Assigning Liability for AI Misconduct

By Vanshika Shukla

Abstract

Modern tort doctrines such as negligence, strict liability, and vicarious liability provide foundational tools for addressing harms caused by artificial intelligence (AI), yet they struggle to accommodate the unique features of autonomous, opaque, and continuously evolving systems. This paper employs a combined doctrinal, comparative, and normative methodology to argue that a hybrid liability architecture is the most appropriate pathway for India. Specifically, this paper advocates for: (i) strict liability for physical, safety-critical harms caused by autonomous AI systems; (ii) a calibrated fault-based regime for informational and reputational harms; and (iii) statutory algorithmic duties supported by procedural and institutional reforms. Drawing on Indian constitutional principles, landmark domestic and international jurisprudence, and the EU Artificial Intelligence Act (EU AI Act), the paper identifies structural doctrinal failures in existing law, critically evaluates the applicability of European regulatory models to the Indian legal context, and proposes concrete legislative provisions tailored to India's socio-legal environment.

I. Introduction

Artificial intelligence has rapidly transitioned from laboratory research to pervasive real-world deployment. Autonomous vehicles, predictive risk algorithms, algorithmic hiring systems, and content-recommendation engines now make consequential decisions across sectors that directly affect life, liberty, and property. When these systems malfunction or cause harm, law must answer a deceptively simple question: who is responsible, and under what normative principles?

The urgency of this question is reflected in empirical data. According to the Stanford AI Index Report 2025, documented AI safety incidents surged from 149 in 2023 to 233 in 2024, resulting in a 56.4% increase in a single year, underscoring that the doctrinal failures identified in this paper are not theoretical concerns but a growing enforcement gap demanding immediate legislative attention.

> *Documented AI safety incidents rose by 56.4% between 2023 and 2024 alone (Stanford HAI, AI Index Report 2025). Reports of malicious actors using AI — including deepfakes and AI-enabled disinformation — grew eightfold from 2022 to 2025, with deepfake-related incidents now outnumbering autonomous vehicles, facial recognition, and content-moderation incidents combined (AI Incident Database, 2025).*

India's exposure to these risks is acute and growing. The Indian AI market, valued at approximately USD 9.51 billion in 2024, is projected to reach USD 130.63 billion by 2032 at a compound annual growth rate of approximately 39%[1]. India also ranks second globally in public generative AI projects on GitHub, with approximately 16% of the world's AI talent, and has been identified as the fastest-growing hub for AI developers[2]. These figures illustrate the pace at which AI systems capable of causing harm will be embedded in critical sectors before any comprehensive liability framework is in place, and they make India's regulatory choices consequential well beyond its borders.

Existing legal frameworks in India corresponding to negligence under common law, absolute liability under M.C. Mehta v. Union of India, product liability under the Consumer Protection Act, 2019, and vicarious liability principles were developed in contexts that presuppose human agency and traceable causal chains. AI disrupts each of these assumptions. Its defining features are algorithmic opacity, continuous post-deployment learning, distributed multi-party development, and the capacity for autonomous decision-making that no single human actor directed or anticipated. These features generate what this paper identifies as four doctrinal failures: (i) the foreseeability gap, where emergent AI behaviour defeats the negligence standard; (ii) the causation lacuna, where distributed supply chains obscure the chain of responsibility; (iii) the temporal mismatch, where product liability's 'time-of-sale' framework cannot accommodate post-deployment algorithmic drift; and (iv) the harm-quantification deficit, where intangible informational and reputational harms resist traditional damage assessment.

Comparative jurisdictions have begun responding. The EU AI Act (Regulation (EU) 2024/1689) introduces a risk-based regulatory architecture, and the revised Product Liability Directive extends strict liability to software and evolving AI systems. These developments are

[1]Fortune Business Insights, 2024
[2]PIB, Government of India

instructive but not directly transplantable: India's constitutional structure, the relative weakness of regulatory capacity, and the vastly different socio-economic context of AI deployment demand a tailored response rather than wholesale legal borrowing.

This paper pursues two aims. First, it analyses why existing Indian legal doctrines fail to adequately address AI misconduct, proceeding beyond descriptive summary to identify the precise doctrinal contradictions that demand reform. Second, it proposes a hybrid liability framework operationalised into structured legal provisions; draft statutory sections, institutional mechanisms, and enforcement standards that can guide Indian legislative action.

II. Methodology

This paper employs three interconnected methodological approaches, each directed at a distinct analytical purpose.

First, doctrinal analysis examines the internal coherence of Indian tort law principles such as negligence, strict liability, vicarious liability, and their constitutional dimensions when applied to AI-caused harms. Rather than describing existing doctrine, the analysis proceeds by identifying structural contradictions: points at which the application of existing rules to AI systems either produces incoherent outcomes or leaves identifiable gaps in victim protection.

Second, comparative analysis draws on the EU AI Act and associated liability instruments to identify regulatory techniques that may be instructive for Indian reform. This analysis is explicitly critical: European models are assessed for their applicability to the Indian legal system, with careful attention to differences in constitutional architecture, regulatory capacity, rule-of-law infrastructure, and the scale and nature of AI deployment in Indian public and private sectors.

