AI HALLUCINATIONS AND SAFETY FRAMEWORK IN INDIA: A COMPARATIVE ANALYSIS OF THE EU AI ACT, NIST AI RMF, AND THE AI (ETHICS AND ACCOUNTABILITY) BILL, 2025

Debarshi Roy Choudhury, B.A. LL.B., Jalpaiguri Law College

ABSTRACT

India is developing one of the world's largest AI ecosystems, yet it possesses one of the least developed AI-specific legal regimes and this paradox is not merely ironic but constitutionally perilous. The core issue addressed in this article is the absence of binding legal obligations in Indian law concerning algorithmic bias, AI hallucinations, and output reliability harms that are already being experienced on a large scale in critical domains such as credit assessment, judicial risk evaluation, and healthcare diagnostics. Three comparative frameworks are analyzed: the EU AI Act 2024, selected for its risk-based regulatory structure and its implementation of the proportionality principle in AI governance; the NIST AI Risk Management Framework for its technical operationalizability and detailed methodologies for measuring bias and hallucination; and the proposed AI (Ethics and Accountability) Bill, 2025, selected as the most pertinent domestic legislative pathway for examination. The primary argument posited is that India's current voluntary, principle-based approach to AI governance, as outlined in the MeitY AI Governance Guidelines, 2025, is both constitutionally and practically insufficient when compared to the significant risks posed by AI bias and hallucinations to fundamental rights enshrined in Articles 14 and 21 of the Constitution. This article contributes uniquely by proposing a four-pillar legislative framework: a statutory risk-classification system, mandatory pre-deployment bias audits and hallucination benchmarking, a right to explanation and algorithmic redress, and a technically mandated AI Safety Institute which are all tailored to India's distinctive constitutional, social, and technological context.

Keywords: Algorithmic Bias, AI Hallucinations, EU AI Act, NIST AI RMF, AI Ethics and Governance, AI Safety, Fundamental Rights

I. INTRODUCTION

In June 2023, a federal Judge in the Southern District of New York found two practising attorneys had submitted a legal brief riddled with citations to non-existent cases. These citations had been generated by an AI tool produced with total confidence, formatted with genuine citations, and indistinguishable in appearance from real ones. The attorney had trusted this tool, even asking it whether the citations were real and the tool had confirmed their authenticity. Judge P. Kevin Castel imposed a $5,000 fine and ordered the attorneys to notify the cited judges.

The story of *Mata v. Avianca, Inc,* is typically told as a cautionary tale about careless lawyers. However, reading this case misses its deeper structural problem. The attorneys were not negligent in the ordinary sense as they were only using a tool designed to produce authoritative outputs, which provided no visible signal of epistemic uncertainty and failed in a way that was, by definition, undetectable without independent verification. The consequences that followed to the lawyers' professional integrity and the dignity of the court were not caused by bad faith but by a fundamental mismatch between what the AI systems appeared to be doing and what it was actually doing.[1]

That mismatch is the subject of this article. And while *Mata v. Avianca* arose in a Manhattan federal courtroom, the legal problem it dramatises is universal and it lands with particular urgency in India, where AI is not a future hypothetical but a present operational reality. Indian courts are experimenting with AI-powered translation and transcription tools. Fintech platforms are making credit decisions about millions of Indians using algorithmic scoring systems trained on data that may not represent the diversity of the country's population. Large employers in the formal sector are using AI-assisted recruitment tools. The government has deployed AI in welfare eligibility determinations and in law enforcement analytics.[2]

Ask, now, the honest question: when these systems are wrong, biased, hallucinated and when they produce confident outputs that are quietly, invisibly, and consequentially incorrect then who is entitled for accountability under Indian law? At present, nobody is. There is no

[1] *Mata v. Avianca, Inc*., No. 22-cv- 1461, 2023 WL 4114965 (S.D.N.Y. June 22, 2023)
[2] Association of Corporate Counsel, *Practical Lessons from the Attorney AI Missteps in Mata v. Avianca* (Aug. 7, 2023), https://www.acc.com/resource-library/practical-lessons-attorney-ai-missteps-mata-v-avianca (last visited 16 June, 2026)

meaningful, enforceable sense that a harmed individual could invoke to obtain a remedy.

