FORECASTING GUILT: DANGEROUSNESS, ALGORITHMIC PREDICTION, AND THE CONSTITUTIONAL LIMITS OF PREVENTIVE JUSTICE

Animesh Singh, Assistant Professor, Law, Babu Banarasi Das University

ABSTRACT

Criminal justice systems are shifting from a reactive model, which punishes conduct already committed, toward a preventive model, which restrains liberty on the basis of anticipated future conduct. This shift is visible in clinical and actuarial risk assessment, in AI-assisted predictive policing and algorithmic sentencing tools, and in long-standing preventive detention regimes. This paper asks whether dangerousness can be predicted with sufficient psychological and scientific reliability to justify depriving a person of liberty before any offence occurs; whether algorithmic decision-making reduces or reproduces human bias in that prediction; whether preventive detention is constitutionally compatible with prediction-based justice; and what safeguards should govern AI-assisted predictive justice. Adopting a doctrinal, comparative, and interdisciplinary methodology centred on India, with comparative reference to the United States and the United Kingdom, the paper traces the psychological science of violence prediction, the architecture and documented biases of algorithmic risk-assessment tools, and India's existing statutory model of preventive detention under the National Security Act 1980, the Unlawful Activities (Prevention) Act 1967, and the Conservation of Foreign Exchange and Prevention of Smuggling Activities Act 1974, read against Article 21 and Article 22 jurisprudence from A.K. Gopalan to Puttaswamy. It argues that dangerousness is an inherently probabilistic construct that cannot, by itself, constitute an independent basis for depriving personal liberty; that artificial intelligence enhances predictive capacity without eliminating the psychological uncertainty and systemic bias underlying dangerousness assessment; and that predictive justice consequently requires a constitutional framework anchored in criminal psychology, due process, and proportionality rather than in confidence borrowed from computation.

Keywords: Dangerousness; Criminal Psychology; Predictive Policing; Preventive Detention; Algorithmic Risk Assessment; Constitutional Law; Due Process.

I. Introduction

Criminal justice has traditionally been reactive: the state investigates, charges, and punishes conduct after it occurs, and liberty is restrained only upon proof of a completed offence. Over the past half-century, and with increasing speed over the past decade, this reactive model has been supplemented by a preventive one, in which the state restrains conduct, movement, or liberty in anticipation of future harm rather than in response to past wrongdoing. Predictive policing tools now allocate patrol resources toward places statistically associated with future crime; algorithmic risk-assessment instruments generate scores that inform bail, sentencing, and parole decisions; and artificial intelligence is increasingly proposed, and in limited respects already used, as a decision-support tool within this predictive architecture. Underlying every one of these developments is a single contested concept: dangerousness — the notion that an identifiable subset of individuals can be reliably distinguished, in advance, as more likely than others to cause future harm.

Dangerousness is not a new foundation for criminal justice intervention. Preventive detention, which authorises the state to deprive a person of liberty without trial on the ground that they are likely to act prejudicially in the future, has existed in India since the colonial period and was expressly preserved by the framers of the Constitution under Article 22. What has changed is the instrumentation available to assess dangerousness: from a magistrate's or psychiatrist's clinical judgment, to statistically validated actuarial instruments, to machine-learning systems trained on historical criminal-justice data. Each instrument promises greater accuracy than its predecessor; none has eliminated the underlying probabilistic uncertainty that dangerousness, as a predictive judgment about a future that has not yet happened, necessarily carries.

Four bodies of scholarship inform this inquiry, and each has largely developed in isolation from the others. Criminal psychology has produced a substantial empirical literature on the prediction of violence and recidivism, from John Monahan's foundational critique of clinical prediction to the subsequent development and validation of actuarial instruments (Monahan, 1981; Monahan et al., 2001). Constitutional scholarship has examined preventive detention as a distinct category of state power, situated uneasily between regulation and punishment, most influentially in the United States Supreme Court's decision in *United States v. Salerno*, 481 U.S. 739 (1987), and in India's own line of Article 21 and Article 22 jurisprudence beginning

with *A.K. Gopalan v. State of Madras*, AIR 1950 SC 27. Predictive policing and algorithmic governance scholarship has documented the migration of risk assessment from clinical to computational settings, most visibly through the litigation surrounding the COMPAS instrument in *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016). And a growing body of work on algorithmic bias, catalysed by ProPublica's 2016 investigation into COMPAS, has demonstrated that computational risk assessment can reproduce, rather than correct, the demographic disparities present in the data on which it is trained (Angwin, Larson, Mattu, & Kirchner, 2016; Dressel & Farid, 2018).

Existing scholarship examines dangerousness from psychological, technological, or legal perspectives largely independently of one another. Limited research analyses dangerousness as a continuous concept evolving from psychological risk assessment, through AI-assisted predictive tools, to preventive detention as practised in India. This paper addresses that gap by treating clinical prediction, actuarial and algorithmic risk assessment, and statutory preventive detention as points along a single continuum of prediction-based justice, asking in each case what psychological and constitutional limits apply.