Third, normative analysis evaluates proposed liability models against four criteria: victim compensation adequacy, innovation-incentive proportionality, institutional enforceability within India's existing administrative framework, and consistency with constitutional rights under Articles 14, 19, and 21. Legislative proposals are assessed against these criteria before they are advanced as recommendations.

Primary sources include Indian statutes and judicial decisions, EU legislative instruments, and regulatory guidance. Secondary sources include peer-reviewed legal scholarship, industry

market analyses, and policy studies including reports by Stanford HAI, NASSCOM, Fortune Business Insights, IMARC Group, the AI Incident Database, and the Urban Institute. The paper deliberately limits reliance on U.S. case law to instances where those decisions illuminate doctrinal questions not resolved by Indian or European authority.

III. Doctrinal Analysis: Where Existing Law Fails

A. Negligence and the Foreseeability Gap

Negligence in Indian tort law requires establishing duty of care, breach, causation, and damage. Applied to AI systems, this framework encounters its first structural failure at the foreseeability stage. In Jacob Mathew v. State of Punjab[3], the Supreme Court affirmed that breach of duty requires proof that the defendant failed to exercise the degree of care that a reasonable person in their position would have exercised. Deep-learning models sometimes act in ways their creators didn't expect. This means a developer could do everything right and still end up with a system that causes harm.

In Donoghue v. Stevenson[4], the neighbour principle established that manufacturers owe a duty to those foreseeably affected by their products. AI systems complicate this in two respects. First, AI decisions affect a much wider and more unpredictable group of people than the simple relationship between a manufacturer and a consumer seen in the Donoghue case. This includes everyone from loan applicants and job seekers to pedestrians in the path of a self-driving car. Second, the foreseeability of specific harms is negated precisely by the opacity of the model: a developer cannot foresee the particular discriminatory output of a neural network trained on historical data if the discriminatory pattern is encoded in latent feature correlations invisible to human inspection.

This foreseeability gap does not simply raise the evidentiary bar for negligence claims; it defeats the conceptual premise of fault-based liability in cases involving high-complexity AI systems. Fault-based liability remains meaningful, however, in cases of informational, reputational, or moderate-risk AI harms where the decision-making process remains partially

[3] Jacob Mathew v. State of Punjab (2005) 6 SCC 1
[4] Donoghue v. Stevenson [1932] AC 562

interpretable and where developers can reasonably be expected to have foreseen categories of harm through impact assessments.

B. The Causation Lacuna in Distributed AI Systems

The second structural failure concerns causation. Traditional but-for causation, as understood in Indian courts following Executors of the Estate of T.R. Venkatarama Iyer v. State of Tamil Nadu[5], requires the plaintiff to establish that the defendant's act or omission was a necessary condition of the harm. Modern AI systems are developed across layered supply chains: data curators provide training datasets, researchers develop base model architectures, platform integrators fine-tune and deploy models, and end-use operators configure and operationalise them. When something goes wrong, like a loan being unfairly denied, it's hard to pinpoint who is responsible. Because an algorithm is involved, the actual cause of the harm becomes almost impossible to untangle.

This is not merely an evidentiary problem that better disclosure rules can resolve; it is a structural one. The harm may be the emergent product of interactions between the training data distribution, the model architecture, the fine-tuning process, and the deployment context, none of which was individually sufficient or necessary to produce the precise output. Indian courts have extended the material contribution test in complex industrial causation cases (see the reasoning in Syad Akbar v. State of Karnataka[6], but even this more flexible standard requires the plaintiff to identify a defendant whose contribution materially increased the risk of harm. In distributed AI supply chains, this attribution remains deeply contested.

The scale of the causation problem is already visible in autonomous vehicle data. NHTSA reported 975 automated driving system incidents in 2025, up from 526 in 2024, over a period during which AV deployment mileage also doubled from 75 million to approximately 145 million miles (Autonomous Vehicle Industry Association, 2025). Of the 2,052 AV incidents reported to the NHTSA that contain narratives, autonomous vehicles were solely at fault in only 4% of cases involving other road users, and hardware or software failures accounted for just 7.8% of at-fault incidents[7]. The vast majority of these issues pop up because of unpredictable interactions between the AI and the world around it. This is a problem for our

[5] T.R. Venkatarama Iyer v. State of Tamil Nadu AIR 1972 SC 1314
[6] Syad Akbar v. State of Karnataka (1979) 4 SCC 175)
[7]NHTSA ADS Standing General Order Data, analysed 2026

current laws, because it doesn't fit the patterns that negligence or strict liability were designed to handle.

> *NHTSA ADS incident reports: 526 incidents (2024) → 975 (2025), a near-doubling in one year. AVs were solely at fault in only 4% of incidents involving other road users. Hardware/software failures accounted for 7.8% of at-fault incidents. (NHTSA Standing General Order; Autonomous Vehicle Industry Association, 2025.)*

C. Product Liability's Temporal Mismatch

The Consumer Protection Act, 2019, and common law product liability principles examine a product's defective condition at or proximate to the time of manufacture or sale. This snapshot approach produces a fundamental temporal mismatch when applied to AI systems that continue learning after deployment. A model deployed without any latent defect may, through reinforcement learning or periodic retraining on new data, develop discriminatory or harmful behaviour months after deployment. Under the current statutory framework, there is no product to which liability can attach at the moment of harm because the 'product' at the time of harm is materially different from the product at the time of sale.

