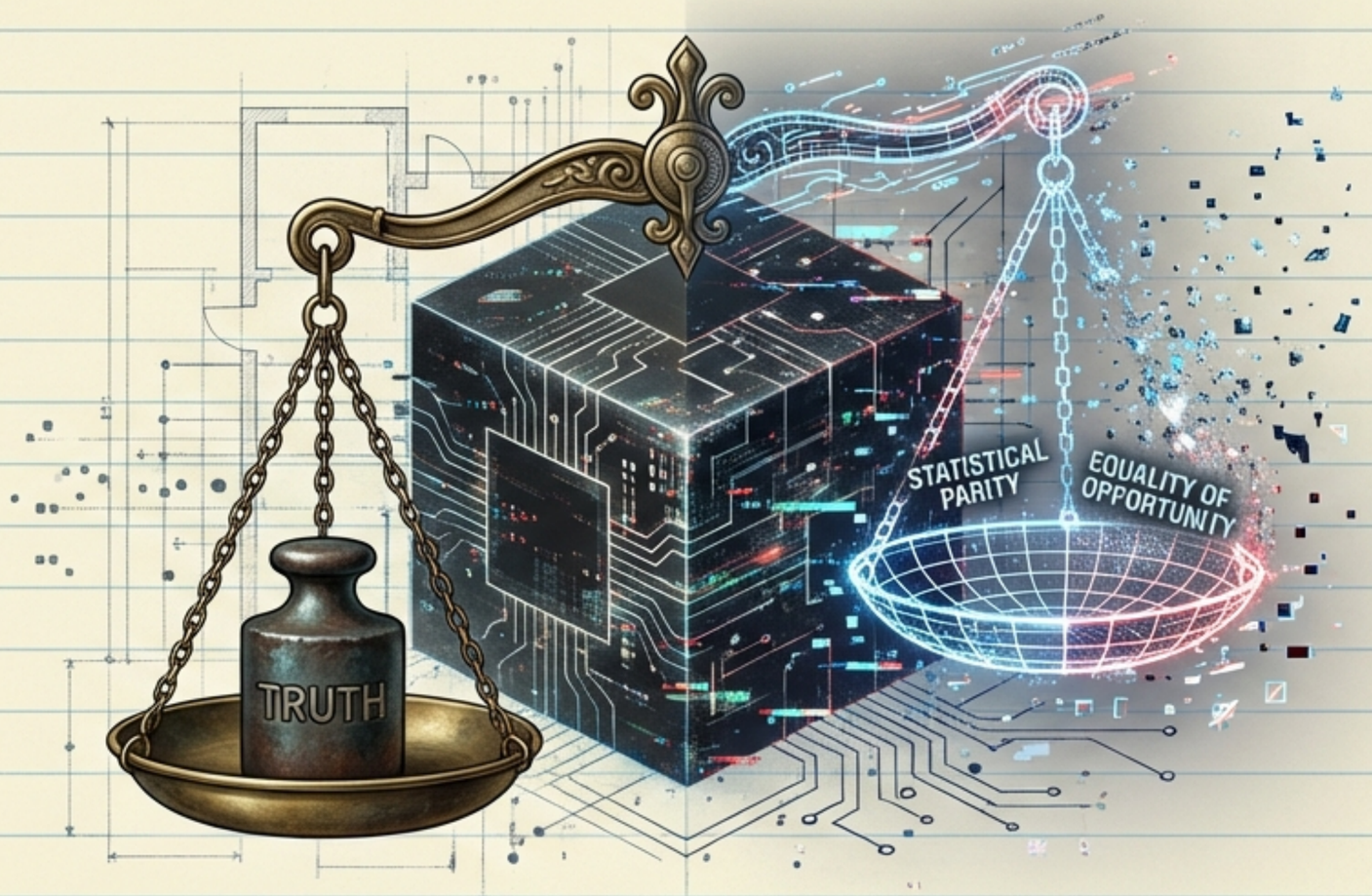
The Algorithm's Forked Tongue

Algorithmic Justice & The Impossibility of Statistical Fairness



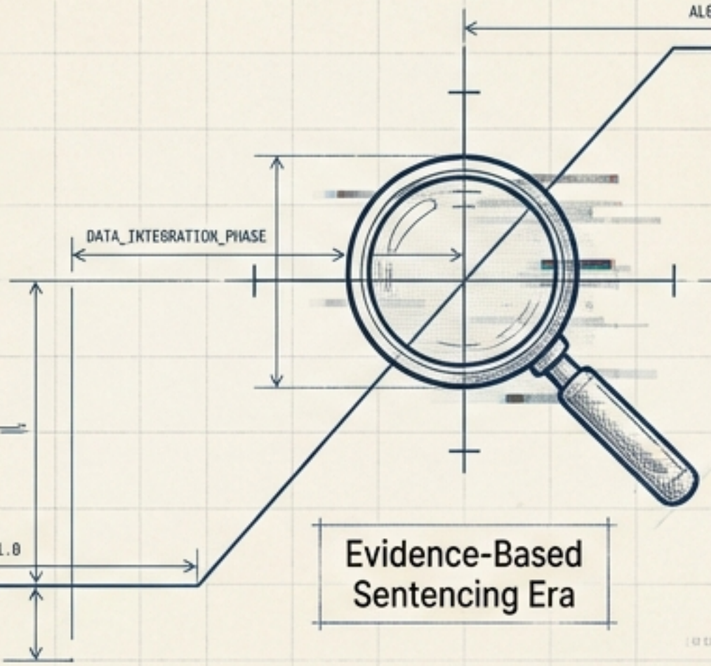
A deep dive into the COMPAS controversy, the limits of automated justice, and the mathematical deadlock between two valid definitions of truth.

The promise of evidence-based sentencing was to cure human bias

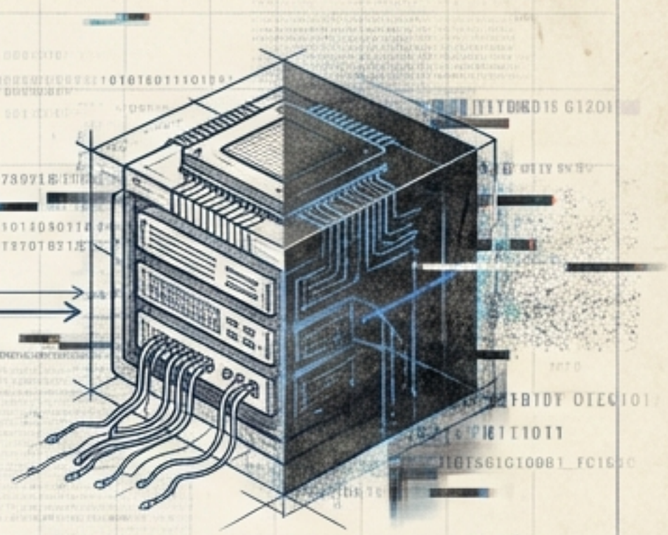
For decades, the American justice system sought to replace the "freewheeling" discretion of the mid-20th century with "actuarial tools."

The Goal: To use data to determine prison sentences, bond amounts, and parole eligibility with scientific rigour and a "veneer of objectivity."

Mid-20th Century:
Judicial Discretion



Evidence-Based
Sentencing Era



Present Day:
Algorithmic Risk Assessment

"These high-stakes decisions are increasingly informed by algorithmic risk assessments... but the shine of automated justice was scrubbed away in 2016."

A collision between two definitions of truth

ProPublica

The Accusation: In 2016, an investigative analysis of Broward County, Florida data revealed the COMPAS algorithm was racially biased.

Dataset: 6,167 defendants analysed via the AI Fairness 360 toolkit.

Northpointe (equivant)

The Defence: The creators of the proprietary software argued the model was mathematically sound based on the data it was fed.