Four questions organise the analysis that follows. Can dangerousness be predicted with sufficient psychological and scientific reliability to justify restraining liberty in advance of an offence? Does algorithmic decision-making reduce or reproduce human bias in that prediction? Is preventive detention constitutionally compatible with a justice system organised around prediction rather than proof? And what safeguards should govern AI-assisted predictive justice going forward? The argument developed here is that dangerousness is an inherently probabilistic concept that cannot constitute an independent basis for depriving personal liberty. Artificial intelligence enhances predictive capacity in narrow, measurable respects, but it cannot eliminate the psychological uncertainty and systemic bias that underlie dangerousness assessments, because both the uncertainty and the bias are substantially inherited from the human judgments and historical data on which any predictive system, however sophisticated, is built. Predictive justice, accordingly, requires a constitutional framework informed by criminal psychology, due process, and proportionality, rather than a framework that defers to prediction because it is framed in the language of science or computation.

The paper adopts a doctrinal, comparative, and interdisciplinary methodology. Its primary focus is India, with comparative reference to the United States and the United Kingdom, chosen

because each represents a distinct constitutional strategy for accommodating preventive state power: the United States through a doctrinally narrow exception to ordinary due process defended in *Salerno*; the United Kingdom through a civil, judicially supervised preventive-measures regime operating in the shadow of Article 5 of the European Convention on Human Rights; and India through preventive detention as an express, constitutionally textualised power under Article 22 rather than an exception grafted onto ordinary criminal procedure. The analysis is limited to predictive policing, preventive detention, and algorithmic risk assessment within criminal justice, and does not extend to related but distinct domains such as immigration detention or civil commitment, except where these illuminate the doctrinal architecture directly at issue.

II. Dangerousness as a Psychological Construct: The Science and Limits of Predicting Criminal Behaviour

The idea that certain individuals are identifiably "dangerous" in advance of any specific act has a long history in criminology, running from nineteenth-century positivist criminology's search for a criminal "type" through to the mid-twentieth-century psychiatric practice of certifying dangerousness as a precondition for civil commitment and parole decisions. What twentieth-century criminal psychology contributed, decisively, was empirical scrutiny of whether clinicians asked to make such judgments could do so accurately. The verdict, delivered most influentially by John Monahan's 1981 monograph *The Clinical Prediction of Violent Behavior*, was sobering: unstructured clinical predictions of dangerousness were wrong roughly twice as often as they were right, meaning that evaluators who identified a subject as dangerous were incorrect in the substantial majority of cases where the base rate of actual violence was low (Monahan, 1981). This finding reframed dangerousness from a fixed trait that a trained clinician could reliably detect into a dynamic, situationally contingent probability that clinical judgment alone systematically overestimated.

That reframing produced two distinct methodological traditions. Clinical risk assessment relies on a trained evaluator's individualised judgment, informed by interview, history, and diagnostic criteria, and its central strength — sensitivity to the particular circumstances of a given individual — is also the source of its principal weakness, since individualised judgment is vulnerable to the cognitive biases examined below. Actuarial risk assessment instead derives a numerical risk score from statistically validated correlates of violence or recidivism drawn

from large historical samples, and a substantial body of comparative research has found that actuarial instruments consistently outperform unstructured clinical judgment in predictive accuracy (Litwack, 2001). A related, more recent development is the structured professional judgment model, which supplements a clinician's assessment with an empirically derived checklist of risk factors, and behavioural prediction models more broadly, including tools such as the MacArthur Violence Risk Assessment Study's iterative classification approach, which combine actuarial and contextual variables to produce probabilistic, rather than categorical, risk classifications (Monahan et al., 2001; Appelbaum, Robbins, & Monahan, 2000). Whatever the method, the object being assessed is functionally identical to what earlier generations termed dangerousness: an estimate of the probability that a given individual will engage in future violent or recidivist conduct.

Even the best-validated actuarial instruments operate within psychological and statistical limits that no amount of instrument refinement fully removes. False positives and false negatives are not incidental design flaws but a structural consequence of predicting a low base-rate event: because serious violence is statistically rare even among populations flagged as high risk, any instrument calibrated to catch most future offenders will necessarily misclassify a substantial number of people who will never offend, and any instrument calibrated to minimise such misclassification will necessarily miss a meaningful share of those who will (Litwack, 2001). Behavioural variability and environmental influence compound this problem, since violence is not a fixed disposition but an outcome shaped by situational triggers, relationships, substance use, and opportunity, meaning that a risk estimate generated at one point in time degrades in accuracy as the person's circumstances change. This is why contemporary risk-assessment literature increasingly distinguishes static risk factors, which are fixed historical facts such as prior convictions, from dynamic risk factors, which are changeable circumstances such as employment, housing stability, or substance use, and treats only continuous reassessment against dynamic factors as methodologically defensible (Appelbaum et al., 2000). The consequence is that any single dangerousness assessment is, at best, a time-bound probability statement rather than an ascertainable fact about a person's character.

Layered onto these scientific limits is a further, distinctly psychological problem: the decision-makers who interpret risk assessments, whether judges, parole boards, or police officers, are themselves subject to well-documented cognitive biases that distort how probabilistic information is used. Confirmation bias leads decision-makers to weigh evidence

consistent with an initial impression of dangerousness more heavily than disconfirming evidence. The availability heuristic causes vivid or memorable instances of violence, such as a recent high-profile offence, to be given disproportionate weight relative to their actual statistical frequency, a bias formally described in the foundational heuristics-and-biases literature developed by Amos Tversky and Daniel Kahneman (Tversky & Kahneman, 1974). Anchoring bias causes an initial risk score or classification to unduly influence a decision-maker's subsequent judgment even where later, more individualised information should displace it. Implicit bias introduces systematic, often unconscious, distortion correlated with a subject's race, class, or other social identity, a phenomenon of particular significance given that many of the historical correlates of "risk" used to validate actuarial instruments are themselves artefacts of biased policing and prosecution patterns rather than neutral measures of underlying propensity. Finally, labelling theory and the self-fulfilling prophecy describe how the formal designation of a person as dangerous can itself alter the social response to that person — through increased surveillance, restricted opportunity, or defensive social exclusion — in ways that make future offending more, rather than less, likely (Merton, 1948). Dangerousness prediction, in other words, is not merely imperfect; in some documented respects, it is partially self-fulfilling.