This gap is not hypothetical. Take the 2025 Florida Tesla Autopilot verdict, where a jury awarded over USD 240 million after a fatal crash. In that case, investigators had to look at software updates that changed how the car behaved between the time it was bought and the day of the accident. Indian product liability law, as currently codified, provides no equivalent mechanism for anchoring liability to the system as it existed at the time of the harmful act, nor any mechanism for allocating responsibility across parties who contributed to post-sale modifications.

D. Intangible Harms and the Quantification Deficit

The fourth area where legal doctrines fail is in how they categorize harm. Physical injuries from AI such as a fatal autonomous vehicle crash, actually fit quite well into the traditional frameworks we already use for personal injury. However, AI systems increasingly cause informational and reputational harms: discriminatory exclusion from employment or credit, wrongful criminal risk profiling, or reputational damage through algorithmic content

distribution. These harms are real and constitutionally cognisable under Articles 14 and 21, but they resist reduction to the pecuniary damage frameworks that Indian courts typically apply.

The real-world magnitude of informational harm from algorithmic systems is illustrated by lending discrimination data. A 2024 Urban Institute analysis of Home Mortgage Disclosure Act data found that minority borrowers were more than twice as likely to be denied credit than comparably situated white applicants even in algorithmically mediated systems. A parallel UC Berkeley study found that algorithm-driven pricing systems in fintech lending imposed higher interest rates on African American and Latinx borrowers amounting to an estimated USD 450 million in excess interest annually (Urban Institute, 2024; UC Berkeley, cited in Kennedy Human Rights Center, 2025). These findings are directly relevant to India: the Indian BFSI sector's AI market reached USD 830 million in 2024 and is projected to approach USD 8.09 billion by 2033 at a CAGR of 28.8% (IMARC Group, 2024), creating a rapidly expanding context in which comparable discriminatory harms may materialise absent statutory bias-impact obligations.

> *India AI in BFSI: USD 830 million (2024) → projected USD 8.09 billion (2033), CAGR 28.8% (IMARC Group, 2024). In the U.S., algorithmic lending discrimination imposed an estimated USD 450 million in excess annual interest costs on minority borrowers (UC Berkeley; Urban Institute, 2024) — a pattern enabled by the same black-box systems increasingly deployed in Indian financial services.*

The absence of clear statutory damage frameworks for algorithmic harms produces a chilling effect on civil litigation. Plaintiffs who suffer quantifiable but difficult-to-prove harm from algorithmic discrimination face both the difficulty of establishing causation and the uncertainty of quantum. Without statutory minimum damages or presumed-harm provisions, courts tend to apply conservative compensation standards, which weakens the deterrent function of tort liability.

This distinction between physical and informational harms has a direct bearing on liability standard selection. Physical harms from safety-critical AI warrant strict liability because the magnitude of potential harm, the difficulty of victim proof, and the enterprise's capacity to invest in prevention jointly justify removing fault as an element.Informational harms should be judged based on fault rather than strict liability. If we're too harsh, we risk discouraging the

development of helpful AI used in public health, education, and the legal system. This is especially true when a harm is real but couldn't have been predicted or prevented, even with proper care. The legal test for the boundary between these categories is proposed in Section V below.

IV. Comparative Frameworks: Critical Evaluation of Applicability to India

A. The EU AI Act: Risk Classification and Its Indian Relevance

The EU AI Act establishes a risk-tiered regulatory framework: prohibited AI practices (Article 5), high-risk AI systems subject to conformity assessments and registration obligations (Annex III), and general-purpose AI models subject to transparency requirements (Article 53). The Act's risk-based architecture is conceptually well-suited to a hybrid liability framework: regulatory obligations attach at the design stage, creating documented standards of care that can later inform fault-based liability claims.

However, three features of the EU model limit its direct applicability to India. First, the EU AI Act relies on a complex regulatory system that India hasn't built yet. This includes specialized groups like the European AI Office, national oversight authorities, and bodies that certify whether a system is safe. Requiring Indian developers to comply with EU-equivalent conformity assessment requirements before deployment would impose costs that are disproportionate relative to India's current regulatory capacity, risking both regulatory evasion and the stifling of domestic AI development.

Second, the EU model was designed for a single internal market with harmonised standards. India's federal structure, the variation in State-level enforcement capacity, and the complexity of regulating AI deployed by public-sector entities including State governments, public sector undertakings, and judicial bodies, require a more differentiated enforcement architecture than the EU model contemplates.

Third, and most critically, the EU AI Act operates primarily as a product-safety and market-regulation instrument; it does not directly create civil liability. The proposed AI Liability Directive, which would have facilitated tort claims by introducing evidentiary presumptions, was stalled and its scope significantly narrowed in the legislative process. India should

therefore not assume that adopting an EU-style regulatory framework will automatically facilitate civil claims; the liability-enabling provisions must be legislated independently.

The compliance data from the EU's own voluntary frameworks reinforce this concern. A McKinsey survey on responsible AI engagement found that while majorities of organisations identified key risks like inaccuracy (64%), regulatory compliance (63%), and cybersecurity (60%), yet significant proportions were not actively addressing them (McKinsey, cited in Stanford HAI AI Index Report 2025). Voluntary compliance is often incomplete, even in places with much stronger oversight than India. This highlights why India needs to build mandatory duties and clear enforcement into its laws, rather than just relying on broad principles or high-level guidance.