The Information Technology Act, 2000 was built for a world of human intermediaries and user-generated content. In 2023, the most sophisticated privacy statute India produced, but it addresses algorithmic inferences derived from data.[3] The India AI Governance Guidelines, 2025 articulate seven admirable principles but create no binding obligations, no enforcement mechanisms, and no remedies.[4] The AI market in India is estimated at USD 12.7 billion in 2025 and projected to reach USD 55.3 billion by 2033 and it is growing at a rate that makes this regulatory vacuum increasingly untenable.[5]

This article attempts to answer a specific question: Does India's architecture comprising the Information Technology Act, 2000, the Digital Personal Data Protection Act, 2023, and the MeitY AI Governance Guidelines, 2025 provide adequate legal remedies and preventive obligations for harms arising from algorithmic bias, hallucinations, and unsafe outputs? If not, what form should a risk-based legislative framework take?

The methodology is doctrinal and comparative. Primary sources include legislative texts of the statutes and frameworks examined judicial decisions, regulatory guidelines, and official government reports. Secondary sources include peer-reviewed scholarship and parliamentary committee reports. This comparative method is used to transplant foreign law or to identify structural principles that can be adapted to India's constitutional framework. Section II examines the nature of algorithmic bias and AI hallucinations as legally cognisable harms. Section III undertakes this comparative analysis. Section IV proposes four recommendations for a robust framework. Section V draws conclusions and proposes reformative structure.

II.UNDERSTANDING THE BIAS AND HALLUCINATIONS

It is necessary to define exactly what kind of harm algorithmic systems produce, and why existing legal precedents fail to capture it. The difficulty with algorithmic bias and AI hallucinations is that both are technical issues embedded inside the system and does not fit well onto the legal categories such as negligence, fraud, and discrimination, through which Indian

[3] Information Technology Act, No. 21 of 2000, Section 79 (India)
[4] India AI Governance Guidelines, *"Enabling Safe and Trusted Innovation"* (Nov. 2025) https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf (last visited June 17, 2026)
[5] Spherical Insights & Consulting, India Artificial Intelligence Market Size, Share & Forecast (2026) , https://www.sphericalinsights.com/reports/india-artificial-intelligence-market (last visited June 17, 2026)

law currently recognises harm. In this section, we will broadly discuss about what counts as "bias" in a machine learning system and that constitute legally applicable harm, as these are questions that engineers and lawyers answer differently and these differences have significant consequences.

A. Algorithmic Bias

Algorithmic bias is not a software bug in the ordinary sense but rather a systematic reproduction of human prejudices through training data and model architecture. These biases are not introduced due to malice or intentional discrimination by the developers; instead, it reflects the societal inequalities present in the data used for training. An AI system learns to replicate these inequalities faithfully and efficiently, mirroring the historical biases inherent in the data.

Consider a hypothetical scenario that is not purely hypothetical, like a hiring algorithm trained on a decade of decisions from a major company known for consistently promoting upper-caste, metropolitan candidates to senior roles. The algorithm learns that these characteristics correlate with success as defined by the company's historical standards. When applied systematically, the algorithm discriminates against candidates from tier-three engineering colleges with regional-language educations, despite their qualifications being identical to those of the favoured candidates. The algorithm has discriminated, even though no individual within the company intended to do so.

This scenario illustrates a significant legal challenge: Article 14 of the Indian Constitution defines certain forms of arbitrary and discriminatory state action, including discrimination based on religion, race, caste, sex, and place of birth. However, these provisions were designed for cases where discrimination can be clearly attributed to a human decision-maker, who's reasoning can be interrogated and whose intent can be proven. In contrast, when a decision is made by a mathematical function trained on biased historical data, the causal chain is often too complex for traditional legal and tort doctrines to reconstruct effectively. The person harmed may not be able to identify a specific discriminatory decision-maker, as there is none in the conventional sense; instead, they face a decision made by an optimization function with an incorrect objective.

This issue extends beyond private employment. Government agencies also use such systems in various contexts, including judicial risk assessments, welfare eligibility determinations, and

police profiling. In these scenarios, the constitutional implications become particularly acute. A citizen affected by an algorithmic state decision has access to judicial review under Article 226, but the question remains: what is being reviewed when the decision-making process is opaque. The lack of transparency means that one cannot meaningfully ask the model why it made a particular decision and receive a causal explanation. Instead, one can only observe the inputs and outputs of the system, making it akin to a mathematical black box.

