

BEYOND HUMAN JUDGMENT: ARTIFICIAL INTELLIGENCE, ALGORITHMIC EVIDENCE, AND FAIR TRIAL GUARANTEES IN CRIMINAL JUSTICE

Prof. Jai Shankar Singh^{*}
Prof. Ishita Chatterjee^{**}
Alben Jicho T.^{***}

ABSTRACT

The integration of artificial intelligence into criminal justice systems has fundamentally altered the landscape of evidence generation, procedural decision-making, and adjudication. From predictive policing algorithms and facial recognition technology to risk-assessment tools used in bail and sentencing, AI-driven mechanisms increasingly shape outcomes once reserved exclusively for human judicial reasoning. This paper critically examines the implications of such technological integration for the constitutional guarantee of a fair trial, with particular emphasis on the admissibility, reliability and contestability of algorithmically generated evidence. It interrogates whether existing evidentiary standards, developed for human testimony and conventional forensic method, are adequate to address the epistemic opacity, or black box nature, of machine-learning system, and explores procedural due process concern including the right to confront and challenge evidence, the burden of algorithmic accountability, and the risk of automation bias among judicial officer. Drawing on comparative jurisprudence from the United States, the European Union and India, the paper analyses emerging regulatory response, including the EU Artificial Intelligence Act and doctrinal development concerning explainability requirement, and argues that the unchecked deployment of AI in evidentiary and procedural context risks eroding foundational fair trial protection unless accompanied by robust frameworks for algorithmic auditability, contestability and judicial oversight. It proposes a normative framework, termed algorithmic due process, resting on mandatory disclosure of algorithmic logic, independent technical review, and judicial AI literacy, intended to reconcile technological efficiency with constitutional safeguard.

KEYWORDS: Artificial Intelligence; Algorithmic Evidence; Fair Trial; Due Process; Automation Bias.

1. INTRODUCTION

Criminal justice has always depended on inference: from a bloodstain to a wound, from a fingerprint to a suspect, from testimony to guilt. What has changed in the past decade is not the reliance on inference itself but the identity of the inferential agent. Increasingly, the entity drawing the inference is not a forensic examiner, a police officer, or a judge, but a statistical model trained on historical data and deployed at scale across police department, prosecutor's office, pre-trial service, and sentencing court.¹ Predictive policing platform allocate patrol resource on the basis of historical crime data; facial recognition system match probe image against database containing millions of face to

^{*} Vice Chancellor, RGNUL, Patiala.

^{**} Professor and Dean, Faculty of Law, Marwadi University, Rajkot.

^{***} Research Scholar, SRMIST, Chennai.

¹ Sarah Brayne, Predict and Surveil: Data, Discretion, and the Future of Policing 12-18 (2020).

generate investigative lead; and actuarial risk-assessment instrument produce numerical score informing decision on bail, probation and sentence length. Each of these technology performs a function once the exclusive province of human judgment, operating inside a criminal process designed, doctrinally and institutionally, around the assumption that the person or mechanism generating evidence could be identified, examined, and, where necessary, discredited.

This paper asks a comparatively narrow but consequential question: can the fair trial guarantee, as elaborated in constitutional and human rights jurisprudence, accommodate an evidentiary and procedural apparatus that is substantially non-human, statistically opaque, and commercially proprietary? The question is narrow because it does not ask whether AI should be used in criminal justice at all, a debate addressed extensively elsewhere; it is consequential because the answer determines whether existing doctrine of admissibility, confrontation and reasoned adjudication can simply be extended to algorithmic tool, or whether an entirely new doctrinal vocabulary, termed here algorithmic due process, is required.

The stake of this inquiry are illustrated starkly by the case of Robert Williams, a Michigan resident wrongfully arrested in 2020 after a facial recognition match wrongly identified him as a shoplifting suspect, a mismatch never independently corroborated before arrest.² They are illustrated equally by the Wisconsin Supreme Court's decision in *State v. Loomis*³, in which a defendant sentenced partly on the basis of a proprietary recidivism score was denied access to the logic underlying that score on the ground that it constituted a trade secret belonging to a private vendor. In both instance, an evidentiary or dispositional outcome traceable to an algorithm intersected with a constitutional guarantee, the right to a fair hearing, that presupposes the possibility of meaningful contestation. The central claim this paper advances is that such contestation becomes structurally difficult, and in some instance impossible, where the reasoning process generating the evidence or risk score is not merely complex but deliberately or inherently opaque.

The paper proceeds in seven further part. Section 2 maps the principal category of AI deployment across the criminal process. Section 3 examines the doctrinal architecture of the fair trial guarantee in the United States, the European Union and Council of Europe system, and India. Section 4 interrogates the black box problem and the specific doctrinal battleground of admissibility and contestability. Section 5 examines automation bias and its implication for independent judicial reasoning. Section 6 surveys comparative regulatory response and three illustrative case study. Section 7 develops the paper's central normative contribution, algorithmic due process, and addresses

² Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. Times (June 24, 2020).

³ *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016).

the framework's own limitation. Section 8 translates that framework into institutional recommendation and concludes.

Two clarification are necessary at the outset. First, this paper does not treat artificial intelligence as a single undifferentiated technology; the due process implication of a rule-based risk matrix differ meaningfully from those of a deep neural network trained on unstructured data, and the analysis attends to those difference where they affect the doctrinal argument. Second, the comparative method is deliberately asymmetric: it treats the United States as the jurisdiction with the most extensive litigation history on algorithmic evidence, the European Union and Council of Europe system as the jurisdiction with the most developed regulatory response, and India as a jurisdiction whose constitutional due process jurisprudence is doctrinally rich but whose regulatory response to AI in criminal justice remains at an early stage.