The cumulative implication of this literature is that criminal psychology does not offer, and by its own internal methodological standards cannot offer, a reliable basis for treating dangerousness as a determinate fact about an individual analogous to guilt established at trial. What the discipline offers instead is a probabilistic estimate, sensitive to base rates, situational context, and the cognitive frailties of whoever interprets it, that degrades over time and is vulnerable to distortion by the very act of prediction. Dangerousness is properly understood as probabilistic rather than deterministic, which means that the absolute prediction of future criminal behaviour that preventive justice regimes implicitly claim to rest upon is not merely difficult to achieve with current methods; it is, on the current state of the science, scientifically unattainable in the individualised, certainty-bearing form that deprivation of liberty would seem to require.

III. From Human Judgment to Algorithmic Prediction: The Rise of Predictive Justice

The migration of dangerousness assessment from clinical judgment to computation has proceeded through recognisable stages. Place-based predictive policing, exemplified by

geographic hotspot-mapping tools developed and evaluated by policy research organisations such as the RAND Corporation, allocates patrol resources toward locations statistically associated with elevated crime risk, treating dangerousness as a property of places rather than persons (Perry et al., 2013). Person-based prediction extends the same logic to individuals, generating risk scores or watch-lists intended to identify people statistically likely to be involved in future violence, whether as offenders or victims. AI-assisted policing represents the most recent stage, in which machine-learning systems trained on large historical datasets generate risk classifications with a degree of statistical sophistication, and a corresponding opacity, that neither clinical judgment nor earlier actuarial instruments possessed.

Within criminal adjudication specifically, algorithmic risk assessment now informs bail, sentencing, and parole decisions in multiple jurisdictions through predictive analytics tools that apply machine-learning techniques to historical case data, and through risk-scoring models that translate a defendant's characteristics into a numerical or categorical risk classification presented to a judge alongside, or in place of, traditional sentencing factors. The most extensively litigated such instrument is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a proprietary tool that generates recidivism-risk scores from a lengthy questionnaire covering criminal history, social environment, and related factors (Harvard Law Review, 2017). Because COMPAS's scoring algorithm is a protected trade secret, the tool has become the central case study in the debate over the explainability and transparency of algorithmic decision-making in criminal justice: a defendant sentenced partly on the strength of a COMPAS score cannot inspect the precise weighting the algorithm assigned to the factors used to classify him (Freeman, 2016).

That debate reached its most significant judicial resolution in *State v. Loomis*, in which the Wisconsin Supreme Court held that a sentencing court's consideration of a COMPAS risk score did not, without more, violate a defendant's due process rights, provided that the score was not the determinative factor in sentencing and that the presentence report accompanying it carried a specified warning about the tool's limitations, including that it had not been cross-validated on the local population and had been the subject of research questioning its accuracy across racial groups (Harvard Law Review, 2017). The court thereby permitted algorithmic risk scores to inform, but constitutionally required them not to determine, deprivations of liberty — a compromise that critics have argued does not resolve the underlying due process problem, since a defendant denied access to the algorithm's internal logic remains unable to meaningfully

contest the accuracy of the classification driving the recommendation the judge in fact received (Freeman, 2016).

The concern that algorithmic tools might reduce rather than reproduce human bias has been substantially undermined by empirical research into their actual operation. ProPublica's 2016 investigation into COMPAS, based on an analysis of scores assigned to more than seven thousand defendants in Broward County, Florida, found that Black defendants who did not subsequently reoffend were nearly twice as likely as white defendants who did not reoffend to have been classified by the algorithm as high risk, while white defendants who did reoffend were substantially more likely than Black defendants who reoffended to have been misclassified as low risk (Angwin et al., 2016; ProPublica, 2016). A subsequent peer-reviewed study published in *Science Advances* extended this concern by demonstrating that COMPAS's predictive accuracy, at roughly sixty-five per cent, was statistically indistinguishable from the accuracy achieved by untrained members of the public given only a handful of variables about each defendant, and was matched by a simple linear classifier using only two variables — age and total number of prior convictions — casting doubt on whether the tool's proprietary complexity contributed any predictive value beyond what simple, transparent statistical models already achieve (Dressel & Farid, 2018). Historical bias embedded in training data, feedback loops in which policing patterns generated by earlier predictive deployments become the training data for future predictions, and the resulting risk of algorithmic discrimination are not, on this evidence, hypothetical concerns but empirically documented features of currently deployed criminal-justice algorithms.