B. What India Can Adapt

Notwithstanding these limitations, three features of the EU framework are worth adapting. First, the risk-classification logic: distinguishing between systems deployed in high-stakes contexts (autonomous vehicles, medical devices, judicial decision-support, credit scoring) and lower-stakes applications, provides a principled basis for differentiating the intensity of liability and regulatory obligations. This classification should be grounded in Indian constitutional rights: AI systems deployed by the State in ways that engage Articles 14, 19, or 21 should automatically be treated as high-risk regardless of the technological complexity of the system involved.

Second, the EU's technical documentation and incident-logging obligations (Articles 11 and 62 of the EU AI Act) provide a model for creating the evidentiary infrastructure that Indian courts currently lack when adjudicating AI claims. Making operational logs a legal requirement instead of just a suggestion does two things: it encourages companies to prevent mistakes before they happen and provides a clear trail of evidence if a case ever goes to court.

Third, the concept of rebuttable presumptions developed in the proposed AI Liability Directive and partially operationalised in the EU Product Liability Directive offers an evidentiary mechanism that Indian courts can adopt through statutory reform. Where a deployer fails to comply with mandatory logging or risk-assessment obligations, the court should be empowered to presume that the non-disclosed information would have been adverse to the deployer's case.

It is also worth noting that voluntary transparency improvements are occurring within industry. The Foundation Model Transparency Index found that average transparency scores among

major model developers increased from 37% in October 2023 to 58% in May 2024[8]. While this is a meaningful improvement, a residual opacity of nearly 42% in even the most scrutinised models underscores that transparency norms alone cannot substitute for the structured evidentiary framework this paper proposes. Voluntary progress and mandatory duty are complements, not substitutes.

V. A Hybrid Liability Framework for India: Structured Legal Provisions

This section advances the paper's central normative argument: that India should adopt a hybrid liability architecture operationalised through specific statutory provisions. The architecture rests on three pillars, each addressing a distinct category of AI-caused harm.

Pillar 1: Strict Liability for Physical Harms from Autonomous Systems

Building on the absolute liability doctrine established in M.C. Mehta v. Union of India[9], this paper proposes a statutory provision along the following lines:

Draft Provision 1 - Absolute Liability for High-Risk Autonomous AI Systems: Any person who deploys, operates, or places in service a high-risk autonomous AI system in a public context shall be absolutely and vicariously liable for all physical harm, death, or grievous injury caused by or through the operation of that system, without proof of fault. No defence of contributory negligence, act of God, or technical compliance shall extinguish liability. Liability shall not be reduced by the conduct of the plaintiff unless the plaintiff intentionally caused the harm.

The definition of 'high-risk autonomous AI system' should include: (i) autonomous or semi-autonomous vehicles capable of operating without continuous human control; (ii) AI systems deployed in medical devices capable of initiating or recommending clinical interventions; (iii) AI systems deployed in industrial machinery operating in proximity to human beings; and (iv) AI systems used in judicial or quasi-judicial risk-assessment, bail, or sentencing recommendations.

[8] Stanford HAI AI Index Report 2025
[9] M.C. Mehta v. Union of India (1987) 1 SCC 395

This strict liability standard is justified by three convergent considerations. First, the magnitude of potential harm in safety-critical deployments is sufficiently catastrophic that the social interest in deterrence outweighs the developer's interest in being protected from liability for unforeseeable failures. Second, in these domains, the developer or deployer is better positioned than the victim to invest in risk-reduction and to obtain liability insurance. Third, the opacity of AI systems makes fault-based proof practically unavailable to victims who lack access to training data, model architecture, or operational logs.

Pillar 2: Fault-Based Liability with Statutory Duties for Informational Harms

Instead of a one-size-fits-all approach, I'm proposing a fault-based regime for harms like discriminatory hiring or unlawful profiling. By turning algorithmic duties into law, we can address these specific informational and reputational risks more effectively.

The applicable legal test for distinguishing information from physical harm should be: does the AI system's output operate directly on the physical world, or does it operate through the mediation of a human decision-maker who retains the formal authority to act upon or override the AI's recommendation? Where the AI output directly causes physical consequence (an autonomous vehicle brakes or accelerates; a drug-dispensing machine administers medication), the physical harm standard applies. Where the AI output is mediated by a human decision (a loan officer reviews an AI credit score; a hiring manager considers an algorithmic ranking), the informational harm standard applies, and fault must be established.

Draft Provision 2 - Statutory Algorithmic Duties: Any person who deploys an AI system that makes or materially influences decisions affecting the legal rights, financial interests, or reputational standing of another person shall: (a) maintain structured operational logs recording all inputs, model versions, decision outputs, and human override events for a minimum period of five years; (b) conduct and document a bias and fairness impact assessment prior to deployment and following any material update; (c) provide, upon request of any affected person, a meaningful explanation of any automated decision that adversely affects that person, in plain language accessible to a non-specialist; and (d) report material incidents causing harm to an identified national regulatory authority within 72 hours of discovery.