2. THE PROLIFERATION OF ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE

Artificial intelligence has entered the criminal process at every stage, from the earliest moment of investigation to the final calculation of a custodial sentence. This section maps four principal category of deployment before turning, in later sections, to the doctrinal consequence of that deployment.

Predictive policing system, exemplified by platform such as PredPol (later Geolitica) and the Chicago Police Department's Strategic Subject List, use historical crime and, in some iteration, social-network data to forecast where crime is likely to occur or which individual are statistically associated with elevated risk of violence.⁴ Because the training data reflect historical policing pattern rather than the true underlying distribution of criminal activity, the model risk reproducing and amplifying pre-existing pattern of over-policing in historically over-surveilled neighbourhood, a phenomenon the technical literature terms a feedback loop; several jurisdiction, Los Angeles among them, have since discontinued such programme following independent audit questioning both their predictive validity and disparate impact.

Facial recognition technology, deployed to compare a probe image, often drawn from closed-circuit television footage of uncertain quality, against a reference database, returns a ranked list of probable match rather than a definitive identification. The gap between what the technology produces, a probabilistic candidate list, and what it is frequently treated as producing by investigating officer, a positive identification, has generated a documented series of wrongful arrest, of which Robert Williams's case is the most widely

⁴ Rashida Richardson, Jason Schultz & Kate Crawford, *Dirty Data, Bad Predictions*, 94 N.Y.U. L. Rev. Online 192 (2019).

reported but not the only instance. Independent testing by the United States National Institute of Standards and Technology has documented demographic differential in false-positive rate across several commercially deployed algorithm, with certain system exhibiting materially higher false-positive rate for Black and Asian face than for white face under equivalent condition.⁵ In the United Kingdom, the Court of Appeal in *Bridges v. Chief Constable of South Wales Police*⁶ held that live automated facial recognition had not been conducted in accordance with a sufficient legal framework, in part because the force had not taken reasonable step to satisfy itself that the software carried no unacceptable bias on ground of race or sex.

Risk-assessment instrument, including COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) and the Public Safety Assessment, generate a numerical score purporting to represent the statistical likelihood that a defendant will reoffend, fail to appear, or engage in violent conduct if released, and these score inform decision on pre-trial detention, bail condition, probation eligibility, and, in some jurisdiction, sentence length. An influential ProPublica investigation published in 2016 concluded that COMPAS, as deployed in Broward County, Florida, was substantially more likely to falsely flag Black defendant as high-risk and white defendant as low-risk, a finding generating considerable methodological controversy but crystallising, for many scholar, the underlying tension between statistical conception of fairness such instrument cannot simultaneously satisfy.⁷

A fourth category, probabilistic forensic analysis, concerns the use of machine-learning technique within traditionally human forensic discipline: probabilistic genotyping software such as TrueAllele and STRmix in DNA mixture interpretation, algorithmic tool in latent fingerprint comparison, and, more recently, machine-learning classifier applied to digital forensic evidence. Unlike risk-assessment tool, these system are typically introduced as aid to rather than substitute for a human examiner, yet the underlying software often remains proprietary, and defence access to source code or validation data has been a recurring point of litigation in the United States.

These category interact across the temporal arc of a single case rather than operating as isolated point of intervention: a predictive policing output may direct patrol attention toward an individual; a facial recognition match generated during the resulting investigation may identify a suspect; that suspect's bail determination may then be informed by a risk-assessment score; and, at sentencing, the same or a related score may again inform the court's

⁵ Patrick Grother, Mei Ngan & Kayee Hanaoka, NIST Interagency Report 8280: Face Recognition Vendor Test Part 3: Demographic Effects (2019).

⁶ *R (Bridges) v. Chief Constable of South Wales Police*, [2020] EWCA Civ 1058.

⁷ Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016).

recidivism assessment. This compounding effect, sometimes termed runaway feedback, means the reliability of a later-stage tool cannot be fully assessed without considering whether the historical data on which it was trained was itself generated, in part, through earlier-stage algorithmic tool whose own reliability may never have been independently tested.⁸ A further structural feature cutting across every category is the degree of market concentration among vendor: a comparatively small number of firm, including Clearview AI and NEC Corporation in facial recognition and Equivant (formerly Northpointe) in risk assessment, supply software deployed across many jurisdiction, meaning a single vendor's design choice, training data and validation practice can shape outcome for defendant across system with otherwise quite different constitutional and evidentiary tradition.

Taken together, these category share a common structural feature central to this paper's argument: in each instance, a decision-relevant output, a patrol allocation, a facial match, a risk score, a probabilistic genotype, is generated by a computational process whose internal logic is not readily inspectable by the person whose liberty or investigatory status it affects. The remainder of this paper examines what that feature means for the constitutional guarantee of a fair trial.

3. FAIR TRIAL GUARANTEES: DOCTRINAL FOUNDATIONS ACROSS JURISDICTIONS

Before assessing whether existing fair trial doctrine can accommodate algorithmic evidence, it is necessary to identify what that doctrine actually requires. Although the United States, the European human rights system, and India differ in constitutional structure and legal tradition, each has developed a recognisably similar set of minimum guarantee clustered around three normative commitment: the right to contest the evidence against oneself, the right to a reasoned and individualised determination, and the right to equality of arms.

In the United States, the fair trial guarantee is distributed across several constitutional provision. The Sixth Amendment's Confrontation Clause guarantees the right to be confronted with the witness against oneself, construed in *Crawford v. Washington*⁹ to require that testimonial hearsay be subject to cross-examination unless the declarant is unavailable and the defendant had a prior opportunity to cross-examine. The Due Process Clause separately guarantees a fair and individualised process, applied to sentencing in *Gardner v.*

⁸ Danielle Ensign et al., Runaway Feedback Loops in Predictive Policing, 81 Proc. Mach. Learning Rsch. 160 (2018).

