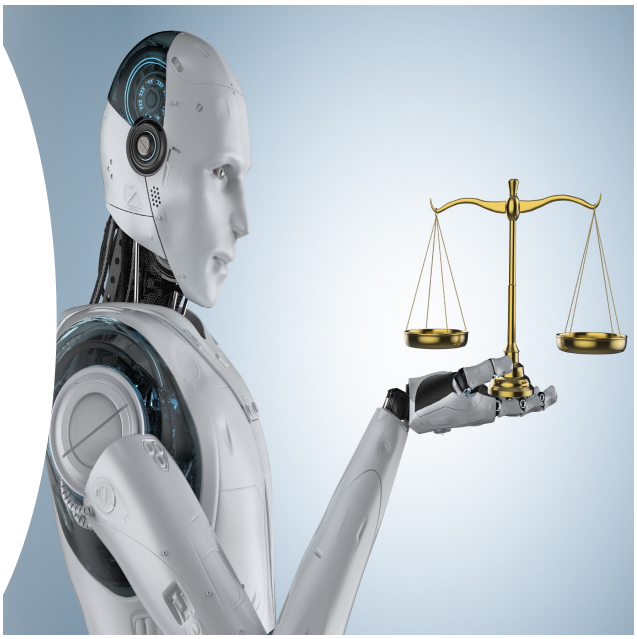
When AI Commands the Bench: Automation Bias and the Future of Judicial Independence

By
Dr Uchenna Nnawuchi
Dr Carlisle George

ALERT Research Group
Middlesex University, London (UK)



1

Application of AI (Machine Learning) in Criminal Justice

(Israni, 2017)
State of Wisconsin v. Eric L. Loomis 2015AP157

Bail decisions **Parole decisions** **Sentencing** **Probation and Supervision**

Rauber et al (2020)

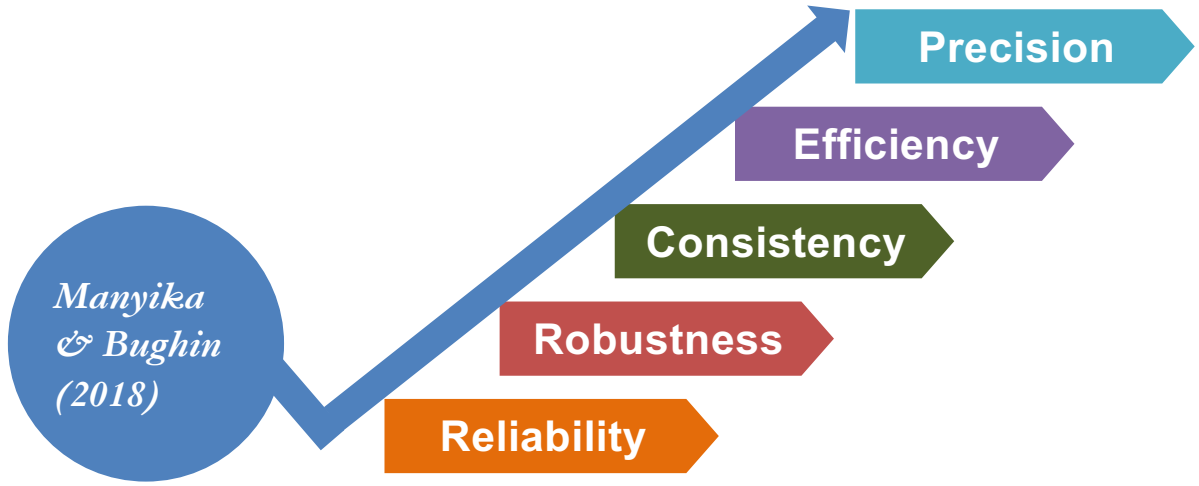
2

Common AI Risk Assessment Tools - CJS

- **UK - OASys (Offender Assessment System):** Widely used in the UK to assess risk of harm and re-offending. It incorporates AI techniques to profile offenders, influencing bail, sentencing, and parole decisions. Assesses over **250,000** people per year.
- **UK - OGRS (Offender Group Re-Conviction Scale):** AI predictive tool used in the UK to estimate the likelihood of re-offending within two years.
- **UK - HART (Harm Assessment Risk Tool):** Used by some police forces in the UK for predicting re-offending risk.
- **USA - COMPAS (Correctional Offender Management Profiling for Alternative Sanctions):** Common in the U.S. for assessing recidivism risk.