B. AI Hallucinations

The term "hallucination," when applied to AI systems, refers to a phenomenon where output generated is factually incorrect yet indistinguishable from correct output. A large language model, relying on the technology, does not know facts but predicts the most statistically probable next word or token based on patterns in its training data. Given that the training data may contain gaps or that the model is queried at the edges of its training distribution, it fills these gaps with the most plausible-seeming continuation, which can appear knowledgeable but is fundamentally wrong.

This technical specificity highlights a legal dimension. For instance, in a Scenario where a junior advocate uses an AI legal tool to identify precedents for a bail application in a sessions court, The tool might confidently cite a case such as Ramesh Kumar v. State of Haryana, (2019) 14 SCC 282, that does not exist. The advocate, trusting the tool's output, includes the false citation in the pleading and presents it to the judge. Under Section 35 of the Advocates Act, 1961, professional misconduct attaches to the advocate for misleading the court. However, the legal responsibility lies with the advocate as the developer bears no identifiable liability under current Indian law.

In another context, a diagnostic AI deployed in an underrepresented district hospital in West Bengal might incorrectly diagnose a patient with viral fever when the condition is actually bacterial meningitis. If the patient receives the wrong treatment, a medical services provider may be liable under Section 2(11) of the Medical Council of India Act, 1956. However, the question arises whether the AI developer is a "service provider" within the meaning of Section 2(42). If the hospital deploys a third-party tool and lacks the technical capacity to evaluate its reliability, the liability may extend to the developer. This issue remains unresolved in Indian law.

In the realm of credit scoring, a first-generation entrepreneur from a marginalized community might be denied a business loan due to an inflated risk assessment generated by a credit scoring system. The system's training data does not match his financial profile, leading to an inaccurate inference.[6] Under Section 8(3) of the Digital Personal Data Protection Bill, 2023, the individual must ensure that personal data used in decisions is complete, accurate, and consistent. However, the denial is not caused by inaccurate data but by an inaccurate inference derived from accurate data. The bill does not provide a right to challenge or have corrected an automated inference as it addresses data inaccuracies. Since data and inferences are legally distinct, this distinction swallows any potential remedy.

The point here is clear, that existing legal categories such as professional misconduct and negligence were designed for a world where human actors make human decisions and where the causal chain between wrongful conduct and harm is, at least in principle, traceable. As AI systems become more prevalent, incremental judicial interpretation of existing statutes will not suffice to address the gaps.

III.CRITICAL ANALYSIS OF THREE REGULATORY FRAMEWORKS

The three frameworks discussed in this section do not operate in isolation, rather they engage in a meaningful dialogue with one another. More critically, each framework addresses the same underlying issues with fundamentally different assumptions regarding how law should regulate risk, the level of trust regulators should place in industry, and the relationship between technical standards and legal obligations. Analyzing the underlying assumptions is more insightful than merely cataloguing provisions.

A. EU AI Act, 2024

The EU AI Act is considered significant because it does not primarily prohibit certain AI systems but rather focuses on how these systems are managed. It does not ask whether a given AI system is harmful in a binary sense. Instead, it examines the context in which the system is deployed, the stakes involved, the margin for error, and the presence of human oversight to catch mistakes. This approach represents a sophisticated legal move. The proportionality principle which was long established in European constitutional law as a requirement that state

[6] Cyrus Farahani & Akbar Akhtar, *Credit Scoring in India,* Indian J. L. & Tech. Blog (April 30, 2025) , https://forum.nls.ac.in/ijlt-blog-post/credit-scoring-in-india/ (last visited June 16, 2026)

measures be appropriate, necessary, and not excessive relative to their legitimate aim is integrated into the governance of technology.[7]