⁹ *Crawford v. Washington*, 541 U.S. 36 (2004).

*Florida*¹⁰, which held a defendant could not be sentenced to death on the basis of a confidential presentence report he had no opportunity to deny or explain. More recently, *Carpenter v. United States*¹¹ recognised that acquisition of historical cell-site location data implicates Fourth Amendment protection notwithstanding the third-party doctrine, signalling a broader judicial sensitivity to how large-scale data analytics can alter the balance between law enforcement capability and individual privacy.

The European system layers two sources of protection. Article 6 of the European Convention on Human Rights guarantees a fair and public hearing, interpreted by the European Court of Human Rights to include adversarial proceeding, requiring each party an opportunity to know and comment on evidence adduced by the other.¹² Superimposed on this is Article 22 of the General Data Protection Regulation, granting individual the right not to be subject to a decision based solely on automated processing producing legal effect, subject to a right to obtain human intervention and contest the decision, and Regulation (EU) 2024/1689, the Artificial Intelligence Act, which classifies AI system used in law enforcement and the administration of justice as high-risk, examined further in Section 6. Article 47 of the Charter of Fundamental Rights separately guarantees an effective remedy, which the Court of Justice of the European Union has held requires, for automated decision-making, that the data subject be given access to the underlying logic sufficient to permit effective challenge, a standard elaborated in automated creditworthiness scoring but readily transposable to criminal risk-assessment.¹³

India's constitutional due process jurisprudence, though textually distinct, converges on similar substantive commitment. Article 21 guarantees that no person shall be deprived of life or personal liberty except by procedure established by law, a procedure *Maneka Gandhi v. Union of India*¹⁴ held must be fair, just and reasonable, not arbitrary, fanciful or oppressive. *Selvi v. State of Karnataka*¹⁵ extended this reasoning to hold that involuntary narcoanalysis, polygraph and brain-mapping test violated both the privilege against self-incrimination under Article 20(3) and substantive due process under Article 21, reasoning that technique of contested scientific reliability cannot displace testimonial autonomy. More recently, *Justice K.S. Puttaswamy v. Union of India*¹⁶ recognised privacy as a fundamental right flowing from Article 21, establishing a proportionality standard that has since informed judicial scrutiny of state surveillance and biometric data-collection programme, including

¹⁰ *Gardner v. Florida*, 430 U.S. 349 (1977).

¹¹ *Carpenter v. United States*, 585 U.S. 296 (2018).

¹² *Ruiz-Mateos v. Spain*, App. No. 12952/87, Eur. Ct. H.R. (1993).

¹³ *Case C-634/21, OQ v. Land Hessen*, ECLI:EU:C:2023:957.

¹⁴ *Maneka Gandhi v. Union of India*, A.I.R. 1978 S.C. 597 (India).

¹⁵ *Selvi v. State of Karnataka*, (2010) 7 S.C.C. 263 (India).

¹⁶ *Justice K.S. Puttaswamy v. Union of India*, (2017) 10 S.C.C. 1 (India).

Aadhaar and, by extension, the National Crime Records Bureau's Automated Facial Recognition System.

A further, less frequently discussed point of convergence concerns the privilege against self-incrimination. Where a risk-assessment instrument's input include information gathered through a defendant's own interview response, as with COMPAS's underlying questionnaire, a due process question arises whether an answer given for one purpose can permissibly be repurposed as input into a score later used to justify a harsher sentence without the defendant being informed, at the time of the interview, that the answer might be so used. *Selvi's* reasoning, that the privilege protects against compelled use of personal response in a manner the person did not anticipate and not merely compelled testimonial communication in the narrow sense, offers a doctrinal resource for interrogating this practice that has not, to date, been squarely tested in the algorithmic risk-assessment context in any jurisdiction examined here.

Three normative commitment recur across these otherwise doctrinally distinct system. Contestability requires a meaningful opportunity to examine, question and rebut evidence; individualisation requires a determination reasoned with reference to the particular defendant rather than an aggregate judgment about a statistical group; equality of arms requires defence access to resource and information roughly commensurate with the prosecution's own. Each of these commitment, the following sections demonstrate, is placed under distinctive strain by proprietary, statistically opaque AI system.

4. THE BLACK BOX PROBLEM: EPISTEMIC OPACITY AND THE LIMITS OF EVIDENTIARY DOCTRINE

The phrase black box describes a system whose input and output are observable but whose internal decision logic is not readily interpretable, either because the model is architecturally too complex for human inspection, as with deep neural network comprising millions of weighted parameter, or because the model, though technically interpretable, is deliberately withheld for reason of commercial confidentiality. These two source of opacity, technical and commercial, generate different legal response and should be kept analytically distinct.

Technical opacity arises because contemporary machine-learning system operate through statistical pattern recognition across high-dimensional feature space rather than humanly articulable rule, such that even a system's own developer frequently cannot state, in terms a lay factfinder could evaluate, why a particular input produced a particular output.¹⁷ Explainable AI research has produced post-hoc interpretability technique, including LIME and SHAP value

¹⁷ Will Knight, *The Dark Secret at the Heart of AI*, MIT Tech. Rev. (Apr. 11, 2017).

analysis, that approximate, after the fact, which input feature most influenced a given output, but these technique generate an approximation rather than a faithful account, and different interpretability method applied to the same model can generate materially different and occasionally contradictory explanation of the same output.