Comparative constitutional responses to prediction-based liberty deprivation illuminate the stakes of this technical debate. In *United States v. Salerno*, the U.S. Supreme Court upheld the Bail Reform Act of 1984's authorisation of pretrial detention based on a judicial finding, by clear and convincing evidence after an adversarial hearing, that no release condition would reasonably assure community safety, reasoning that such detention was regulatory rather than punitive and therefore did not require the full protections of a criminal trial (Cornell Law School Legal Information Institute, n.d.; Justia, n.d.). Justice Marshall's dissent, by contrast, argued that pretrial detention based on predicted future dangerousness was categorically incompatible with the presumption of innocence, since it authorised the state to imprison a person not for what they had done but for what a judge believed they might do (Justia, n.d.). *Salerno* remains the doctrinal high-water mark for prediction-based detention in American

constitutional law, and its reasoning — that procedural safeguards can render prediction-based detention regulatory rather than punitive — recurs, in a different institutional register, in both the United Kingdom's post-2011 preventive measures regime and India's preventive detention statutes discussed below.

The United Kingdom's Terrorism Prevention and Investigation Measures (TPIMs), introduced by the Terrorism Prevention and Investigation Measures Act 2011 to replace the earlier control-order regime, restrict the liberty of persons believed to be involved in terrorism-related activity who cannot be prosecuted or, for foreign nationals, deported, through measures such as residence requirements, travel restrictions, and communication limits, imposed on the Home Secretary's assessment of risk rather than following conviction (UK Government, 2011). The predecessor control-order regime was repeatedly found by British courts to violate Article 5 of the European Convention on Human Rights, the right to liberty and security, where its cumulative restrictions amounted to a de facto deprivation of liberty, and the shift to TPIMs was explicitly designed to impose lighter restrictions precisely so as to avoid the Article 5 threshold rather than to alter the underlying logic of prediction-based restraint (Explanatory Notes to the Terrorism Prevention and Investigation Measures Act 2011, in UK Government, 2011). Comparative scholarship on this transition has observed that the human-rights concerns raised by control orders substantially recur, in attenuated form, under TPIMs, because both regimes ultimately rest on an executive assessment of future risk that the reviewing court can scrutinise only for reasonableness, not for the kind of evidentiary certainty ordinary criminal process would require (Hunt, 2014).

What unites *Loomis*, *Salerno*, and the TPIM regime is a shared institutional move: each replaces, or supplements, individualised proof of past conduct with an assessment of probable future conduct, and each constructs procedural safeguards — a sentencing warning, an adversarial hearing, a time-limited notice subject to judicial review — designed to render that substitution constitutionally tolerable rather than to eliminate the underlying predictive uncertainty. Artificial intelligence does not disrupt this pattern; it intensifies it, by replacing the individualised judgment that at least nominally undergirds clinical and even much actuarial risk assessment with a statistical classification whose internal logic may be inaccessible even to the court applying it. AI replaces human discretion with algorithmic discretion, but that algorithmic discretion remains dependent upon imperfect historical data, probabilistic reasoning that cannot be converted into certainty by additional computation, and the same

institutional biases that shaped the criminal-justice data on which any such system is necessarily trained.

IV. Preventive Detention: India's Existing Model of Predictive Justice

India's preventive detention regime is best understood not as a novel response to the challenges of algorithmic prediction but as a decades-old, constitutionally entrenched instance of exactly the predictive logic examined above, applied through human rather than computational judgment. Preventive detention differs from punitive detention in its temporal orientation and evidentiary basis: punitive detention follows a judicial determination of guilt for a specific completed offence, established through trial and proof beyond reasonable doubt, whereas preventive detention authorises the executive to detain a person, without trial, on the ground that their future conduct is likely to prejudice defined public interests. India's use of this power did not originate with independence; colonial-era measures such as the Bengal Regulation III of 1818 already empowered the government to detain individuals for the defence or maintenance of public order without recourse to ordinary judicial process, and post-independence India, unusually among constitutional democracies, chose to preserve and constitutionally entrench this power rather than treat it as an emergency exception to ordinary criminal procedure.

The principal statutory instantiations of this power illustrate its breadth. The National Security Act, 1980, empowers the Central and State Governments to detain a person, for up to twelve months, where satisfied that detention is necessary to prevent acts prejudicial to the defence of India, India's security, the maintenance of public order, or the maintenance of essential supplies and services, based on the detaining authority's subjective satisfaction rather than proof of a specific act (National Security Act, 1980, ss. 3, 8–13). The Unlawful Activities (Prevention) Act, 1967, originally enacted to address secessionist activity, now provides the principal statutory framework for countering terrorism-related conduct and, following its 2019 amendment, permits the executive to designate individuals, not only organisations, as terrorists, extending the Act's predictive logic from group membership to individual classification. The Conservation of Foreign Exchange and Prevention of Smuggling Activities Act, 1974, permits preventive detention, for an initial period extending to six months, of persons believed likely to engage in smuggling or foreign-exchange violations, illustrating that India's preventive detention architecture extends well beyond national-security contexts into economic offences

(Conservation of Foreign Exchange and Prevention of Smuggling Activities Act, 1974, s. 3). Several States additionally maintain Public Safety Act or "goonda" legislation directed at habitual offenders and bootleggers, extending the same predictive architecture to ordinary policing well below the threshold of national security.