3

Some Benefits of AI (Machine Learning)



4

Some Challenges of ML in Criminal Justice

ISSUES	REASON	EFFECTS
Bias	• Proxies of historical data, Race, Geography, Economic Status	• False Positives & False Negatives
Accuracy	• Veracity of training data	
Opacity	• Black Box effect, Neural Networks	• Trade secrets • Undermines due process • Inability to challenge decision-making • Undermines fairness and transparency

Fortes (2020)

5

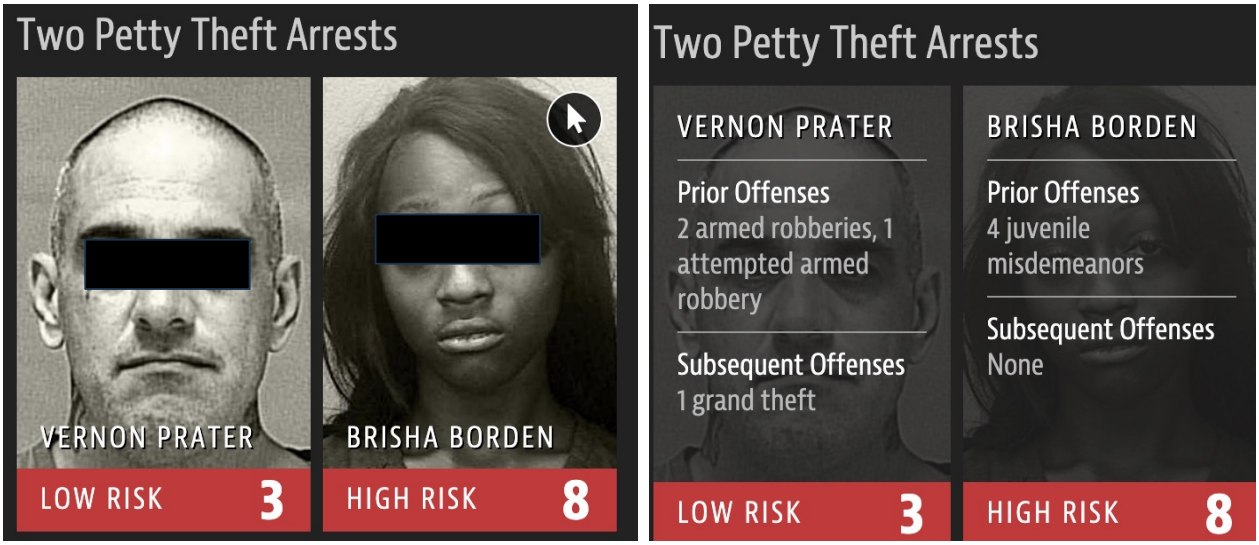
Case study - ProPublica Study of Machine Bias

Angwin et al (2016)

- Focus on algorithms used in the USA CJS for predicting risk assessment used by judges in criminal sentencing.
- Software developed by a for-profit company "Northpointe" - among the most widely used assessment tools in the USA
- Risk scores analysed for 7,000 people arrested in Broward County, Florida (2014-2016) showed that algorithms were unreliable in regard to forecasting violent crime. **Only 20% accuracy** (over the next 2 years).
- Algorithms were **likely to falsely flag black defendants** as future criminals, **wrongly labelling them** this way at almost **twice the rate as white defendants**.
- White defendants were **mislabelled as low risk** more often than black defendants.