Commercial opacity is, by contrast, not an inherent property of the technology but a contractual and proprietary choice. Vendor of risk-assessment and forensic software frequently assert trade-secret protection over internal weighting and source code, resisting disclosure on the ground that such disclosure would compromise a legitimately protected commercial asset. *State v. Loomis* is the paradigmatic illustration: the defendant argued his right to be sentenced on accurate information was violated because COMPAS's proprietary nature prevented him from assessing the accuracy or scientific validity of the score supporting his sentence, an argument the Wisconsin Supreme Court rejected, holding the defendant's due process right adequately protected because he could verify the accuracy of data he personally supplied and the sentencing court retained ultimate discretion, while nonetheless requiring future presentence report incorporating a COMPAS score to include cautionary disclosure regarding the tool's proprietary nature and documented concern regarding disparate racial impact.¹⁸ The decision has attracted extensive academic criticism precisely because it permits reliance on a score whose internal derivation remains, as a matter of law, unreviewable by the very person whose liberty it affects.

The doctrinal significance of this distinction is considerable: commercial opacity is remediable through legal compulsion, a court or regulator can require disclosure of source code, training data provenance and validation methodology to a suitably qualified independent reviewer without altering the technology at all, whereas technical opacity is not similarly remediable by legal fiat, since no amount of compelled disclosure can render a million-parameter neural network intelligible to a lay judge in the way a forensic examiner's stated methodology can be made intelligible through testimony and cross-examination. This suggests different doctrinal response are appropriate to each: mandatory disclosure regime for commercial opacity, and a combination of independent technical auditing, validation-standard requirement, and, in appropriate case, categorical exclusion for irreducible technical opacity.

A related distinction, often elided, concerns a model's overall predictive accuracy against its calibration across relevant subgroup. A tool may achieve respectable aggregate accuracy while being poorly calibrated for a particular subgroup, meaning a given score corresponds to a materially different actual probability of reoffending depending on race, gender or age. The statistical

¹⁸ *State v. Loomis*, 881 N.W.2d 749, 760-61, 769-70 (Wis. 2016).

literature responding to the ProPublica investigation demonstrated precisely this dynamic for COMPAS: the tool could simultaneously satisfy equal predictive value across racial group while failing equal error rate across racial group, a tension Kleinberg, Mullainathan and Raghavan proved formally cannot be resolved except in the trivial case of a perfectly predictive model, wherever the base rate of the predicted outcome differs across the group being compared.¹⁹ A reliability inquiry asking only whether a tool is accurate in the aggregate, without separately asking how it is calibrated across the jurisdiction's specific defendant population, will systematically understate its potential for disparate impact on individual defendant.

Turning to admissibility doctrine specifically, in the United States this is governed, in federal court and most states, by the standard in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, directing trial judge to assess whether proffered expert testimony is reliable by reference to testability, peer review, known error rate, and general acceptance within the relevant scientific community, a minority of states continuing to apply the older *Frye* standard of general acceptance alone.²⁰ Both standard presuppose a testable, publicly describable methodology, precisely what commercial confidentiality claim are designed to withhold, and several scholar have accordingly argued courts apply a substantially attenuated version of *Daubert* scrutiny to algorithmic tool, admitting risk score and probabilistic genotyping output with comparatively little independent testing of the underlying validation data.

The Confrontation Clause present a related but conceptually distinct difficulty: courts and commentator have long debated whether a machine-generated output constitutes a statement by a witness capable of triggering confrontation-clause scrutiny at all, given the doctrine's historical focus on human declarant. Scholarship on so-called machine testimony has argued this human-declarant framing is doctrinally under-inclusive, since what matters from a fair trial perspective is not whether the evidence's source is human or mechanical but whether it embeds assertion the defendant has no meaningful opportunity to test, a condition machine-generated risk score plainly satisfy however technically classified.²¹ Where courts have permitted the human analyst who operated or interpreted the software to be cross-examined in lieu of the software itself, the underlying difficulty persists, since an analyst can testify to how a tool was used but not to why the underlying model reached a particular internal weighting.

¹⁹ Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 8 *Proc. Innovations in Theoretical Computer Sci.* (2017).

²⁰ *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993); *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

²¹ Andrea Roth, *Machine Testimony*, 126 *Yale L.J.* 1972 (2017).

India's evidentiary framework, principally the Bharatiya Sakshya Adhiniyam, which succeeded the Indian Evidence Act, 1872, does not yet contain provision specifically addressed to algorithmic evidence, and courts have approached such evidence largely through the existing framework governing electronic record and expert opinion.²² The reasoning in *Selvi*, though concerned with a different technology, offers a doctrinal template of some relevance, since the Court there declined to defer to proponents' claim of scientific validity, insisting instead on independent judicial assessment of reliability before permitting a technology to bear on a person's guilt or liberty, an approach that, if extended to algorithmic risk-assessment and identification tool, would require Indian courts to resist the kind of largely unexamined deference that has characterised some American sentencing court's treatment of COMPAS-type instrument.

The comparative picture that emerges across these doctrine is one of strain rather than collapse. Existing reliability-screening and confrontation doctrine were not designed with commercially opaque, statistically generated evidence in mind, and courts applying them have generally not adapted the underlying test to compensate for the distinctive opacity of algorithmic tool, resulting in a pattern of admission with comparatively light independent scrutiny, a pattern this paper argues is incompatible with the contestability and equality-of-arms commitment identified in Section 3.