This statutory architecture operates within, and is expressly authorised by, Article 22 of the Constitution, which stands in marked contrast to Article 21's general guarantee of life and personal liberty. Article 22(3)(b) exempts persons detained under preventive detention law from the procedural protections otherwise available to arrested persons, including the ordinary right to be informed of the grounds of arrest and to consult counsel, while Article 22(4) imposes an outer limit: no preventive detention law may authorise detention beyond three months unless an Advisory Board, composed of persons qualified to be High Court judges, reports that there is sufficient cause for continued detention (National Security Act, 1980, s. 9). The resulting procedural safeguards are real but narrow: the detaining authority must ordinarily communicate the grounds of detention within five days, extendable to fifteen in exceptional circumstances, while withholding any material it considers contrary to the public interest; the detainee may make a representation against the order but has no statutory right to legal representation before the Advisory Board itself. The safeguard structure, in other words, is calibrated to make executive prediction of dangerousness procedurally accountable without subjecting it to the evidentiary standards that would apply to a criminal charge.

Judicial interpretation of this framework has evolved substantially since independence, tracing almost exactly the same due-process trajectory later replicated, in compressed form, by the American COMPAS litigation. In *A.K. Gopalan v. State of Madras*, the Supreme Court, by a four-to-one majority, upheld the Preventive Detention Act, 1950 against a challenge grounded in Articles 14, 19, and 21, holding that Article 21 required only "procedure established by law," not substantive due process, and that fundamental rights under different constitutional provisions operated in mutually exclusive, textually sealed compartments (A.K. Gopalan v. State of Madras, AIR 1950 SC 27). This narrow, textualist reading meant that so long as a validly enacted statute prescribed a procedure, and that procedure was followed, the substantive fairness of the resulting deprivation of liberty was constitutionally irrelevant. *Maneka Gandhi v. Union of India* decisively overruled this approach, holding that the "procedure established by law" under Article 21 must itself be fair, just, and reasonable, and that Articles 14, 19, and 21 must be read together rather than as isolated, self-contained

guarantees (Maneka Gandhi v. Union of India, AIR 1978 SC 597). This shift imported something functionally close to substantive due process into Indian constitutional law and, in doing so, subjected preventive detention to a standard of procedural fairness considerably closer to that developed by the U.S. Supreme Court in *Salerno*, albeit through a different doctrinal route.

Khudiram Das v. State of West Bengal addressed the specific evidentiary structure of preventive detention: the detaining authority's "subjective satisfaction" that detention is necessary. The Supreme Court held that while such satisfaction is inherently subjective and courts will not substitute their own assessment of the adequacy of the grounds for that of the detaining authority, the satisfaction is not wholly immune from judicial review; courts retain the power and the duty to examine whether all the basic facts and materials that in fact influenced the detaining authority were communicated to the detainee, and whether the detaining authority considered only relevant material rather than extraneous or improper considerations (Khudiram Das v. The State of West Bengal & Ors., 26 November 1974). This holding is analytically significant for the argument developed here because it establishes, within Indian preventive detention law, the same distinction between output review and process review that later characterised the American courts' more cautious approach to reviewing algorithmic risk scores: just as a court applying *Loomis* cannot interrogate COMPAS's internal weighting but can insist on procedural safeguards surrounding its use, an Indian court applying *Khudiram Das* cannot substitute its judgment for the detaining authority's subjective assessment of dangerousness but can insist that the process by which that assessment was reached meet defined standards of disclosure and relevance.

Justice K.S. Puttaswamy v. Union of India, in which a nine-judge bench unanimously recognised privacy as a fundamental right traceable to Articles 14, 19, and 21, is not a preventive detention case but establishes doctrinal machinery of direct relevance to it: any state action restricting a recognised fundamental right must satisfy a three-part proportionality test requiring legality, a legitimate state aim, and proportionality between the means adopted and the end pursued (Global Freedom of Expression, 2019). Because preventive detention self-evidently restricts personal liberty, and because AI-assisted dangerousness assessment would additionally implicate privacy interests in the personal and behavioural data such systems require, *Puttaswamy*'s proportionality framework supplies the most readily available doctrinal vehicle through which Indian courts could scrutinise the introduction of algorithmic prediction

into a preventive detention architecture historically reviewed under the more deferential subjective-satisfaction standard of *Khudiram Das*.

The introduction of artificial intelligence into this architecture would not, on the analysis developed here, present a categorically new constitutional problem so much as an intensified version of the existing one. Preventive detention has always operated as a prediction-based legal mechanism, resting on a detaining authority's assessment of future dangerousness rather than proof of a completed offence; what AI would change is the method by which that prediction is generated and the degree to which it can be meaningfully reviewed. A human detaining authority's subjective satisfaction, however imperfect, can at least in principle be interrogated through the disclosure obligations recognised in *Khudiram Das*. An algorithmic risk classification generated by a proprietary or technically opaque system risks being effectively unreviewable in the same way that COMPAS's internal weighting proved unreviewable in *Loomis*, converting a constitutionally uneasy but at least partially transparent form of executive prediction into one that satisfies neither the fairness standard of *Maneka Gandhi* nor the disclosure standard of *Khudiram Das*, while offering no compensating gain in actual predictive reliability given the psychological limits documented in Part II and the empirical bias findings documented in Part III.

V. Constitutional and Ethical Challenges of Predictive Justice

Predictive justice inverts a foundational structural premise of criminal law: that liberty may be restricted only upon proof, to a defined evidentiary standard, of conduct already committed. Where dangerousness assessment substitutes for proof, probability replaces proof and future conduct replaces past conduct as the operative basis for state action, a substitution with consequences that extend well beyond the specific procedural context of preventive detention. The presumption of innocence, understood as the principle that a person is not to be treated as culpable until guilt is established through fair process, sits uneasily beside a system that authorises restraint on the basis of a statistical likelihood of future wrongdoing, because the very concept of a risk score presupposes a population-level base rate that says nothing conclusive about the specific individual being assessed, yet is applied to that individual as though it did. This tension is not unique to India: it is precisely the tension Justice Marshall identified in his *Salerno* dissent, and it is the tension implicit in the elaborate warning label the Wisconsin Supreme Court required to accompany every COMPAS score used at sentencing.