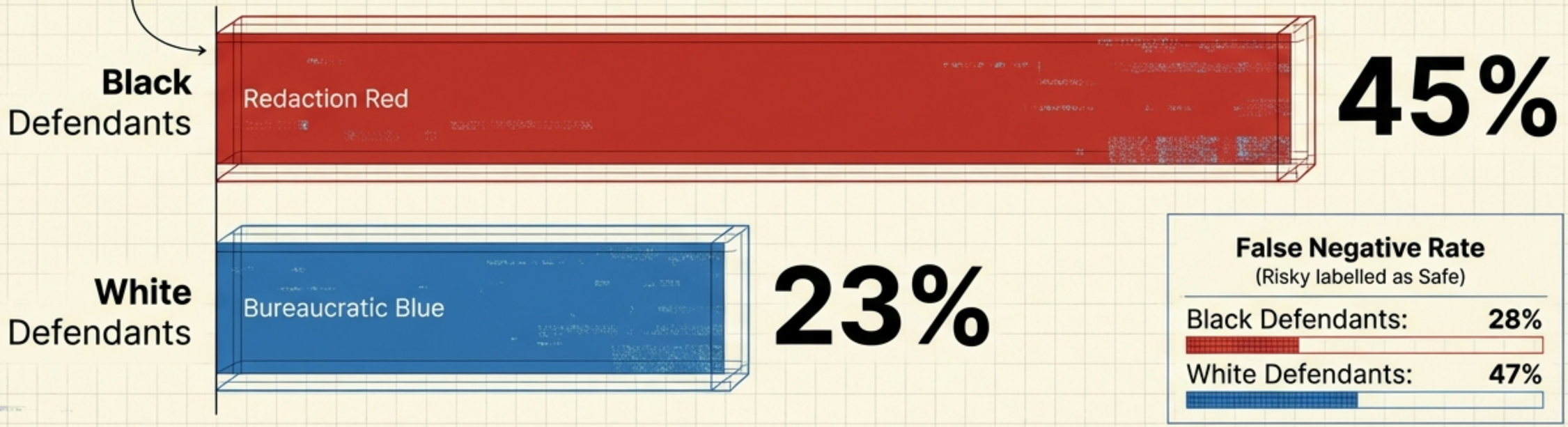
The Claim: 'Accuracy Equity' and predictive validity.

Key Insight: This wasn't just a technical glitch; it was a fundamental disagreement on what 'fairness' mathematically means.

The algorithm distributes errors unevenly along racial lines

Defendants labelled High Risk who did **NOT** reoffend.