5. AUTOMATION BIAS AND THE EROSION OF INDEPENDENT JUDICIAL REASONING

A distinct but related concern arises not from the formal admissibility of algorithmic evidence but from the psychological dynamic of its use once admitted. Automation bias is the well-documented tendency of human decision-maker to over-rely on automated decision aid, to give algorithmic recommendation greater weight than an equivalent recommendation from a human source, and to be less likely to seek disconfirming information once an automated recommendation has been provided.²³ First documented extensively in aviation and medical decision-making context, studies found operators presented with an automated recommendation both more likely to follow an incorrect recommendation and less likely to detect an error the automation had made, than operators performing the same task unassisted.

Applied to criminal adjudication, automation bias raises a specific institutional concern: a judge confronted with a numerical risk score presented as the output of a validated statistical model may find it professionally and

²² Bharatiya Sakshya Adhiniyam, 2023, No. 47, Acts of Parliament, 2023 (India).

²³ Linda J. Skitka, Kathleen L. Mosier & Mark Burdick, Does Automation Bias Decision-Making?, 51 Int'l J. Human-Computer Stud. 991 (1999).

psychologically difficult to depart from that score, particularly where departure requires articulating, in a written judgment, a specific reason for disagreeing with a tool the court itself lacks the technical training to evaluate. Empirical studies of bail and sentencing decision-making in jurisdictions using actuarial risk instrument have found that, notwithstanding formal doctrine preserving judicial discretion, risk score exert a measurable anchoring effect on outcome, with judicial decision clustering closely around the tool's recommendation even where case-specific factor might have justified a different outcome.²⁴ The Wisconsin Supreme Court's insistence in *Loomis* that the sentencing court retained ultimate discretion, and its accompanying cautionary disclosure, can be understood as a partial doctrinal response to precisely this concern, an attempt to interrupt the anchoring effect by reminding the court, in writing, of the tool's limitation before relying on it; whether such disclosure sufficiently counteracts a well-documented cognitive bias, as opposed to functioning as a formal safeguard satisfied by mere recitation, remains an open empirical question on which the available evidence is, at best, mixed.

A second dimension concerns appellate review. Where a trial court's determination has been informed by an algorithmic score, an appellate court faces the same interpretive difficulty, compounded by a more limited factual record; absent underlying model documentation, validation data, or an independent expert assessment in the trial record, appellate review of the algorithmic component is reduced to a formal check on whether the tool was used, rather than a substantive check on whether its output was sound, insulating the algorithmic component from meaningful appellate scrutiny even in system otherwise providing robust right of appeal.

The behavioural literature on debiasing intervention offers qualified ground for optimism. Experimental studies examining blind review protocol, requiring an independent assessment before viewing an algorithmic recommendation, and requiring a written justification addressed specifically to any point of disagreement, have found modest but measurable reductions in automation bias compared with unstructured exposure, though the same studies caution such intervention are considerably less effective under high caseload pressure and time constraint, condition closely describing the operating environment of many bail and remand court.²⁵ This carries a direct institutional implication, examined further in Section 8: judicial training alone, without accompanying structural reform to caseload and time allocation, is unlikely to achieve the reduction in automation bias the training is intended to produce.

²⁴ Megan T. Stevenson, *Assessing Risk Assessment in Action*, 103 Minn. L. Rev. 303 (2018).

²⁵ Ben Green & Yiling Chen, *Disparate Interactions*, in *Proc. Conf. on Fairness, Accountability, and Transparency* 90 (2019).

A further dimension operates well before any judicial hearing, at the stage of plea negotiation. Because risk score are frequently disclosed to both prosecution and defence counsel in advance of a plea, and because defence counsel operating under significant caseload pressure may treat a high-risk score as effectively determinative of a court's likely posture, an algorithmic output can shape the term on which a plea is negotiated long before any judicial officer exercises independent discretion at all. This pre-adjudicative dimension is particularly resistant to the doctrinal remedy discussed in Section 4, which address principally the admissibility of evidence at a contested hearing, because the vast majority of criminal case in jurisdiction such as the United States are resolved by plea rather than trial, meaning whatever distortive effect a risk score exerts on negotiating leverage will, in most case, never be tested by a court at all.

Automation bias reinforces, rather than duplicates, the black-box concern examined in Section 4. Even a perfectly disclosed, technically interpretable algorithm would remain vulnerable to automation bias if the human decision-maker lacks either the training or institutional incentive to scrutinise its output critically. Disclosure and interpretability are accordingly necessary but not sufficient condition for preserving independent judicial reasoning; they must be accompanied by structured decision-making protocol and judicial training addressing the psychological, rather than merely the technical, dimension of algorithmic reliance.

6. COMPARATIVE REGULATORY RESPONSES AND CASE STUDIES

Regulatory response to AI in criminal justice have developed unevenly across jurisdiction, reflecting difference in constitutional structure, legislative capacity, and political prioritisation.

The European Union has adopted the most comprehensive statutory response. Regulation (EU) 2024/1689, the Artificial Intelligence Act, adopted on 13 June 2024, establishes a risk-tiered architecture classifying AI system by the degree of risk to health, safety and fundamental right.²⁶ System used by law enforcement to assess the risk of a person offending or reoffending, to assess personality trait for predicting criminal risk, or to assess the risk of becoming a victim, together with system used by judicial authority to research and interpret fact and law, are classified high-risk, triggering obligation of risk-management, training-data quality assurance, technical documentation, human oversight, and accuracy and robustness standard throughout the system's lifecycle; certain use, including real-time remote biometric identification in public space for law enforcement, are prohibited outright subject to narrow exception. The Council of Europe has separately adopted a Framework Convention on Artificial

²⁶ Regulation (EU) 2024/1689, 2024 O.J. (L) 1689.