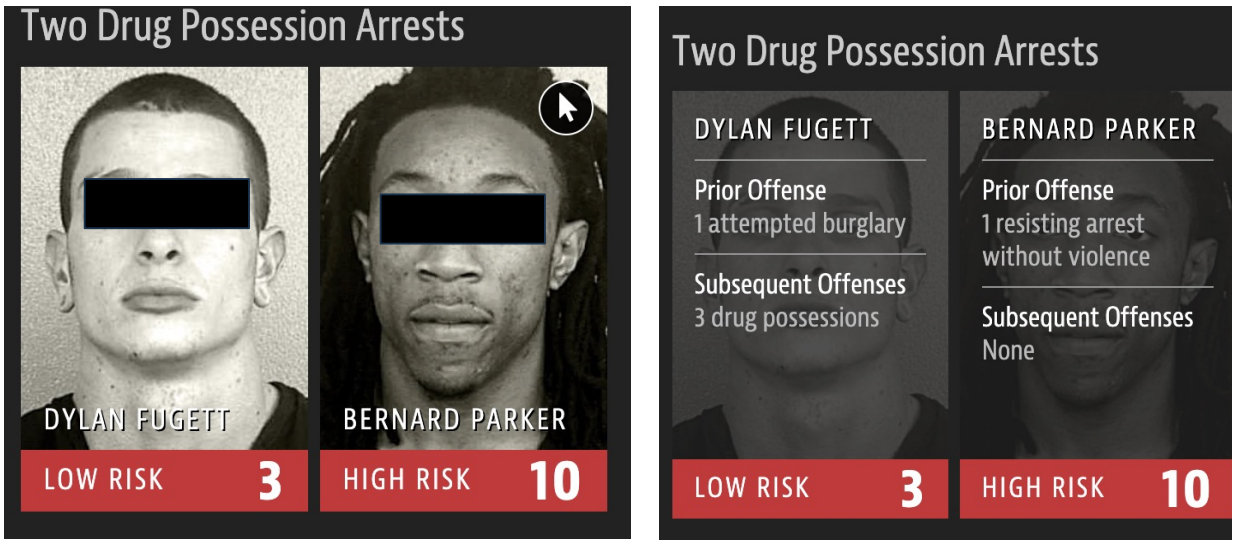
6

ProPublica Study of Machine Bias – Florida (USA) – 1
Angwin et al (2016)



7

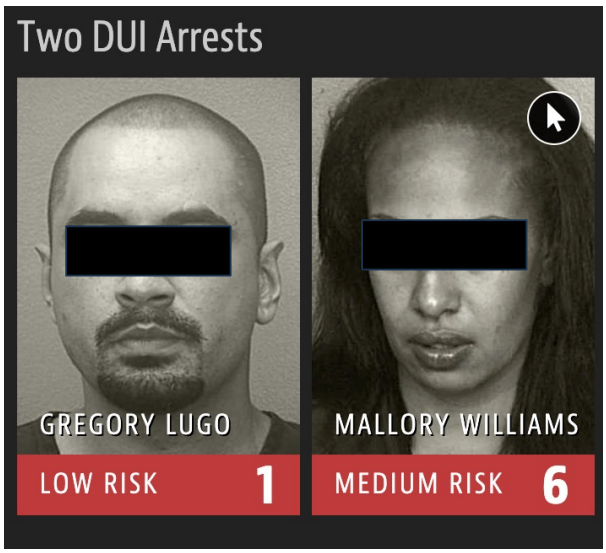
ProPublica Study of Machine Bias – Florida (USA) – 2
Angwin et al (2016)



8

ProPublica Study of Machine Bias – Florida (USA) – 3

Angwin et al (2016)



9

ProPublica Study of Machine Bias – Florida (USA) – 4

Angwin et al (2016)

10

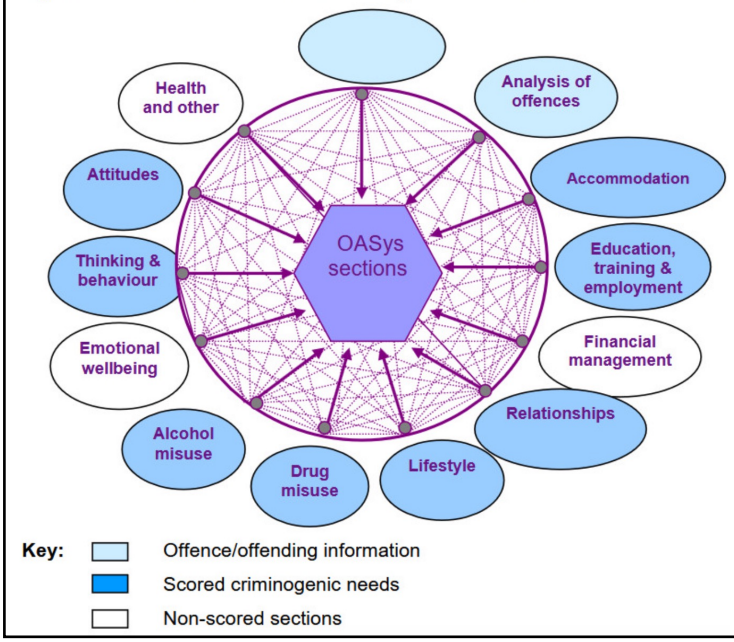
ProPublica Study of Machine Bias – Explanations?