False Positive Rate

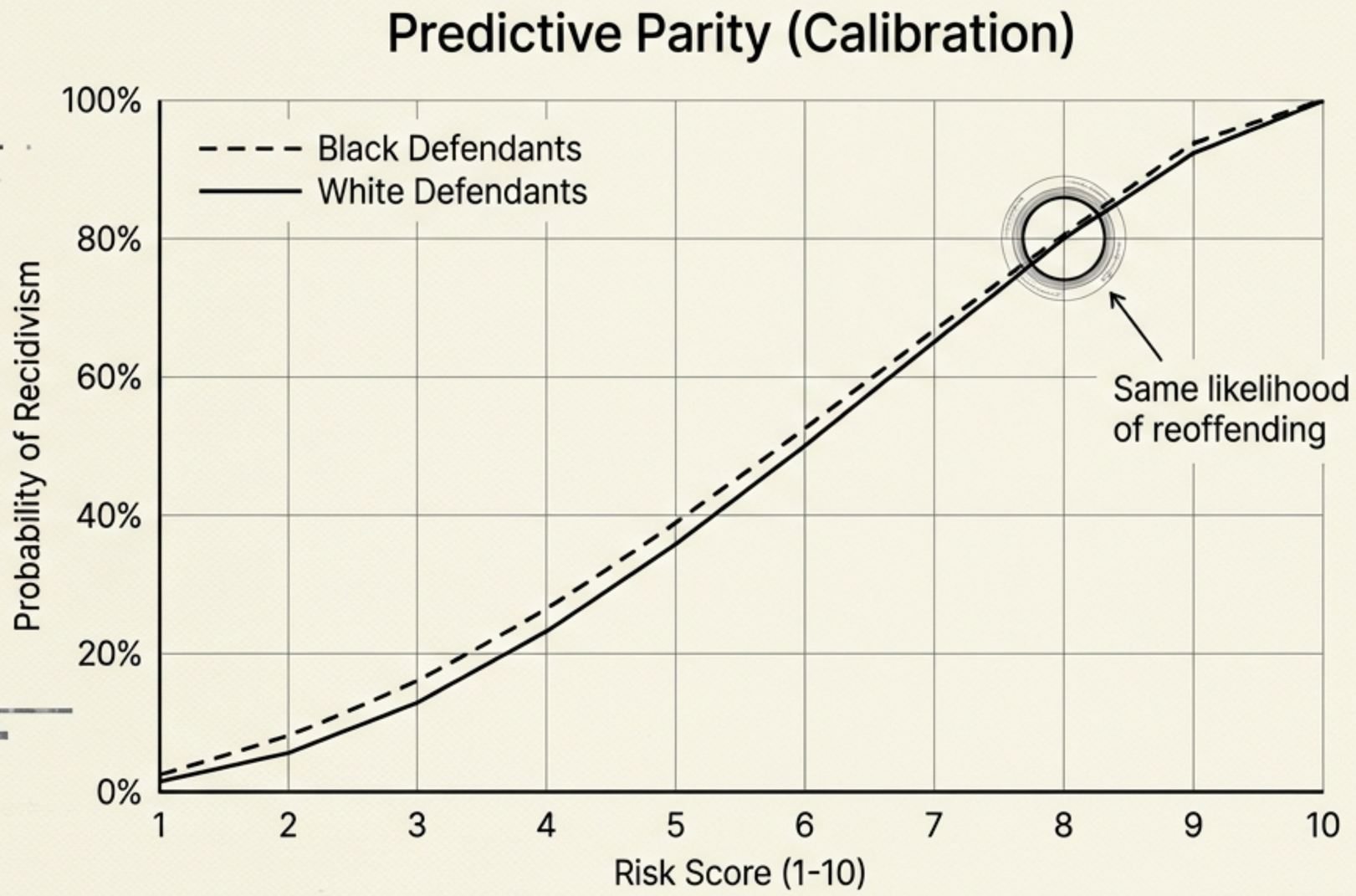


False Negative Rate
(Risky labelled as Safe)

Black Defendants:	**28%**
White Defendants:	**47%**

ProPublica's definition of fairness is "Equalized Odds" (Error Rate Balance). Their findings show that Black defendants are nearly twice as likely to be wrongly flagged as risky, while the algorithm is significantly more "forgiving" to White defendants.

A score of ‘8’ means the same probability regardless of race



The Defence:

Northpointe argues the model has “accuracy equity.” If a Black defendant and a White defendant both score an 8, they have the exact same likelihood of recidivism based on the data provided.

Within the model’s own logic, it treats both groups “equally”.

The Impossibility Theorem ensures a permanent mathematical conflict

IF (Base Rates ≠ Equal) **AND** (Accuracy < 100%) **THEN** (Fairness A ≠ Fairness B)

It is mathematically impossible to satisfy both definitions simultaneously because 'ground truth' data shows higher arrest rates for Black communities (driven by systemic factors). Both parties are technically 'right' within their own silos. The conflict is a mathematical inevitability.