Algorithmic discrimination raises the equality dimension of this same problem into sharper relief. Article 14 guarantees equality before the law and equal protection of the laws, and the empirical findings discussed in Part III, particularly the racially disparate false-positive and false-negative rates documented in the ProPublica and *Science Advances* studies of COMPAS, demonstrate that algorithmic risk-assessment tools can reproduce and formalise pre-existing discriminatory patterns in policing and prosecution data under the appearance of neutral, quantified objectivity (Angwin et al., 2016; Dressel & Farid, 2018). In the Indian context, where preventive detention statutes already operate through subjective executive satisfaction rather than judicially tested proof, and where documented patterns of disproportionate criminal-justice engagement with particular religious, caste, and economic groups exist independently of any algorithmic tool, the introduction of AI-assisted risk assessment carries a specific and foreseeable risk: that historical patterns of profiling, rather than being corrected by the apparent objectivity of a computational system, would instead be laundered through it, rendering discriminatory outcomes harder rather than easier to identify and challenge. Algorithmic fairness, understood as a technical design goal, cannot resolve this problem unilaterally, because fairness metrics themselves involve contested value trade-offs — for instance, between equalising false-positive rates across groups and equalising predictive accuracy across groups — that are properly constitutional and political questions, not merely engineering ones (Dressel & Farid, 2018).

Liberty, privacy, and due process considerations converge most directly on the question of explainability. Article 21's guarantee of personal liberty, read through *Maneka Gandhi*'s requirement that any procedure depriving a person of liberty be fair, just, and reasonable, and through *Khudiram Das*'s requirement that the basic facts and materials influencing a detention decision be disclosed to the detainee, is difficult to reconcile with any predictive tool whose internal logic cannot be meaningfully explained to the person it is used against. This is not a hypothetical difficulty: it is the precise objection Eric Loomis raised, unsuccessfully, against the use of a proprietary risk score at his sentencing, and it would apply with equal or greater force to any AI-assisted dangerousness assessment used to justify preventive detention under Indian law, given that Indian preventive detention procedure already withholds legal representation before the Advisory Board and permits withholding of material considered contrary to the public interest. Layering an inexplicable algorithmic classification onto a review process that already limits disclosure and denies counsel would compound, rather than mitigate, the due process concerns that *Maneka Gandhi* and *Khudiram Das* were designed to

address. *Puttaswamy*'s proportionality test — requiring legality, a legitimate aim, and proportionality between means and ends — supplies a doctrinal standard against which such compounding could be challenged, but only if courts are prepared to treat algorithmic opacity itself as a proportionality defect rather than as a permissible feature of an otherwise legitimate predictive tool.

Beyond these specific doctrinal concerns lie broader ethical challenges that recur across every jurisdiction examined in this paper. Accountability becomes diffuse when a detention or sentencing decision is attributed jointly to a human decision-maker and an algorithmic recommendation, since each can plausibly point to the other as the operative cause of an erroneous outcome, a dynamic the Harvard Law Review's analysis of *Loomis* specifically identified as a form of mistaken accountability and attribution built into the architecture of algorithmically assisted sentencing (Harvard Law Review, 2017). Transparency is structurally compromised where risk-assessment tools are protected as commercial trade secrets, as COMPAS was, since the proprietary interest of the tool's developer directly conflicts with the defendant's interest in contesting the basis of a decision restricting their liberty. Human dignity is implicated whenever a person is treated primarily as an instance of a statistical category rather than as an individual whose particular circumstances are entitled to individualised consideration, a concern the Wisconsin Supreme Court itself partially acknowledged in requiring courts to treat COMPAS scores as reflecting group data rather than individual assessment. The rule of law is strained wherever the operative standard governing a deprivation of liberty is embedded in code inaccessible to the courts applying it, rather than in law publicly promulgated and judicially interpretable. And the democratic legitimacy of predictive governance more broadly is open to question wherever risk-assessment tools, developed and validated by private technology vendors or opaque executive processes, come to exercise what is functionally a quasi-adjudicative role without having been subject to the legislative scrutiny that governs the statutory preventive detention regimes examined in Part IV.

Predictive justice, taken as a whole, fundamentally challenges constitutional guarantees not because prediction is illegitimate as an input to criminal justice decision-making — risk has always informed bail, sentencing, and parole practice in every jurisdiction examined here — but because it permits restrictions on liberty premised on speculative assessments of future behaviour to substitute for, rather than merely supplement, the established criminal liability that has traditionally been the precondition for such restrictions. The constitutional and ethical

challenge is one of degree and of institutional design: how much weight prediction may bear, through what procedural safeguards, and subject to what standard of review, rather than whether prediction may play any role in criminal justice at all.