Intelligence and Human Rights, Democracy and the Rule of Law, opened for signature in September 2024, establishing cross-cutting obligation of proportionality, non-discrimination and access to remedy applicable beyond the EU's own membership, ratified by signatory including the United Kingdom and the United States, signalling an emerging international consensus on minimum procedural safeguard even where domestic constitutional doctrine has not yet caught up.²⁷

The United States regulatory landscape is, by contrast, fragmented and predominantly sub-constitutional, with no federal statute directly analogous to the EU Act. Regulatory activity has occurred principally at state and municipal level: Illinois and Washington have enacted statute requiring disclosure of the general characteristic and validation data of risk-assessment instrument used in pre-trial or sentencing decision, and municipality including several within California and Massachusetts have restricted or banned municipal facial recognition use. The National Institute of Standards and Technology's ongoing testing programme provides an important, though voluntary, source of independent validation data, leaving the doctrinal burden of accommodating algorithmic evidence disproportionately on courts applying doctrine, Daubert and the Confrontation Clause, not designed for that purpose.

India's regulatory posture remains comparatively nascent. The Digital Personal Data Protection Act, 2023 establishes general obligation concerning personal data processing but contains no provision specifically addressed to algorithmic evidence, automated decision-making in criminal proceeding, or law enforcement use of facial recognition and predictive analytics.²⁸ The National Crime Records Bureau's Automated Facial Recognition System has been the subject of parliamentary question and civil society advocacy concerning the absence of a specific statutory framework governing accuracy standard, permissible use and independent audit, but has not, to date, been squarely tested before the Supreme Court on due process ground analogous to *Puttaswamy* and *Selvi*. Indian scholarship has accordingly argued that the constitutional resource for an algorithmic due process jurisprudence already exist, in the proportionality framework *Puttaswamy* articulates and the reliability-scrutiny approach *Selvi* articulates, but await a suitable case and a suitably resourced litigant through which to be judicially elaborated. A further dimension of the Indian regulatory gap concerns access to justice more broadly: because a substantial proportion of criminal defendant rely on legal aid counsel operating under severe resource and caseload constraint, the practical capacity of the defence bar to identify, let alone contest, algorithmically generated evidence is likely considerably more limited than in jurisdiction with better-

²⁷ Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, C.E.T.S. No. 225 (2024).

²⁸ Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India).

resourced public defender system, meaning any effective Indian regulatory response will need to be paired with corresponding investment in legal aid capacity specifically trained to identify and challenge such evidence.

The doctrinal analysis above can be sharpened by three illustrative case study, drawn from the United States, the United Kingdom and India, which together illustrate a common pattern: courts and regulator are capable of identifying the due process risk posed by algorithmic tool with considerable sophistication, but the remedy actually granted has fallen short of the disclosure-review-training framework proposed in Section 7.

In *Loomis*, the Wisconsin Supreme Court engaged seriously with the defendant's due process argument, canvassing academic literature on COMPAS's limitation and acknowledging the proprietary nature of the tool prevented full verification of its methodology, yet the remedy fashioned, cautionary disclosure in future presentence report, operates entirely at the level of notice rather than independent verification: the sentencing court is told the tool has limitation but given no institutional mechanism, expert, audit body or standardised validation report, through which to assess whether those limitation are present, and to what degree, in the specific case before it. The gap between the Court's diagnosis and the modesty of its remedy is, in microcosm, the gap this paper identifies throughout.

Bridges illustrates a structurally different but related dynamic. The Court of Appeal held South Wales Police's live facial recognition use unlawful principally because the force had failed its public sector equality duty, not taking reasonable step to satisfy itself the software carried no unacceptable bias, notwithstanding the NIST testing evidence discussed in Section 2 that such bias was a live and known risk. The remedy was procedural rather than substantive, requiring a documented equality impact assessment before resuming deployment rather than mandating a specific accuracy threshold or ongoing external verification, and South Wales Police has since been able to resume deployment on the strength of an internally commissioned assessment, a outcome falling considerably short of the independent technical review this paper argues necessary.

The Indian position, by contrast, is one of doctrinal potential rather than doctrinal engagement. No reported Supreme Court decision has yet addressed the due process implication of the NCRB's facial recognition system or any comparable risk-assessment tool, notwithstanding the existence of a constitutional proportionality framework, articulated in *Puttaswamy*, doctrinally well suited to the task: that test, requiring state interference with a fundamental right to be backed by law, pursue a legitimate aim, be necessary in the sense that no less restrictive alternative would achieve the aim, and be proportionate in the narrow sense, maps closely onto the disclosure and review obligation

proposed in Section 7, and a facial recognition or risk-assessment system deployed without validation data specific to the Indian population would struggle to satisfy the necessity limb given the ready availability of the less restrictive alternative of requiring such validation as a precondition of deployment. The absence to date of a case squarely presenting this argument reflects less a doctrinal gap than a gap in litigation capacity and public awareness.

Two lessons emerge. First, the doctrinal capacity to recognise the fair trial risk posed by algorithmic tool already exists, in varying form, across all three jurisdiction; what is lacking, consistently, is the institutional machinery capable of translating that recognition into an enforceable, substantively meaningful safeguard. Second, the remedy courts have fashioned to date have tended toward procedural formalism rather than the substantive, independently verified reliability threshold the framework in Section 7 argues is necessary.

7. TOWARDS ALGORITHMIC DUE PROCESS: A NORMATIVE FRAMEWORK AND ITS CRITICS

The preceding sections have identified a structural mismatch between the evidentiary and procedural apparatus of fair trial doctrine and the technological characteristic of AI system now deployed across the criminal process. This section develops a normative framework, algorithmic due process, intended to reconcile the legitimate efficiency gain AI can offer with the three core commitment identified in Section 3, resting on three interlocking pillar.

