Annotated Timeline , Information Manipulation

Munongedzi Mabhoko, Clarkson University, mabhokm@clarkson.edu

1. Early 2000s : Risk Assessment Becomes the Administrative Default

Event. Criminal-justice agencies begin incorporating actuarial risk instruments to categorize individuals by their likelihood of reoffending. These tools emerge as a response to overcrowded prisons, budget constraints, and a desire for more standardized decision-making.

This shift marks the foundational step toward algorithmic governance in sentencing and parole. Although promoted as neutral, these tools rely on historical criminal-justice data that already reflect disproportionate surveillance of marginalized communities. The very idea of "risk" is built atop structural inequalities, meaning bias enters the system before algorithms are even introduced. At this stage, opacity arises not from code but from untreated assumptions: that past enforcement patterns can be treated as objective predictors of future behavior.[8, 9]

2. 2009 COMPAS Positioned as a Modern, Data-Driven Solution

Event. COMPAS is formally validated in academic literature as a next-generation risk assessment tool integrating criminal history with behavioral and psychosocial indicators.

Its public framing emphasizes technical rigor, but its internal mechanics remain proprietary. Even official practitioner documentation provides only high-level descriptions of how COMPAS organizes risk domains and decile conversions, without revealing the computational pathways connecting individual inputs to final scores. This marks the institutionalization of intentional opacity, where crucial details about feature weighting, interactions, and calibration are inaccessible to the legal actors who depend on them. [8, 10]

3. 2010 Risk Identified as a Racial Proxy

Event. Scholars highlight that even when an algorithm excludes race explicitly, the variables used (such as prior arrests, neighborhood instability, or employment patterns) act as indirect encodings of racialized exposure to policing.

This insight demonstrates that formal neutrality does not imply functional neutrality. A system can avoid using race as a field while still reproducing racially patterned outcomes because the data environment itself is unequal. This represents the first major recognition that algorithmic bias emerges from structural conditions rather than malicious design.[9]

4. 2013–2014:COMPAS Adopted in High-Stakes Pretrial Decision-Making

Event. Local jurisdictions begin relying on COMPAS to inform bail decisions and detention recommendations soon after arrest.

These assessments rely on a 1–10 decile scale that simplifies multidimensional risk indicators into categorical recommendations, but the mapping from questionnaire responses to risk classes remains undisclosed. Even publicly available FAQ and technical overview materials describe the scoring structure without exposing the underlying logic, reinforcing procedural opacity at the point where liberty decisions are made. [3, 11]

5. 2016 (Early) :Explainability Research Intensifies (LIME Introduced)

Event. Explanatory frameworks designed to make black-box predictions more intelligible emerge, emphasizing localized interpretability.

This represents a methodological response to growing discomfort with opaque models. The field begins recognizing that high-performance systems cannot remain inscrutable when they inform decisions about punishment and release. Explainability research underscores that complex models often rely on unpredictable or socially unacceptable features unless constrained. This stage exposes deeper layers of opacity: even when developers aim for fairness, nontransparent interactions between variables can produce unintuitive, sometimes harmful decision boundaries.[2]

6. 2016 (May): ProPublica Reveals Systematic Racial Disparities in COMPAS

Event. Analyses of thousands of real-world cases reveal that Black defendants are far more likely to be incorrectly flagged as “high risk,” while white defendants are more frequently misclassified in the opposite direction.

The public learns that accuracy alone is a misleading metric. Two groups may exhibit similar overall accuracy while experiencing drastically different error distributions. Since false positives translate into detention, higher bail, or longer supervision, the burden of algorithmic uncertainty falls disproportionately on communities already over-policed.

The investigation highlights that proprietary systems shield not only their algorithms but also their embedded value judgments from democratic oversight.[3]

7. 2016 (July): Judicial Acceptance of Black-Box Decision Tools (State v. Loomis)

Event. The Wisconsin Supreme Court upholds the use of COMPAS in sentencing, while acknowledging the defendant cannot examine how the score was derived.
This decision normalizes a model of justice where individuals can be punished based on evidence they are not permitted to evaluate an inversion of traditional due process norms. The ruling exemplifies how opacity becomes codified in legal practice. The judiciary treats algorithmic assessments as authoritative despite lacking epistemic access to their internal consistency, fairness tradeoffs, or error patterns. This marks the consolidation of institutional opacity where the state endorses proprietary algorithms as legitimate arbiters of human liberty.[12]

8. 2016: Theoretical Foundations of Fairness Trade-offs Established

Event. Multiple research teams independently demonstrate that widely accepted fairness goals such as equal accuracy across groups and equal error rates cannot all be satisfied when underlying crime or arrest rates differ.
This constitutes one of the most important conceptual breakthroughs in algorithmic ethics. It formalizes that fairness is not a singular property but a landscape of competing criteria. When base rates differ, insisting on one fairness metric mathematically requires sacrificing another. Bias therefore cannot be eliminated through technical optimization alone; it stems from social conditions reflected in the data. These results force policymakers to confront a central ethical question: Which group should bear the cost of prediction error?[5, 6, 7]