The four-tier architecture is particularly noteworthy.[8] At the top tier, certain AI applications are prohibited due to their unacceptable risk profile such as government-operated social scoring systems, real-time biometric identification in public spaces by law enforcement, systems that exploit psychological vulnerabilities to manipulate behaviour, and systems that infer sensitive characteristics from biometric data. The second tier, high-risk AI systems, includes those used in critical infrastructure, hiring and employment decisions, credit scoring, educational and vocational training, law enforcement, migration, and administration of justice. High-Risk systems, prior to deployment, must establish risk management systems, demonstrate data adequacy, maintain technical documentation, implement human oversight mechanisms, and achieve appropriate levels of accuracy, robustness, and cyber security.[9] The third tier, limited risk, applies to chatbots and generative AI content, imposing disclosure obligations so users know that they are interacting with an AI system. The fourth tier, minimal risk, covers systems like spam filters and games, which face no regulatory obligations.[10]

The most important quality of The EU AI Act, for the purposes of this article, is its anticipatory logic. This ensures that high-risk systems cannot be deployed until they have undergone conformity assessment which is a structured pre-deployment evaluation against statutory requirements. This represents a fundamental shift from the ex-post liability regime that traditionally dominates Indian law, where the legal system responds only after harm has occurred. Some harm are severe enough, and the connection between system design and harm is predictable enough, and for that prevention is indeed better than cure.

However, The EU AI Act comes with notable resource requirements such as conformity assessments, and notified body infrastructure. While India has not yet fully developed, the Act addresses accuracy and robustness requirements for high-risk systems which do not prescribe a specific hallucination rate benchmark or output verification standard; instead, it says high-risk systems must be accurate and robust, but does not specify how accurate they need to be or

[7] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689) 1 (EU), arts. 6-17.
[8] EU AI Act arts. 5(1) (a)-(h)
[9] EU AI Act, arts. 10, 15
[10] European Commission, AI Act: *The European Approach to Artificial Intelligence*, (2024) https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (last visited June 17, 2026)

how robustness against hallucination should be tested. As a legal instrument that aims to be technically specific, this identifies gaps that require further development. Interestingly, The NIST framework is well-positioned to address these gaps.

B. NIST AI Risk Management Framework

The NIST AI RMF is not a legally binding document that should be stated plainly and immediately because the temptation in a comparative legal analysis, the requirement is to treat frameworks as functionally equivalent regardless of their legal status. The NIST AI RMF is a voluntary guidance document developed by the United States National Institute of Standards and Technology, intended for organizations developing and deploying AI systems. Its value for this article is not as a binding legal model but as something The EU AI Act precisely lacks, which is a technically operationalized standard that tells you what to measure and how to measure it.[11]

The framework organizes around four core functions such as Govern, Map, Measure, and Manage. Those together constitute a continuous risk management cycle rather than a one-time compliance checkpoint. The Govern function establishes organizational accountability structures such as documented roles and responsibilities, red-teaming and adversarial testing requirements, and incident response planning. The Map function establishes context and identifies what kinds of risk a given AI system poses in its specific deployment environments. The Measure function prescribes testing for fairness, reliability, and accuracy at multiple stages of the model lifecycle. The Manage function prioritizes identified risks and defines treatment strategies.[12]

For AI hallucinations and bias specifically, the framework's 2025 updates, which introduced a Generative AI Profile extending the core framework are particularly significant. The Profile identifies twelve risk categories specific to generative AI systems, including hallucination, data poisoning, prompt injection, intellectual property concerns, and over-reliance. It provides guidance on measuring hallucination using benchmark datasets and adversarial testing. It also introduced broader threat taxonomies, stronger supply chain and third-party risk management requirements, and maturity model guidance encouraging organizations to measure their AI risk

[11] National Institute of Standards & Tech., Artificial Intelligence Risk Management Framework (AI RMF 1.0) (Jan. 2023) https://doi.org/10.6028/NIST.AI.100-1 (last visited 16 June, 2026)
[12] NIST AI RMF, supra note 11, at 8-9 (Govern, Map, Measure, Manage functions)

management progress rather than treating compliance as binary.

The critical limitation is inherent in the framework's voluntary character. Because compliance is optional, The NIST AI RMF has not produced uniform industry practice. Large technology companies with substantial legal and compliance teams engage seriously with it. Smaller developers and deployers who constitute the overwhelming majority of India's AI ecosystem, which spans everything from fintech startups to agritech platforms to judicial software vendors often do not. The framework's value to India lies not in its voluntary character but in its technical substance such as the metrics, methodologies, and measurement approaches it has developed can be extracted and imported into binding Indian regulation without adopting the American approach's non-binding nature.