The first pillar is mandatory disclosure of algorithmic logic, responding directly to the commercial-opacity problem identified in Section 4. It requires that any AI system used to generate evidence or inform a dispositional decision be accompanied, as a condition of admissibility or use, by documentation sufficient to permit an independent technical reviewer, though not necessarily the defendant personally, to assess the system's design, training data provenance, validation methodology, and known error rate and demographic performance differential. Mandatory disclosure of this kind need not require wholesale publication of proprietary source code; it can be structured, following model already used in intellectual property and national security litigation, through protective order permitting disclosure to a court-appointed or defence-retained technical expert bound by confidentiality obligation. The critical doctrinal shift this pillar demands is rejection of the reasoning in *Loomis* insofar as that decision treated trade-secret confidentiality as a sufficient reason to deny any independent technical scrutiny whatsoever, rather than merely a reason to structure such scrutiny through appropriately confidential channel.

The second pillar is independent technical review. Disclosure alone accomplishes little if the recipient lacks the expertise or institutional capacity to

evaluate what has been disclosed; this pillar proposes a standing technical review capacity, analogous to the office of a court-appointed forensic expert but specifically resourced and trained to assess machine-learning system, empowered to evaluate a tool's validation methodology, assess whether it has been validated on a population comparable to the jurisdiction of deployment, and flag known limitation including demographic performance differential of the kind documented in the NIST and ProPublica material discussed in Section 2. Such a body could operate along the line of existing specialised judicial support institution, adapted to the distinctive technical demand of machine-learning system. Where independent review concludes a system's error characteristic or validation gap render its output unreliable for the purpose proposed, the framework contemplates exclusion of the evidence notwithstanding its algorithmic packaging.

The third pillar is judicial training on AI literacy, responding directly to the automation-bias concern examined in Section 5. It requires that judicial officer called upon to assess, weigh or rely upon algorithmic evidence receive structured training, developed in consultation with independent technical expert rather than vendor, covering the basic statistical and methodological concept necessary to interrogate a validation report, the documented psychology of automation bias and debiasing technique such as structured decision protocol requiring an affirmative, individualised statement of reason before an algorithmic recommendation is adopted, and the jurisdiction-specific disclosure and review mechanism established under the first two pillar. Comparative experience with judicial training on other technically demanding evidentiary category, DNA evidence and complex financial fraud among them, suggests structured training of this kind can meaningfully improve the quality of judicial engagement with technical evidence, provided it is mandatory, recurrent, and delivered independently of any party with a commercial interest in the tool's continued use.

These three pillar operate cumulatively rather than as alternative: disclosure without independent review risks becoming a formal but substantively empty compliance exercise; independent review without judicial training risks producing expert report the deciding judge is not equipped to weigh appropriately; and judicial training without the underlying disclosure and review infrastructure risks equipping judge with critical faculty they have no factual basis on which to exercise. The three-pillar structure is not without precedent: forensic science oversight reform undertaken in several jurisdiction following high-profile wrongful-conviction scandal involving unreliable forensic discipline followed a broadly similar structure, mandatory accreditation and disclosure requirement, independent standard-setting and audit body, and mandatory continuing education for both examiner and judge, meaning the framework proposed here extends an institutional model criminal justice system

have already found workable rather than requiring invention of an entirely new regulatory paradigm.²⁹

7.1 Critiques and Limits of the Framework

No normative framework of this kind is beyond legitimate criticism, and intellectual honesty requires several such criticism be addressed directly rather than assumed away.

The first is institutional feasibility: establishing standing independent technical review capacity requires resource many criminal justice system, particularly in lower- and middle-income jurisdiction, may struggle to provide, especially where such system already face chronic under-resourcing of conventional forensic science capacity. This criticism has real force; the response offered is modest rather than triumphant, since the first and third pillar can be implemented at comparatively lower cost and can meaningfully improve the status quo in advance of the second pillar's full realisation, and regional or international pooling of technical review capacity, discussed in Section 8, offers a means of achieving economies of scale a single under-resourced jurisdiction could not achieve alone.

The second is potential over-correction: a regime of mandatory disclosure and independent review, implemented without sensitivity to case volume and proportionality, risks imposing delay and cost on the ordinary run of case in which an algorithmic tool's output is not seriously contested, potentially undermining a defendant's own interest in timely resolution, particularly for those held in pre-trial detention. The response suggested here is that the obligation proposed need not be triggered as a matter of course in every case; they are more properly conceived as a standing entitlement a defendant, or the court itself, may invoke where the algorithmic evidence is genuinely material to the outcome, calibrated as existing expert-evidence disclosure regime already calibrate the burden of contesting conventional forensic evidence to its significance in the particular case.

A third, more fundamental criticism questions whether disclosure and review can ever adequately substitute for the understanding a human factfinder derives from observing and cross-examining a live witness, suggesting the more honest response to irreducibly opaque system is simply to exclude them from the criminal process altogether.³⁰ This paper's response is that categorical exclusion remains available, and indeed required, under the framework proposed wherever independent technical review concludes a given system's opacity is irreducible in the sense that no validation methodology can render its

²⁹ President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (2016).

³⁰ Rebecca Wexler, *Life, Liberty, and Trade Secrets*, 70 *Stan. L. Rev.* 1343 (2018).

error characteristic assessable; the framework does not assume every tool can be rendered fair-trial-compatible through disclosure and review, only that the question of whether a given tool can be so rendered ought to be settled through independent technical assessment rather than blanket administrative or judicial assumption in either direction.