Draft Provision 3 - Rebuttable Presumption: Failure to comply with any obligation under Provision 2 shall give rise to a rebuttable presumption that the defendant's AI system caused or materially contributed to the harm alleged. The defendant may rebut this presumption by

adducing evidence that the non-disclosed or non-documented information would not have altered the outcome.

These provisions are calibrated to the Indian litigation context. Right now, victims often have to hire expensive experts just to prove an AI was at fault, which is simply too costly for most people. My framework shifts that burden, requiring the company to provide the documentation or prove they weren't responsible if records are missing.

Pillar 3: Institutional and Enforcement Architecture

Substantive liability rules are effective only to the extent that enforcement mechanisms are functional. This paper identifies three institutional reforms as essential to the operationalisation of the proposed framework.

First, a designated AI Liability and Algorithmic Accountability Authority should be established, either as a standalone statutory body or as a division of an existing regulator such as the Information Commission or the proposed Data Protection Board under the Digital Personal Data Protection Act, 2023. This Authority should have power to: (i) maintain a public register of high-risk AI systems; (ii) investigate complaints of algorithmic harm; (iii) impose administrative penalties (calibrated to deployer revenue, not fixed absolute amounts, to ensure proportionate deterrence across firms of different sizes); and (iv) issue compliance guidance.

Second, specialised AI benches should be established within the High Courts of major jurisdictions, modelled on the Commercial Courts Act, 2015 structure, to adjudicate civil claims arising from AI-caused harm. These benches should be empowered to appoint court-expert technical advisors under section 45 of the Indian Evidence Act, 1872 (or its successor under the Bharatiya Sakshya Adhiniyam, 2023) to assist in the interpretation of algorithmic evidence. Technical evidence disclosed in these proceedings should be subject to in-camera review and protective orders where necessary to safeguard legitimate trade secrets.

Third, mandatory liability insurance should be required for deployers of high-risk AI systems, with minimum coverage amounts prescribed by the Authority and scaled to the scale and risk profile of the system. Insurance requirements create an immediate market incentive for risk reduction, ensure that victims can obtain compensation even where corporate defendants are insolvent, and generate actuarial data that over time informs the calibration of regulatory standards.

VI. Indian Constitutional and Jurisprudential Grounding

The proposed framework is not advanced in isolation from Indian legal tradition; it is grounded in existing constitutional doctrine and emerging judicial engagement with algorithmic systems.

The 'absolute liability' rule from the 1987 M. C. Mehta[10] case is the foundation for the first pillar of my proposal. In that ruling, the Supreme Court decided that any business involved in dangerous activities is fully responsible for any harm they cause, with no exceptions allowed. The Court's reasoning was explicitly enterprise-liability in character: the enterprise which creates the risk should bear the cost of the harm that materialises from it. Autonomous AI systems deployed in safety-critical public contexts are paradigmatic cases to which this reasoning applies.

The right to reasoned decisions implicit in Article 21's 'procedure established by law' standard is directly engaged by AI systems that make or influence decisions affecting life or personal liberty. In Maneka Gandhi v. Union of India (1978) 1 SCC 248, the Supreme Court held that the procedure established by law must be fair, just, and reasonable. An automated decision affecting the liberty of a person, made by a system whose internal logic is mathematically insulated from scrutiny, cannot satisfy this standard. Indian courts have yet to directly apply this reasoning to private-sector AI systems, but the constitutional principle is broad enough to encompass State-contracted or State-deployed AI decision-support tools.

The right to equality under Article 14 is engaged by AI systems that reproduce or amplify historical patterns of discrimination. In the Navtej Singh Johar case[11], the Court confirmed that a rule can violate the right to equality even if it seems neutral on the surface. If the rule's actual effect unfairly hits a protected group in a significant way, it can be ruled unconstitutional under Article 14. Algorithmic systems trained on historically biased datasets may constitute precisely this form of disparate-impact discrimination. Even the most advanced AI systems struggle with implicit bias. According to the Stanford AI Index Report 2025, several widely used commercial models still show racial and gender prejudice, often linking negative words to minorities or suggesting men are better suited for leadership, even after safety measures were put in place. If such patterns persist in frontier models designed for global markets, the risk of similar bias in AI systems trained on India's historically stratified social and economic data is substantial.

[10] M.C. Mehta v. Union of India (1987) 1 SCC 395
[11] Navtej Singh Johar v. Union of India (2018) 10 SCC 1

The proposed statutory duty to conduct bias impact assessments (Provision 2(b)) operationalises this constitutional requirement in the AI context.

Indian courts have not yet decided a case squarely on AI liability, but the judiciary's engagement with algorithmic systems is beginning. The Supreme Court's directions in Arnesh Kumar v. State of Bihar case[12] which limited routine arrest powers and requiring reasoned satisfaction before arrest reflected a jurisprudential instinct toward requiring articulable justification for decisions affecting liberty. Extending this instinct to algorithmic bail, risk-scoring, and sentencing tools is the natural next doctrinal step, and the proposed framework for specialised AI benches and expert-assisted evidence review provides the institutional mechanism to operationalise it.

VII. Addressing Enforcement Challenges

The practical enforceability of the proposed framework is as important as its theoretical coherence. This section identifies the primary enforcement challenges and proposes mitigation measures tailored to India's regulatory and judicial context.