When statistics become sentences, the benefit of the doubt is lost

The Aggravating Factor

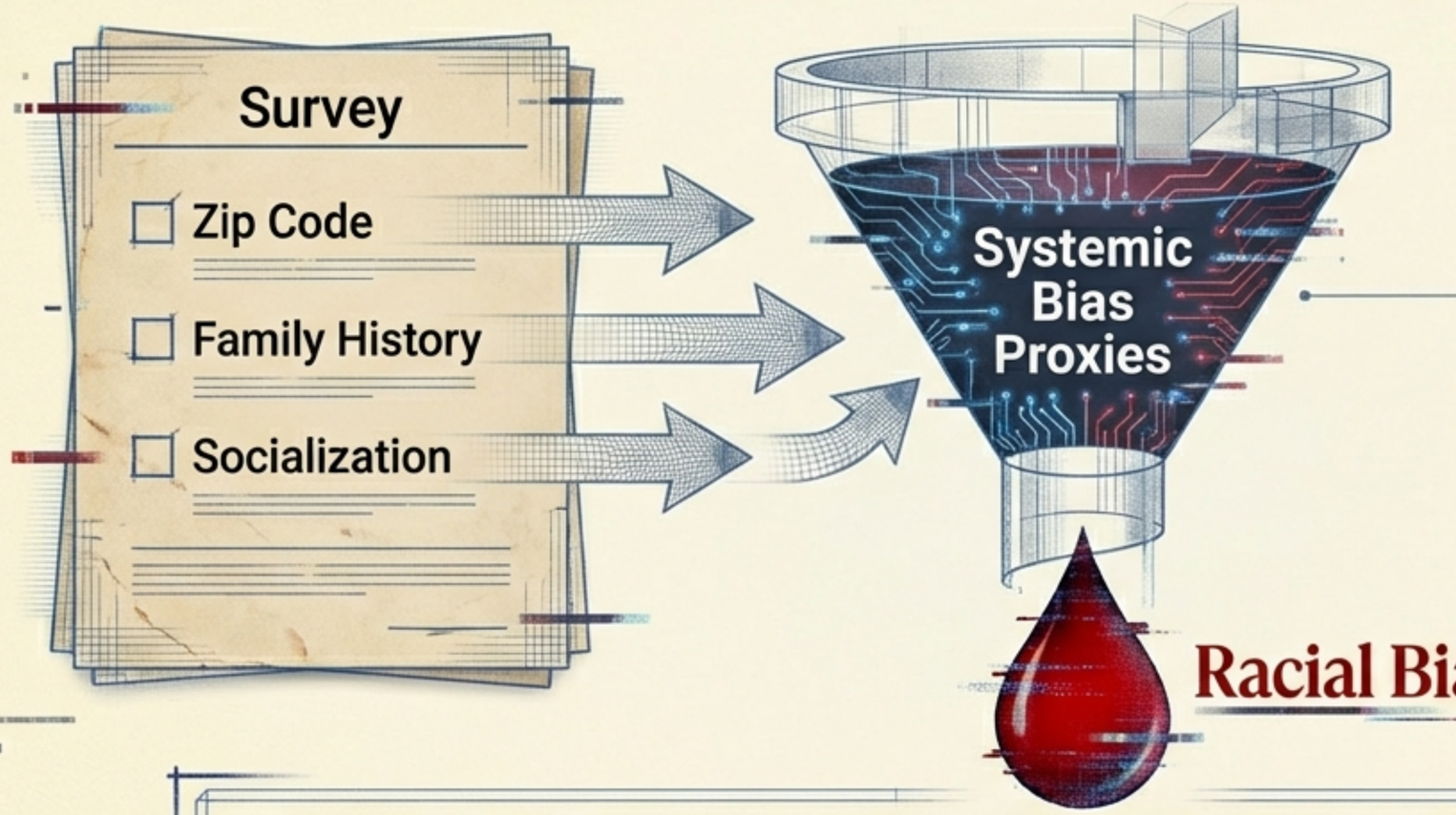
In cases like State v. Gauthier and State v. Loomis, a ‘High Risk’ label justifies higher bond amounts, denied parole, and extended incarceration.