VI. Towards a Psychology-Informed Constitutional Framework for Predictive Justice

The analysis developed in this paper points toward a framework organised around a single governing distinction: between psychological probability and legal certainty. Criminal psychology, as documented in Part II, offers probabilistic estimates of risk that are genuinely useful for resource allocation, treatment planning, and supervision design, but that do not and cannot amount to the individualised certainty that has traditionally justified criminal deprivation of liberty. A constitutionally sound predictive justice framework must therefore begin by refusing to treat a risk score, however statistically sophisticated, as functionally equivalent to proof of guilt or as an independent legal finding of dangerousness; it must instead be treated as one input among several into a decision that remains subject to full legal and procedural scrutiny. Re-evaluating dangerousness as a legal standard along these lines would require Indian courts, drawing on the disclosure principles already established in *Khudiram Das*, to extend the same searching inquiry into the basis of any algorithmically informed detention decision that is currently applied to human subjective satisfaction, rather than treating computational output as inherently more authoritative than the human judgment it supplements.

Constitutional safeguards for AI-assisted criminal justice should be built around five interlocking requirements. Human oversight in decision-making must be substantive rather than nominal, meaning that a human decision-maker must retain genuine discretion to depart from an algorithmic recommendation and must be required to state independent reasons for a detention or sentencing decision rather than merely ratifying a machine-generated score, mirroring the Wisconsin Supreme Court's requirement in *Loomis* that a COMPAS score never be the determinative factor in sentencing. Explainable and transparent artificial intelligence should be a precondition for the admissibility of algorithmic evidence in any liberty-restricting decision, such that a tool whose classification logic cannot be disclosed to the affected individual, even in general terms, should not be permitted to inform detention decisions at all, addressing directly the trade-secret opacity that proved decisive in the *Loomis* litigation and that would sit especially uneasily alongside Indian preventive detention's existing public-interest non-disclosure exception. Independent algorithmic audits, conducted by bodies

institutionally separate from both the tool's developer and the agency deploying it, should periodically test any risk-assessment instrument used in the Indian criminal justice system for the kind of demographic disparity documented in the ProPublica and *Science Advances* studies of COMPAS, with adverse findings triggering mandatory recalibration or withdrawal. Periodic review of predictive assessments should be built into any detention regime that relies on dynamic risk factors, recognising that a risk classification generated at one point necessarily degrades as circumstances change, consistent with the dynamic-risk-factor literature discussed in Part II. And judicial scrutiny of algorithmic evidence should be structured through an evidentiary standard specific to computational risk assessment, requiring courts to assess an instrument's validation methodology, its demographic performance, and its degree of explainability before any weight is given to its output, rather than admitting such evidence under the same relatively permissive standards applied to conventional expert testimony.

These safeguards translate into concrete legislative and policy recommendations. Comprehensive regulation of predictive policing technologies, including a statutory registration and validation requirement before any risk-assessment tool may be deployed within Indian criminal justice, would close the current governance gap in which such tools, where used, operate without dedicated legal oversight distinct from the general administrative law principles applicable to executive decision-making. Constitutional standards for algorithmic decision-making, ideally codified through amendment to procedural criminal legislation rather than left to case-by-case judicial development, should require, at minimum, disclosure of the categories of data and variables used by any risk-assessment tool informing a detention or sentencing decision, consistent with the disclosure principle *Khudiram Das* already applies to human subjective satisfaction. Integration of behavioural science into judicial risk assessment should proceed through structured judicial training in the psychological limits of dangerousness prediction documented in Part II, so that judges evaluating either human or algorithmic risk assessments understand the base-rate, false-positive, and dynamic-risk-factor problems inherent in any such assessment rather than treating either clinical authority or computational sophistication as a proxy for reliability. And safeguards against arbitrary preventive detention based on predictive technologies should extend the *Puttaswamy* proportionality framework explicitly to any preventive detention order substantially informed by algorithmic risk assessment, requiring the detaining authority to demonstrate not only that detention serves a legitimate aim but that no less restrictive, non-predictive alternative was reasonably available.

Underlying each of these recommendations is a single normative commitment: criminal psychology should function as a constitutional restraint on predictive state power, illuminating the limits of what dangerousness assessment can reliably establish and thereby constraining how much weight the state may place on it, rather than functioning as a scientific or technological justification for expanding the reach of preventive governance. A framework that inverts this relationship — that treats advances in risk-assessment methodology as grounds for relaxing rather than tightening the procedural safeguards surrounding preventive detention — would repeat, in an intensified and less reviewable form, the very error that *Maneka Gandhi* corrected when it rejected *A.K. Gopalan*'s purely procedural conception of Article 21.

VII. Conclusion

Dangerousness remains an uncertain psychological construct rather than a legally ascertainable fact. The clinical and actuarial risk-assessment literature reviewed in Part II establishes that even the best-validated instruments generate probabilistic estimates bounded by base-rate limitations, dynamic risk factors, and the cognitive biases of whoever interprets them, not the individualised certainty that deprivation of liberty has traditionally required. Neither human judgment nor algorithmic prediction eliminates this uncertainty: the migration from clinical to actuarial to algorithmic risk assessment, traced in Part III, has improved statistical accuracy at the margins while leaving the underlying probabilistic character of the enterprise, and its vulnerability to demographic bias, substantially intact, as the COMPAS litigation and the ProPublica and *Science Advances* research demonstrate.