The first challenge is regulatory capacity. India does not currently have an AI-specific regulator with technical capacity, jurisdiction, and independence. The proposed AI Liability and Algorithmic Accountability Authority must be adequately staffed and funded from inception; a body with mandate but without capacity will generate expectations it cannot meet, undermining public trust in the regulatory framework. To manage technical duties like system audits and compliance guidance, the Authority should collaborate with universities. This ensures they have access to expert research in areas where the government lacks its own specialists.

The second challenge concerns jurisdiction over foreign AI systems. Many of the AI systems deployed in India that affect Indian users and rights are developed and operated by entities incorporated outside India. The proposed framework should include a clear deemed-deployment rule: any AI system that makes or influences decisions affecting persons in India, regardless of where the system is developed or operated, shall be subject to Indian liability law to the extent permitted by applicable private international law principles. Deployers without a

[12] Arnesh Kumar v. State of Bihar (2014) 8 SCC 273

registered Indian presence who generate revenue from Indian users should be required to appoint an authorised representative for regulatory and litigation purposes, modelled on the requirement in Article 27 of the EU AI Act.

The third challenge is litigation cost and access. Even a well-designed liability framework fails if victims cannot practically access courts. The statutory rebuttable presumption mechanism (Provision 3) addresses the evidentiary cost barrier by shifting the burden of production to the deployer. The establishment of specialised AI benches should be accompanied by procedural rules providing for expedited case management and, in appropriate cases, cost-shifting to defendants where the claim involves a systemic harm affecting multiple persons. Representative actions by consumer organisations or civil society bodies should be expressly permitted for systemic algorithmic harms, modelled on the representative complaint mechanism in India's consumer protection framework.

The fourth challenge is the risk of regulatory capture. A sectoral AI regulator that is too closely aligned with industry may calibrate standards in ways that protect deployers rather than victims. The Authority's governance structure should include independent members drawn from civil society, the legal profession, and academia, with mandatory public consultation requirements for major guidance and standard-setting activities.

VIII. The Black Box Problem: Doctrinal and Technical Dimensions

The preceding sections have identified the black-box phenomenon as a common thread running through all four doctrinal failures. This section analyses the doctrinal and technical dimensions of this problem in greater depth, with the aim of demonstrating why technical solutions alone are insufficient and why the proposed legal framework is necessary.

Modern deep-learning models operate in high-dimensional feature spaces employing non-linear transformations across multiple layers. Unlike classical algorithmic logic, which maps inputs to outputs through traceable conditional steps, neural network outputs are the product of billions of weighted parameter interactions that are not human-readable. This structural opacity creates what this paper terms a causal blank: even the original creators cannot fully articulate why a particular output was produced. This is not a temporary limitation of current technology that better engineering will resolve; it is, to a significant degree, a constitutive feature of the statistical methods that underlie the most capable contemporary AI systems.

The persistence of algorithmic bias even in the most extensively evaluated models illustrates the limits of technical self-regulation as a substitute for statutory duty. Research cited in the Stanford AI Index Report 2025 found that leading large language models continued to exhibit implicit biases along race and gender lines despite measures designed to mitigate them. Furthermore, documented AI safety incidents rose by 56.4% between 2023 and 2024 to reach a record 233 incidents, and the AI Incident Database records an eightfold growth in malicious-use incidents including deepfakes and AI-enabled disinformation from 2022 to 2025, with deepfake incidents now outnumbering AV, facial recognition, and content-moderation incidents combined. These trajectories collectively demonstrate that voluntary technical improvements and industry self-regulation, while valuable, have not kept pace with the expansion of AI-caused harm.

> *Stanford AI Index Report 2025: documented AI safety incidents = 233 in 2024 (record high; +56.4% from 2023). AI Incident Database: malicious-use AI incidents grew 8x from 2022–2025; deepfake incidents now exceed AV + facial recognition + content-moderation incidents combined. Foundation Model Transparency Index: average transparency score rose from 37% (Oct 2023) to 58% (May 2024) — meaningful progress, but a residual opacity of ~42% in the most scrutinised models confirms that voluntary transparency is insufficient substitute for statutory evidentiary duties.*

The legal consequences are direct. In State v. Loomis[13], the Wisconsin Supreme Court permitted the use of the proprietary COMPAS actuarial risk tool in criminal sentencing while acknowledging that its internal mechanics were undisclosed. The court conditioned its approval on the requirement that COMPAS scores not be used as the determinative factor, but the practical effect was to allow an opaque algorithmic input to influence a decision directly affecting the liberty of the defendant. In India, equivalent dependence on undisclosed algorithmic scoring in bail or sentencing applications would constitute a prima facie violation of the Article 21 requirements elaborated in Maneka Gandhi, because the 'procedure established by law' cannot be fair and reasonable if it incorporates a reasoning process that is structurally immune from scrutiny.

Technical approaches to explainability; LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations) values, for example- offer post-hoc

[13] State v. Loomis, 881 N.W.2d 749 (Wis. 2016)

approximations of decision logic. These are useful for generating plausible accounts of a model's behaviour but do not provide access to the model's mechanical reasoning. In high-stakes legal contexts, a plausible account is insufficient; what is required is a legally cognisable record of the basis for the decision. This is precisely the function served by the AI passport and operational logging requirements proposed in this paper's Pillar 2: they do not eliminate opacity but create a structured evidentiary record that courts can interrogate using expert assistance.