The Cost

A 45% false positive rate means nearly half of Black defendants flagged as risky would have remained law-abiding but were stripped of liberty based on a flawed prediction.

> “The ‘benefit of the doubt’ is distributed unevenly.”

The well is poisoned by proxy variables before the code is even written



1. **Zip Codes:** Proxies for segregated, heavily policed neighbourhoods.
2. **Family/Friend Victim History:** Captures the nature of systemic racism rather than individual intent.
3. **Socialization:** Penalises those facing socioeconomic disruption.

Key Insight: If the training data is a mirror of historical oppression, the algorithm becomes a high-speed engine for reinforcing it.

Due process cannot exist inside a proprietary Black Box

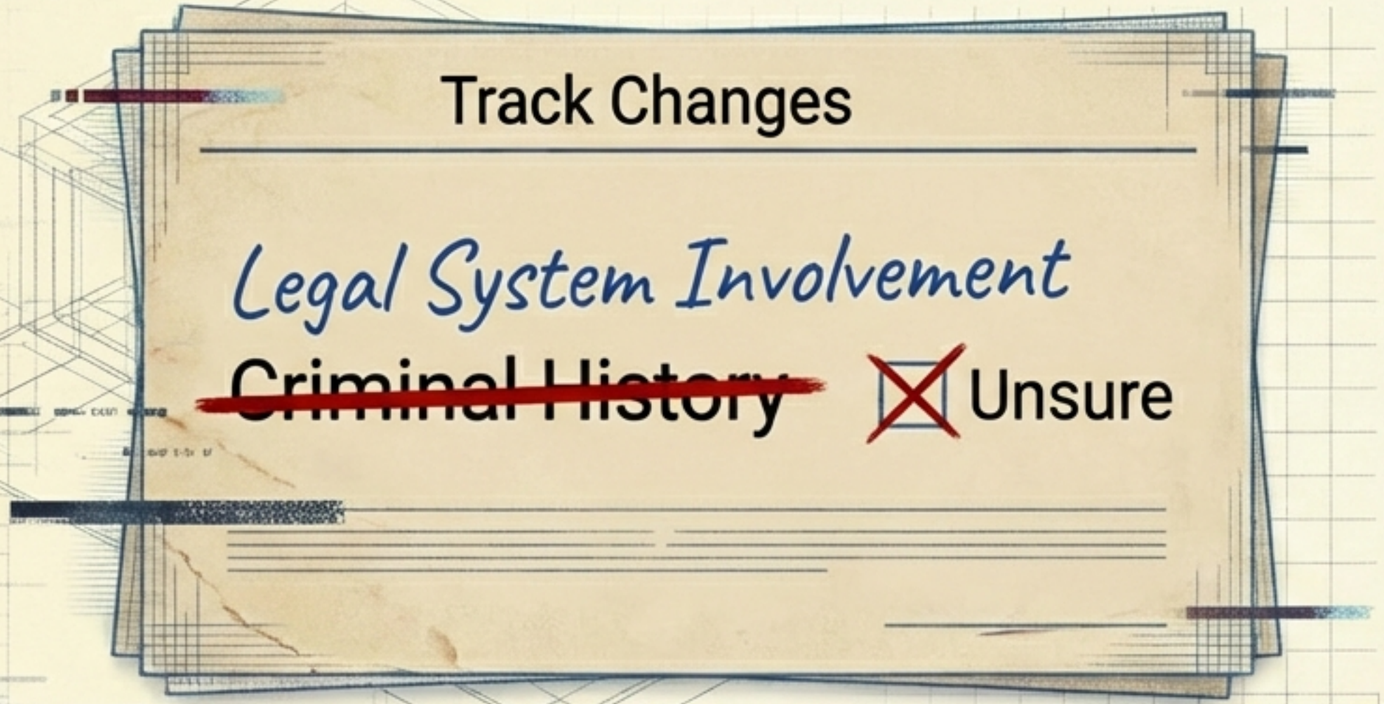
Case Study: State v. Loomis

The Wisconsin Supreme Court allowed the use of COMPAS despite its secret code, requiring only "disclaimers" as a warning.