India's preventive detention regime, examined in Part IV, illustrates the constitutional risks inherent in prediction-based justice more starkly than any hypothetical AI deployment, because it has operated at scale, under express constitutional authorisation, for over seven decades, generating a substantial jurisprudential record — from *A.K. Gopalan*'s narrow textualism, through *Maneka Gandhi*'s substantive due process, to *Khudiram Das*'s qualified review of subjective satisfaction and *Puttaswamy*'s proportionality framework — that traces the same due process anxieties AI-assisted risk assessment now raises in a computational register. Artificial intelligence does not introduce these anxieties for the first time; it amplifies concerns relating to bias, transparency, and accountability that were already latent within human-administered preventive detention, by adding a further layer of technical opacity to a decision-making process whose disclosure obligations were already narrowly drawn.

The theoretical contribution of this paper lies in treating clinical prediction, algorithmic risk assessment, and statutory preventive detention not as separate domains but as a single continuum of prediction-based justice governed by a common set of psychological limits and a common constitutional question: how much weight the state may place on a prediction about the future before that reliance becomes indistinguishable from punishing a person for a crime they have not committed. Its policy implications favour a framework in which behavioural science operates as a constraint on, rather than a licence for, the expansion of preventive state power — requiring explainability and independent audit of any algorithmic tool before it may inform a liberty-restricting decision, extending existing disclosure and proportionality doctrine explicitly to algorithmic evidence, and maintaining meaningful, substantive human decision-making authority at every stage.

Future research should examine how Indian courts and detaining authorities currently use, or might use, algorithmic risk-assessment tools in practice, since this paper's analysis has necessarily proceeded from comparative and doctrinal material given the early stage of such deployment in India; should ask whether the disclosure and proportionality standards developed in *Khudiram Das* and *Puttaswamy* are, in practice, adequate to the distinct opacity problems algorithmic tools present; and should evaluate, empirically and over time, whether the safeguards proposed in Part VI meaningfully reduce demographic disparity in preventive detention outcomes rather than merely formalising existing patterns under new procedural cover. A constitutionally compliant response to the rise of predictive justice requires ensuring that behavioural science and artificial intelligence strengthen, rather than substitute for, the foundational principles of due process, equality, and personal liberty that a prediction, however sophisticated, cannot by itself establish.

References

A.K. Gopalan v. State of Madras, AIR 1950 SC 27 (India).

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*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Appelbaum, P. S., Robbins, P. C., & Monahan, J. (2000). Developing a clinically useful actuarial tool for assessing violence risk. *British Journal of Psychiatry, 176*(4), 312–319. Cambridge University Press. https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/developing-a-clinically-useful-actuarial-tool-for-assessing-violence-risk/

Conservation of Foreign Exchange and Prevention of Smuggling Activities Act, 1974, Act No. 52 of 1974 (India). https://www.indiacode.nic.in/bitstream/123456789/15382/1/the_conservation_of_foreign_exchange_and.pdf

Cornell Law School Legal Information Institute. (n.d.). *United States v. Salerno, 481 U.S. 739 (1987)*. https://www.law.cornell.edu/supremecourt/text/481/739

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580. https://www.science.org/doi/full/10.1126/sciadv.aao5580

Freeman, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in *State v. Loomis*. *North Carolina Journal of Law & Technology, 18*(5), 75–121. https://scholarship.law.unc.edu/ncjolt/vol18/iss5/3

Global Freedom of Expression, Columbia University. (2019). *Puttaswamy v. Union of India (II)*. https://globalfreedomofexpression.columbia.edu/cases/puttaswamy-v-union-of-india-ii/

Harvard Law Review. (2017). State v. Loomis: Wisconsin Supreme Court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review, 130*(5), 1530–1537. https://harvardlawreview.org/print/vol-130/state-v-loomis/

Hunt, A. (2014). From control orders to TPIMs: Variations on a number of themes in British legal responses to terrorism. *Crime, Law and Social Change, 62*(3), 289–321. Springer. https://doi.org/10.1007/s10611-013-9467-5

Justia U.S. Supreme Court Center. (n.d.). *United States v. Salerno, 481 U.S. 739 (1987)*. https://supreme.justia.com/cases/federal/us/481/739/

Khudiram Das v. The State of West Bengal & Ors., (1975) 2 SCC 81 (India, decided 26

November 1974). https://indiankanoon.org/doc/679149/

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law, 7*(2), 409–443.

Maneka Gandhi v. Union of India, AIR 1978 SC 597 (India).

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review, 8*(2), 193–210.

Monahan, J. (1981). *The clinical prediction of violent behavior*. U.S. Government Printing Office.

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Roth, L. H., Grisso, T., & Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. Oxford University Press.

National Security Act, 1980, Act No. 65 of 1980 (India). https://www.mha.gov.in/sites/default/files/2022-08/ISdivII_NSAAct1980_20122018%5B1%5D.pdf

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. RAND Corporation.

ProPublica. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

State v. Loomis, 881 N.W.2d 749 (Wis. 2016). https://law.justia.com/cases/wisconsin/supreme-court/2016/2015ap000157-cr.html

Terrorism Prevention and Investigation Measures Act 2011, c. 23 (UK). https://www.legislation.gov.uk/ukpga/2011/23/contents

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Unlawful Activities (Prevention) Act, 1967, Act No. 37 of 1967 (India), as amended by the Unlawful Activities (Prevention) Amendment Act, 2019.

United States v. Salerno, 481 U.S. 739 (1987).

UK Government. (2011). *Terrorism Prevention and Investigation Measures Act 2011: Explanatory notes*. https://www.legislation.gov.uk/ukpga/2011/23/notes