IX. Synthesis of Findings

The doctrinal, comparative, and normative analyses conducted in this paper support the following conclusions.

Existing Indian doctrines provide partial coverage at best. Negligence is applicable in principle but fails at the foreseeability and causation stages for high-complexity AI systems. Product liability under the Consumer Protection Act, 2019 applies to AI embedded in tangible goods but cannot accommodate AI-as-a-service models or systems that evolve post-deployment. Vicarious liability supports holding deployers accountable where they exercise operational control but does not resolve the challenge of distributed supply-chain causation. These are not gaps that creative judicial interpretation can fully close; statutory reform is required.

Strict liability is constitutionally grounded and empirically appropriate for safety-critical AI harms. The absolute liability doctrine in M.C. Mehta provides the necessary foundation. The 2025 Florida Tesla Autopilot verdict, though decided under U.S. law, demonstrates that senior courts are prepared to impose substantial liability for autonomous system failures even where precise causal attribution is contested.

Regulatory duties facilitate civil claims but cannot substitute for them. The EU AI Act's experience demonstrates that preventive regulatory obligations create documented standards of care that subsequently ease fault-based liability claims. India needs to put tools like logging requirements and incident reporting directly into its AI laws. We can't just assume that a government regulator will be enough to help people get justice on their own.

Procedural and institutional reform is indispensable. The decisions in Loomis and Bridges case[14] both highlight the degree to which existing judicial processes are ill-equipped to manage algorithmic evidence. Specialised benches, expert-assisted review, and algorithmic discovery protocols are not optional enhancements; they are prerequisites for the liability framework to function.

X. Conclusion and Legislative Recommendations

AI liability reform in India requires three simultaneous interventions: doctrinal evolution within the existing tort framework, statutory codification of new liability rules and algorithmic duties, and the creation of institutional infrastructure capable of implementing and enforcing those rules. A hybrid liability regime calibrated to harm type, supported by robust procedural mechanisms and grounded in constitutional rights, offers the most principled and practically achievable path forward.

The empirical record underscores why reform cannot be deferred. India's AI market is projected to grow nearly fourteenfold, from USD 9.51 billion in 2024 to USD 130.63 billion by 2032, while global documented AI incidents have grown at 56.4% per year and malicious-use incidents have increased eightfold since 2022. A regulatory gap of this size, in a market of this scale, represents a predictable source of compounding harm that reactive tort adjudication will be unable to address at speed or scale.

This paper recommends that the Indian Parliament enact a dedicated AI Liability and Algorithmic Accountability Act. The Act should contain the following principal provisions:

- Section 1 - Absolute Liability: Deployers of high-risk autonomous AI systems in safety-critical public contexts shall be absolutely liable for physical harm, without proof of fault and without the defences available in common law strict liability, substantially along the lines of Draft Provision 1 above.
- Section 2 - Statutory Algorithmic Duties: Deployers of AI systems that make or materially influence decisions affecting legal rights, financial interests, or reputational standing shall comply with operational logging, bias impact assessment, explainability, and incident-reporting obligations as specified in Draft Provision 2.

[14] Loomis and Bridges v. Chief Constable of South Wales Police [2020] EWCA Civ 1058

- Section 3 - Rebuttable Presumption: Failure to comply with Section 2 obligations shall give rise to the rebuttable presumption specified in Draft Provision 3.
- Section 4 - Mandatory Insurance: Deployers of high-risk AI systems shall maintain liability insurance at levels prescribed by the AI Liability and Algorithmic Accountability Authority.
- Section 5 - AI Liability and Algorithmic Accountability Authority: The Act shall establish the Authority with powers of investigation, standard-setting, administrative penalty, and public register maintenance, with governance structures ensuring independence from industry.
- Section 6 - Specialised AI Benches: The Act shall authorise the establishment of specialised AI liability benches within the High Courts, with powers to appoint court-expert technical advisors and to make algorithmic discovery orders under protective conditions.
- Section 7 - Representative Actions: Consumer organisations and civil society bodies meeting criteria prescribed by the Act may bring representative claims on behalf of groups of persons affected by systemic algorithmic harms.
- Section 8 - Deemed Deployment and Foreign Deployers: AI systems affecting persons in India shall be subject to this Act regardless of where they are developed or operated. Foreign deployers without an Indian registered presence shall appoint an authorised representative as a condition of operating in India.

This framework preserves the space for innovation by calibrating strict liability to contexts where the risk-benefit calculus clearly justifies it, while ensuring that victims of informational harms have access to a practically functional remedy. It grounds reform in existing constitutional doctrine rather than importing foreign frameworks wholesale. And it provides the enforcement mechanisms like the Authority, the specialised benches, the mandatory insurance requirement, without which even the most elegantly drafted substantive rules remain aspirational.

India’s evolving AI governance landscape presents a genuine policy window for structured legal reform. The framework proposed here is designed to be both principled in its doctrinal foundations and realistic in its institutional demands. The question is no longer whether reform is necessary; it is whether the legal system will act before the harms it is designed to prevent become irreversible.