The Issue: Trade secret protections prevent defendants from challenging the logic used to imprison them.

Should a private company's intellectual property rights trump a human being's right to liberty?

Sanitising the language does not remove the bias



Attempted Intervention:

- **Neutral Language**: Renaming terms to to sound less accusatory.

- **Data Cleaning**: Removing 'Unsure' options to force binary responses.

- **Experiments**: Testing the removal of low-level offences like marijuana possession.

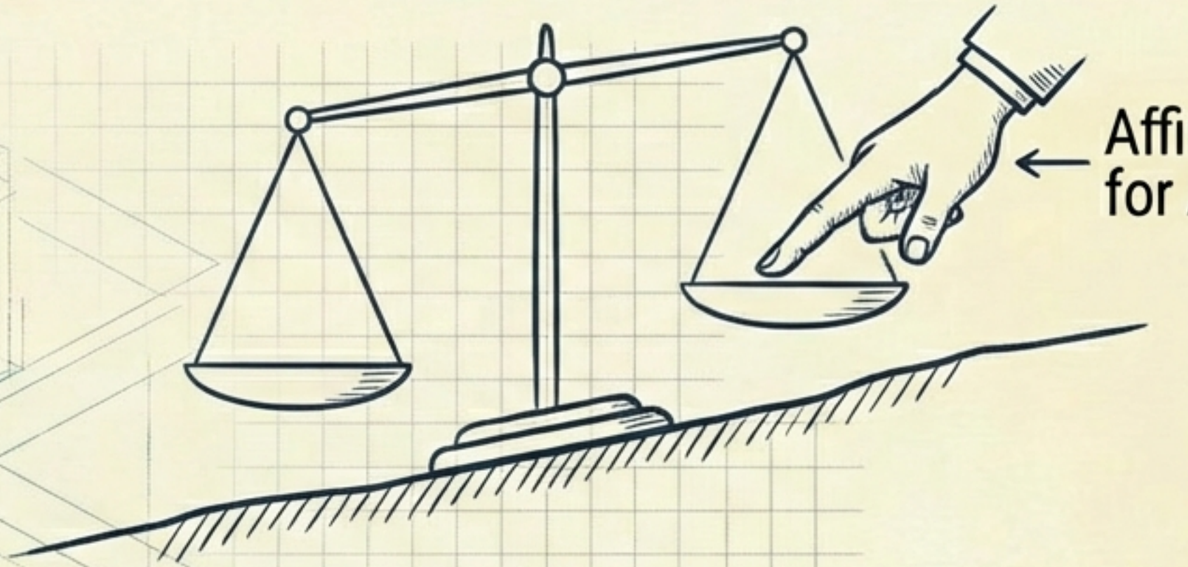
Verdict: These are surface-level adjustments to a structural problem. They do not solve the ‘base rate’ disparity.

Technical interventions force a trade-off between fairness and accuracy

Reweighing	Prejudice Remover	Calibrated Equalized Odds
Aligns metrics but ignores bias in initial data collection.	Enforces independence from race. Result: Accuracy plummeted to 38% (worse than a coin flip).	Reduces bias but increases false negatives, potentially endangering public safety.

Insight: You can technically ‘debias’ the machine, but usually at the cost of breaking its utility.

If the system is tilted, the algorithm must be tilted back



A Radical Proposal: Using race as a ‘plus factor’ to lower risk scores for unprivileged groups.

The Logic: Since ‘priors’ and ‘residential stability’ are heavily weighted against Black defendants due to systemic history, the only way to achieve true equity is to mathematically handicap the score in the opposite direction.

Group Fairness often comes at the expense of Individual Fairness.