C. AI (Ethics and Accountability) Bill, 2025

India's proposed AI (Ethics and Accountability) Bill, 2025 introduced on December 17, 2025, as a Private Member's Bill represents a significant step toward dedicated AI governance and deserves credit for that alone. [13]Moving from entirely voluntary MeitY Guidelines to proposed binding legislation is a notable advancement in the right direction.

Several aspects of the Bill deserve recognition for their substantive contributions. Notably, it establishes a statutory Ethics Committee for Artificial Intelligence, endowed with the power to develop ethical guidelines, monitor compliance, review cases of bias and misuse, and handle complaints from affected individuals. It also imposes transparency obligations on AI developers, including the disclosure of training data sources, methodologies, intended purposes, and limitations. Furthermore, it mandates heightened scrutiny for AI systems deployed in critical decision-making contexts such as law enforcement, financial credit, and employment. These systems must not discriminate based on race, religion, or gender and must undergo ethical review before deployment. The definition of "AI system" covering computer systems capable of performing tasks requiring human-level cognition, including decision-making, language processing, and visual perception is broad enough to encompass most modern machine learning architectures, which is more comprehensive than previous

[13] The Artificial Intelligence (Ethics and Accountability) Bill, 2025 (Private Member's Bill introduced in Lok Sabha, Dec. 17, 2025)

legislation.[14]

However, when examined against the specific harms identified in Section II, the Bill falls short. The structural nature of these harms is evident, and lacks specific provisions addressing algorithmic bias.

While it recognizes the need for transparency and accuracy, it does not adequately distinguish between confidently wrong AI outputs and more ordinary errors and inaccuracies that current laws already address imperfectly. These errors that present as certainty and provide no visible signal of their own incorrectness pose a unique danger that requires a more nuanced legal framework.

The Bill lacks a mandatory pre-deployment bias audit requirement comparable to The EU AI Act's conformity assessment regime. Although The Ethics Committee can review AI systems and investigate complaints, it does not appear to mandate that high-risk AI systems be audited and certified before deployment. This represents a fundamental architectural gap, as it means the legal system responds to bias after it has affected real people in real decisions, rather than catching it before it does.

Moreover, The Bill lacks an institutional framework with a technical standards infrastructure. While The Ethics Committee is equipped with regulatory authority, there is no statutory body mandated to develop technical benchmarks against which AI systems will be assessed. This is the reverse of The NIST approach, which provides technical substance without regulatory authority.

While The AI (Ethics and Accountability) Bill, 2025, is better than nothing and significantly better than purely voluntary MeitY Guidelines that that preceded it, it is not yet sufficient. It is a legal solution designed at a level of abstraction that does not engage with the specific, concrete, technically identifiable ways in which AI systems fail.

IV.RECOMMENDATIONS FOR A ROBUST FRAMEWORK

This framework is rooted in established constitutional values, bolstered by the comparative analysis presented in Section IV, and created to address the gaps highlighted in Section II. The

[14] SCC Online, Artificial Intelligence (Ethics and Accountability) Bill, 2025- Legal Update (Dec. 18, 2025), https://www.scconline.com/blog/post/2025/12/19 (last visited 16 June, 2026)

objective is not to replicate the EU AI Act verbatim into Indian legislation, but to distil the underlying structural principles of each framework component and adapt them to India's unique context which is a federal system with constitutional guarantees of substantive equality, a diverse and underserved population facing significant harm from biased AI, and an AI landscape ranging from multinational tech giants to local fintech start-ups in tier-three cities.

A. Risk-Classification System

India should enact, via legislation, a risk-classification system inspired by but distinct from the EU AI Act. The EU model employs four categories, whereas India's system will consist of three, adjusted to reflect the practicalities of AI deployment in India.

First tier defines prohibited applications which include government-operated social scoring systems, mass biometric surveillance in public spaces without individualized judicial authorization, and AI systems exploiting psychologically identified vulnerabilities to manipulate individuals' decisions without their knowledge. These applications pose real risks and would be among the earliest governmental uses of advanced AI in a country with a history of surveillance and exclusion. Their prohibition must be absolute and statutory, not merely advisory.