Bibliography

A. Primary Legal Sources

- Anderson v. TikTok, Inc. (2024) 3rd Cir., No. 22-3061 (United States Court of Appeals for the Third Circuit).
- Arnesh Kumar v. State of Bihar (2014) 8 SCC 273 (Supreme Court of India).
- Bharatiya Sakshya Adhiniyam 2023 (India).
- Commercial Courts Act 2015 (India).
- Consumer Protection Act 2019 (India).
- Digital Personal Data Protection Act 2023 (India).
- Directive (EU) 2024/... on liability for defective products (Revised Product Liability Directive).
- Donoghue v. Stevenson [1932] AC 562 (House of Lords).
- Gonzalez v. Google LLC (2023) 598 U.S. ___ (Supreme Court of the United States).
- Indian Evidence Act 1872 (India).
- Information Technology Act 2000 (India).
- Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 (India).
- Jacob Mathew v. State of Punjab (2005) 6 SCC 1 (Supreme Court of India).
- M.C. Mehta v. Union of India (1987) 1 SCC 395 (Supreme Court of India).
- Maneka Gandhi v. Union of India (1978) 1 SCC 248 (Supreme Court of India).
- Navtej Singh Johar v. Union of India (2018) 10 SCC 1 (Supreme Court of India).
- Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) COM(2022) 496 final (European Commission).
- R (Bridges) v Chief Constable of South Wales Police [2020] EWCA Civ 1058 (Court of Appeal).
- Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- Rylands v. Fletcher (1868) LR 3 HL 330 (House of Lords).
- State v. Loomis, 881 N.W.2d 749 (Wis. 2016) (Supreme Court of Wisconsin).
- Syad Akbar v. State of Karnataka (1979) 4 SCC 175 (Supreme Court of India).

B. Institutional Reports, Market Data, and Statistical Sources

- AI Incident Database (AIID) (2025). Incident Reports 2022–2025. Available at: https://incidentdatabase.ai [Accessed April 2026].
- Autonomous Vehicle Industry Association (AVIA) (2025). AV Miles Driven Report 2024–2025. Washington, DC: AVIA.
- Fortune Business Insights (2024). India Artificial Intelligence Market Size, Share & Industry Analysis, 2025–2032. Available at: https://www.fortunebusinessinsights.com/india-artificial-intelligence-market-113969 [Accessed April 2026].
- Foundation Model Transparency Index (2024). Transparency Scores: October 2023 and May 2024 Updates. Stanford University. Cited in: Stanford HAI, AI Index Report 2025.
- Government of India, Press Information Bureau (PIB) (2024). India's AI Talent and Developer Ecosystem. New Delhi: Government of India.
- IMARC Group (2024). India Artificial Intelligence in BFSI Market: Size, Trends and Forecast 2025–2033. Available at: https://www.imarcgroup.com/india-artificial-intelligence-in-bfsi-market [Accessed April 2026].
- McKinsey & Company (2024). The State of AI in 2024: Generative AI's Breakout Year. Cited in: Stanford HAI, AI Index Report 2025.
- NASSCOM (2024). AI Adoption Index 2.0: Tracking India's Sectoral Progress in AI Adoption. New Delhi: NASSCOM. Available at: https://nasscom.in [Accessed April 2026].
- National Highway Traffic Safety Administration (NHTSA) (2025). Automated Driving System Incident Reports: Standing General Order Data, 2021–2025. Washington, DC: U.S. Department of Transportation.
- OECD (2024). OECD AI Incidents Monitor (AIM). Paris: OECD. Available at: https://oecd.ai/en/incidents [Accessed April 2026].
- Robert F. Kennedy Human Rights Center (2025). Bias in Code: Algorithm Discrimination in Financial Systems. Available at: https://rfkhumanrights.org [Accessed April 2026].
- Sajadieh, S., et al. (2026). The AI Index 2026 Annual Report. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

- Sajadieh, S., et al. (2025). The AI Index 2025 Annual Report. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. Available at: https://hai.stanford.edu/ai-index/2025-ai-index-report [Accessed April 2026].
- Urban Institute (2024). Analysis of Home Mortgage Disclosure Act Data: Racial Disparities in Algorithmic Lending. Washington, DC: Urban Institute.
- U.S. Department of Justice (2022). Settlement with Meta Platforms. Washington, DC: Department of Justice.

C. Legal Scholarship and Policy Studies

- AP News (2025). Tesla ordered to pay $240M in Autopilot crash verdict. Available at: AP News [Accessed April 2026].
- European Parliament (2025). Artificial Intelligence and Civil Liability. Brussels: European Parliament.
- Norton Rose Fulbright (2024). Artificial intelligence and liability: Key takeaways from the AI Act and revised Product Liability Directive. London: Norton Rose Fulbright.
- Responsible AI Labs (2025). AI Safety Incidents of 2024: Lessons from Real-World Failures. Available at: https://responsibleailabs.ai [Accessed April 2026].
- Surfshark Research (2024). 2023 Was a Record Year for AI Incidents. Available at: https://surfshark.com/research/chart/ai-incidents-2023 [Accessed April 2026].
- Time Magazine (2026). What the Numbers Show About AI's Harms. Available at: https://time.com/7346091/ai-harm-risk [Accessed April 2026].