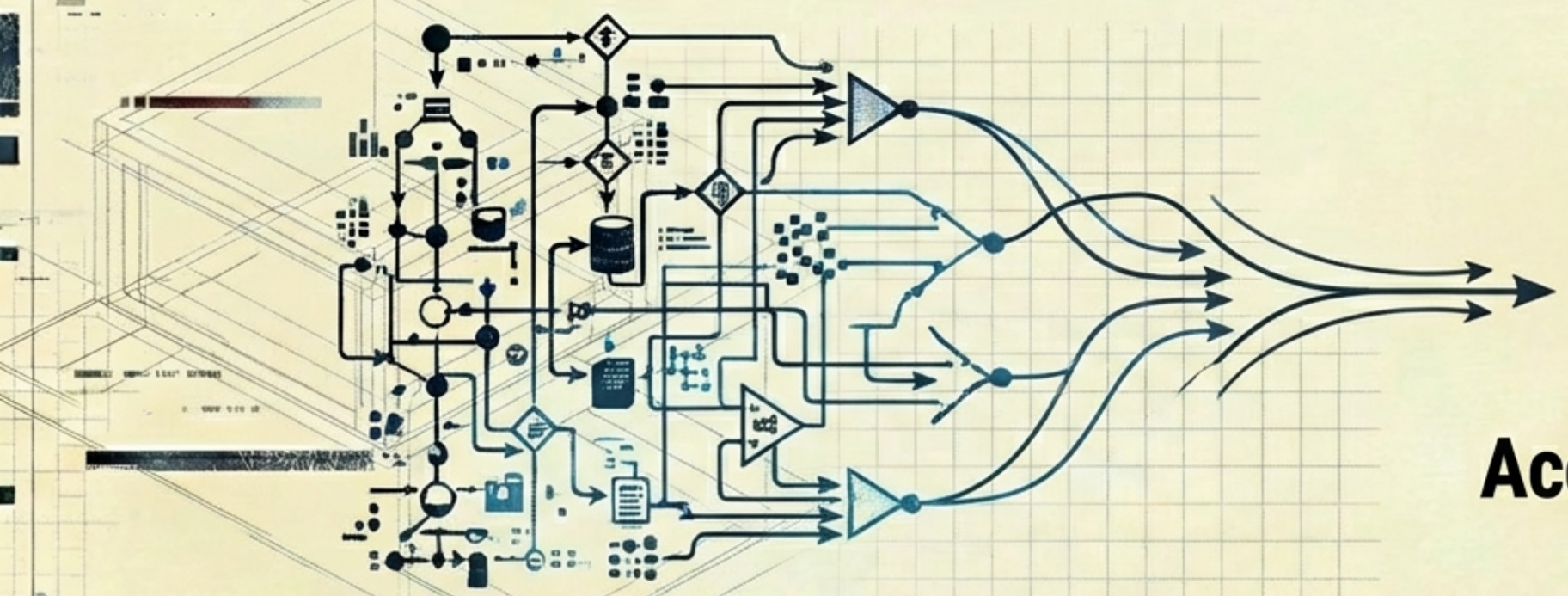
Exporting these models acts as 'Technological Solutionism'

Residential Stability: In informal economies, moving frequently is a survival strategy, not a risk factor.

Vocational History: In conflict zones, gaps in employment are not character flaws.

The Danger: Importing a model trained on Western 'Socialization History' into a post-colonial context treats victims of historical oppression as high-risk criminals.

We are outsourcing moral responsibility to a coin flip



Accuracy ≈ 65%

1. If accuracy is near a coin flip, why use it to determine prison time?
2. Who in a democracy gets to decide which fairness definition ‘counts’?
3. How can we call it justice if the evidence is a ‘trade secret’?

Human accountability must remain the final check on human freedom

Conclusion: Technical interventions only treat symptoms. We cannot ‘fix’ justice with code while ignoring the over-policing and over-criminalization that produce the poisoned data.

We cannot allow the Black Box to be the final word. The ‘benefit of the doubt’ is a human concept, and it requires human conscience to administer it.