- Could the disparity between results for black vs white defendants be explained by defendants' prior crimes or the type of crimes they were arrested for? **The study concluded "No".**
- A statistical test was run that isolated the effect of race from criminal history and recidivism, as well as from defendants' age and gender.
  - Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind.
- Basis of the future crime formula is 137 questions.
  - **Race is not one of the questions.**
  - It includes factors such as education levels, whether a defendant has a job, whether parents were previously in jail, whether friends/acquaintances are taking drugs illegally, frequency of getting in fights while at school etc.

11

UK - The Offender Assessment System (OASys) – 1

(Statewatch, 2025)



OASys (**AI system**) in use by the Home Office since 2001 to 'predict' the risk of re-offending of thousands of prisoners and people on probation every week.

Over 1,300 people profiled daily

Risk scores used to decide:

- bail;
- sentencing;
- the type of prison allocated;
- access to education & rehabilitation programmes.

12

UK - The Offender Assessment System (OASys) – 2

(Statewatch, 2025)

- Official evaluation of OASys found discrepancies in accuracy based on gender, age and ethnicity.
- OASys scores were disproportionately less accurate (lower predictive validity) for racialised people than white people, and especially so for Black and mixed-race people. Among all offenders, actual (proven) reoffending was significantly below the predicted rate.
- Minorities reported inaccurate and false information entered into the system about them (e.g. gang link). (Suspected bias and discrimination, carelessness, or insufficient training)
- Structural racism and other forms of systemic bias may be coded into OASys risk scores—both directly and indirectly. Data entered may be influenced by biased policing and over-surveillance of certain communities.
- Higher risk scores = harsher decisions on sentencing, bail, categorisation, release.
- Despite serious concerns for equality and discrimination, fair trial rights, accuracy, and opportunities for redress, the Ministry of Justice continues to use OASys assessments across the prison and probation services. To be replaced by new system ARNS (Assess Risks, Needs, and Strengths) this year (2026) but bias and discrimination issues not addressed in the new system.

13

Concerns

- Judges are usually unable to understand the rationale behind algorithmic decisions (whose inner workings remain inaccessible and unchallengeable)
- Judges increasingly rely on risk assessment tools in sentencing, making human judgement subordinate to algorithmic decisions.
- Difficult to contest or question algorithmic decisions (even if they often lack transparency).
- Systems with poor predictive validity and bias pose grave risks to fundamental human rights, particularly the right to liberty.
- Erosion of constitutional and ethical foundations of adjudication. Lack of fairness and contestability.

14

Some Solutions

- Ensure AI-Assisted Outcomes Remain Fully Contestable and Reviewable.
- Use Interpretable models for high-risk systems or in critical domains to ensure explainability.
- Mandate AI Transparency Measures to Preserve Fairness and Accountability.
- Mandate that Courts' decisions must be independent of AI systems' recommendations.
- Mandate More Judicial Training and Awareness Regarding AI systems
- Mandate Regular Auditing of AI Systems to Identify Problems and Assess Predictive Validity.

15

References

- Israni (2017). Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis. https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1
- Angwin et al (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks' (ProPublica, 23 May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Andreas Rauber, Roberto Trasarti, and Fosca Giannotti (2020), 'Transparency in Algorithm Decision Making', Transparency in Algorithmic Decision Making - Introduction to the Special Theme (ercim.eu)
- Rubim Borges Fortes P. Paths to Digital Justice: Judicial Robots, Algorithmic Decision-Making, and Due Process. *Asian Journal of Law and Society*. 2020;7(3):453-469. doi:10.1017/als.2020.12
- Statewatch (2025). UK: Over 1,300 people profiled daily by Ministry of Justice AI system to 'predict' re-offending risk https://www.statewatch.org/news/2025/april/uk-over-1-300-people-profiled-daily-by-ministry-of-justice-ai-system-to-predict-re-offending-risk/
- Nnawuchi, U., George, C. Decoding accountability: the importance of explainability in liability frameworks for smart border systems. *Discov Computing* 28, 64 (2025). https://doi.org/10.1007/s10791-025-09559-5

16