Finally, it might be objected that the framework, by requiring demonstrable reliability as a precondition of admission, will simply drive law enforcement and prosecutorial use of AI underground, informing investigative strategy and charging decision without ever being formally introduced as evidence, thereby evading the obligation proposed entirely. This is a genuine risk this paper does not claim to resolve completely; it suggests the appropriate response is to extend disclosure obligation, as several state statute discussed in Section 6 already do in nascent form, to investigative use that materially informs a charging or detention decision even where the tool's output is never formally tendered as evidence at trial.

8. INSTITUTIONAL RECOMMENDATIONS AND CONCLUSION

Translating the normative framework into concrete institutional reform requires attention to legislative, judicial and administrative mechanism.

- At the legislative level, jurisdiction without a comprehensive statutory framework, most notably the United States at federal level and India, should enact legislation establishing a mandatory pre-deployment validation and disclosure obligation applicable to any AI system procured or used by a law enforcement or judicial body, modelled on the EU Act's high-risk obligation but adapted to the evidentiary posture of criminal proceeding; a statutory right, enforceable by the defence, to independent technical review of any algorithmic tool materially informing an investigatory, charging, bail or sentencing decision, structured through protective-order mechanism preserving legitimate commercial confidentiality while guaranteeing meaningful scrutiny; and a statutory obligation on procuring agency to maintain and periodically update jurisdiction-specific validation data, rather than relying solely on validation study conducted on demographically different population elsewhere, a gap the Loomis court itself identified as unresolved in COMPAS's Wisconsin deployment.
- At the judicial level, courts of final appeal should consider articulating, through ordinary interpretive process, a presumption against admissibility of algorithmic evidence or risk score whose underlying validation methodology and error-rate data have not been made available, in confidentiality-protected form, to an independent

technical reviewer, reversing the practical burden decisions such as Loomis have placed on defendants to demonstrate unreliability, and instead requiring the proponent, ordinarily the state, to demonstrate reliability as a precondition of admission. Appellate courts should further require that any judgment relying on an algorithmic score include a specific, individualised statement explaining why the court found the output reliable as applied to the particular defendant, operationalising the individualisation commitment identified in Section 3.

- At the administrative level, judicial training academy, India's National Judicial Academy and the United States' Federal Judicial Center among them, should incorporate structured, vendor-independent AI literacy training into mandatory continuing judicial education, covering both the statistical concept necessary to interrogate a validation report and the documented psychology of automation bias. Procuring agency should be required to conduct and publish periodic post-deployment audit of any AI tool in continued use, given the dynamic character of contemporary machine-learning system, ensuring a validation finding made at initial deployment does not become stale as the underlying model, or the population against which it operates, evolves.

A recurring practical objection concerns funding for independent technical review capacity, and one workable model, adapted from the funding structure of several existing forensic science accreditation body, is a mandatory per-deployment licensing fee, paid by any vendor seeking approval for use in the criminal justice system of the relevant jurisdiction, into a ring-fenced fund administered independently of both the procuring agency and the vendor and used exclusively to finance the review body's operation, aligning the cost of independent review with the party best placed to bear it while structurally insulating the reviewing body's funding from vendor-by-vendor commercial relationship that might otherwise compromise its independence.

Given the transnational character of much of the underlying technology, with the same commercially developed facial recognition, risk-assessment and forensic software deployed across multiple jurisdiction, there is a strong case for international coordination on minimum validation and disclosure standard, building on the Council of Europe Framework Convention, so that jurisdiction with less developed domestic regulatory capacity are not left to rely entirely on vendor-supplied assurance untested against local condition. Such coordination would not require doctrinal harmonisation, neither realistic nor, given genuine difference in constitutional structure, necessarily desirable; it would require only a shared technical baseline against which any jurisdiction's court could

assess a given tool's reliability, leaving the ultimate question of admissibility and evidentiary weight to each jurisdiction's own fair trial doctrine.

This paper has argued that the deployment of artificial intelligence across the investigatory, adjudicative and dispositional stage of the criminal process places genuine and, to date, insufficiently addressed strain on the fair trial guarantee as elaborated in the United States, the European human rights system, and India. That strain arises not from any single defect in existing doctrine but from a structural mismatch: evidentiary and procedural rule developed around the assumption of a human, or at least humanly describable, source of evidence are being asked to accommodate output generated by statistically opaque and, in significant part, commercially withheld computational process. The result, documented across the case law and empirical literature surveyed throughout, is a pattern in which algorithmic evidence is frequently admitted and relied upon with a degree of independent scrutiny that would be regarded as inadequate if applied to a comparably consequential piece of conventional forensic or testimonial evidence.

The normative framework developed in Section 7, algorithmic due process, resting on mandatory disclosure, independent technical review, and judicial AI literacy, is offered not as a wholesale rejection of AI's role in criminal justice but as the doctrinal and institutional precondition on which that role can be reconciled with the constitutional commitment to contestability, individualisation and equality of arms the fair trial guarantee exists to protect. The case study examined in Section 6, drawn from three jurisdiction with markedly different constitutional tradition, demonstrate that the doctrinal resource needed to recognise the fair trial risk of algorithmic evidence already exist in each system; what each currently lacks, in varying degree, is the institutional machinery capable of converting that doctrinal recognition into a substantively enforceable safeguard. The broader claim underlying this paper is modest but, it is submitted, important: that the efficiency gain AI can offer the criminal process are neither inherently incompatible with, nor automatically reconcilable with, fair trial guarantee, and that the difference between these two outcome lies substantially in whether the institutional reform this paper has proposed are undertaken before, rather than after, algorithmic tool become an entrenched and unexamined feature of ordinary criminal adjudication.