9. 2016–2017 : Empirical and Legal Scholars Reevaluate Validity and Governance

Event. Academic critiques refine analyses of COMPAS, disentangling issues of accuracy, calibration, and disparate impact. Legal scholars scrutinize the compatibility of opaque algorithms with constitutional protections.
The debate shifts from whether algorithms “work” to whether they are compatible with democratic accountability. Scholars argue that when systems exert coercive power, transparency, contestability, and procedural fairness are not optional, they are structural

prerequisites. This period also marks a shift toward recognizing epistemic inequality, developers and vendors possess full visibility into the system, while those subjected to its judgments have none.[13, 14]

10. 2017–2020: Deeper Understanding of Systemic Effects

Event. Analyses highlight how risk scores influence not just one decision (such as bail) but entire trajectories: plea negotiations, supervision conditions, treatment eligibility, and parole.

Risk assessments are shown to act as amplifiers of initial disparities. A misclassification at the start of the process can trigger cascading disadvantages. This demonstrates that algorithmic bias is not an isolated outcome but a systemic multiplier of inequalities. The focus expands from algorithmic error to structural harm, underscoring the need for governance frameworks that consider long-term, compounding consequences. [15]

11. 2020–Present: Algorithmic Governance Emerges as a Global Human-Rights Issue

Event. International bodies, civil-society organizations, and policy institutes frame algorithmic risk scoring as part of broader concerns involving AI accountability, discrimination, and due process.

This stage recognizes that algorithmic opacity and structural bias challenge fundamental human-rights principles: equality before the law, non-discrimination, and the right to understand and contest evidence used against one's liberty. Calls for transparency shift from technical suggestions to demands for legal oversight, independent audits, open documentation, and mechanisms that permit individuals to challenge algorithmic findings. The issue becomes part of the global conversation about ethical AI and justice reform. [4, 1, 14]

How Bias and Opacity Enter at Every Layer

1. Data Bias

Historical records reflect unequal policing. Any model trained on such data reproduces these patterns. [8, 9]

2. Feature and Label Choices

Selecting variables tied to socioeconomic disadvantage embeds systemic inequalities into the model. [1, 8]

3. Fairness Criterion Selection
Choosing one fairness metric over another inherently prioritizes certain groups' interests [5, 6, 7]

4. Proprietary Design
Trade-secret protections prevent scrutiny of assumptions and flaw detection. [1, 12]

5. Institutional Use
When opaque tools are deployed in bail or sentencing, informational asymmetry becomes coercive and harms accumulate across legal processes.[3, 14, 15]

6. System Effects
Initial misclassifications propagate through a person's entire justice-system trajectory, magnifying harm. [15]

7. Human-Rights Implications
Opacity undermines due process, fairness, and public trust in decision-making systems.[4, 14]

References

[1] Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data & Society.SAGE Journals+1

[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD / NAACL demo.arXiv+1

[3] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.ProPublica+1

[4] Angwin, J., & Larson, J. (2016, Dec 30). Bias in criminal risk scores is mathematically inevitable, researchers say. ProPublica.ProPublica+1

[5] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.arXiv+1

[6] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807.arXiv+1

[7] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. NeurIPS 29, 3315–3323.arXiv+1

[8] Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behavior, 36(1), 21–40.SAGE Journals+1

[9] Harcourt, B. E. (2010). Risk as a proxy for race: The dangers of risk assessment. University of Chicago Public Law & Legal Theory Working Paper No. 323.Chicago Unbound+1

[10] Equivant (Northpointe). (2017). Practitioner's Guide to COMPAS Core.cjdata.tooltrack.org+1

[11] Equivant (Northpointe). COMPAS FAQ / COMPAS Core Risk & Need Assessment System. (Various technical and FAQ documents detailing decile scoring and risk categories.)Scribd+1

[12] State v. Loomis, 881 N.W.2d 749 (Wis. 2016); cert. denied, 137 S. Ct. 2290 (2017).Harvard Law Review+1

[13] Flores, A. W., Lowenkamp, C. T., & Bechtel, K. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine Bias." (Crime and Delinquency / technical rejoinder).United States Courts+1

[14] Washington, A. L. (2019). Lessons from the COMPAS–ProPublica debate. Colorado Technology Law Journal, 17(1).SCIRP+1

[15] O'Brien, T. (2021). The cascading effect of algorithmic bias in risk assessment. Georgetown Journal on Poverty Law & Policy (or related criminal-justice review).Georgetown Law