Second tier defines high-risk applications that encompass AI systems used in judicial decision support, financial credit scoring, healthcare diagnostics in public facilities, welfare eligibility determination, and employment screening. These contexts carry the most severe and irreversible harms for individuals. Providers of high-risk AI systems must comply with mandatory pre-deployment obligations, require human oversight, and undertake post-deployment monitoring.

Third tier defines general purpose AI applications which include AI systems deployed in consumer-facing contexts where significant individual decisions are not being made, subject to transparency and incident-reporting obligations.

Crucially, India's framework must diverge from the EU model by giving specific and explicit consideration to applications deployed in contexts of existing social vulnerability such as caste, gender, class, disability, and linguistic minority status. A credit scoring model that disproportionately fails historically marginalized communities is not just a technical issue

rather it is a constitutional concern. India's framework must recognize that the distribution of AI harm follows existing fault lines of inequality and the legal response must be calibrated accordingly. This is a constitutional duty under Articles 15 and 16, not merely a policy preference.

B. Bias Audit and Hallucination Benchmarking

High-risk and AI systems must undergo independent third-party audits before deployment in India, evaluating the demographic representativeness of training data, with particular focus on caste, gender, region, and linguistic background; differential outcome rates across protected categories at statistically significant confidence levels; and hallucination rates on domain-relevant benchmark tasks tailored to the specific application context.

The audit methodology should be established by a statutory technical standards body, as outlined in section 4 by leveraging NIST AI RMF measurement methodologies and adapting them for Indian languages, social categories, and specific use cases. The NIST framework's Measure function offers precisely the technical infrastructure needed for benchmark datasets, fairness metrics, adversarial testing protocols, and continuous monitoring frameworks. The challenge is not merely methodological rather it is institutional. A body endowed with the technical competence to develop and maintain these benchmarks must be created by statute, not through ministerial discretion.

Continuous monitoring obligations should mandate that deployers of high-risk AI systems must strictly maintain audit logs of automated decisions, conduct monitoring of output distributions across protected categories, and report any anomalies to the regulatory authority within a specified timeframe. The EU AI Act already imposes logging requirements as a condition for high-risk AI deployment. India's framework should incorporate this principle and include a specific hallucination monitoring requirement that the EU Act omits.

C. Accountability and Algorithmic Redress

India should enact legislation whether through amending the DPDP Act, passing the AI (Ethics and Accountability) Bill, or creating a standalone statute to confer a three-part right on individuals impacted by high-risk AI decisions.

First, the right to notification which entitles any individual whose decision is made by, or

significantly influenced by, a high-risk AI system must be informed that AI played a role in that decision. This is more than just a transparency requirement as it is essential for exercising other rights meaningfully. Without knowing that an algorithm was involved, one cannot challenge the decision effectively.

Second, the accountability rights which entitle the individual should receive a clear, understandable explanation of the factors and logic behind the AI system's output, detailed enough to allow them to challenge the basis of the decision. This explanation should focus on the specific factors that the system considered, rather than the model architecture or training data composition, which are proprietary and serve legitimate commercial interests. India should codify this right, requiring that explanations be provided in the individual's preferred language, given the country's linguistic diversity, which necessitates a statutory obligation to offer explanations in regional languages where necessary.

Third, the right to human review which states that the individual must have the right to have the AI-influenced decision reviewed by a human reviewer who has the authority to override the AI output, the competence to assess the relevant factors, and no institutional bias towards rubber-stamping the algorithmic result. This right is established by GDPR Article 22 and is notably absent from the DPDP Act.[15] In the Indian constitutional context, this aligns with the right against arbitrary decision-making under Article 14 and the informational autonomy dimension of Puttaswamy under Article 21.[16] A decision made solely by an algorithm, without any meaningful human oversight, and without a mechanism for the affected person to seek human reconsideration, would be constitutionally questionable under both provisions.

This right should extend beyond state actors. The most significant AI-driven decisions affecting individual Indians are often made by private financial institutions, employers, and healthcare providers. A rights-protective framework that only covers public-sector AI deployment would fail to address the majority of the harm that occurs.[17]

D. AI Safety Institute with Technical Approach

The MeitY AI Governance Guidelines, 2025 have already proposed the creation of an AI Safety

[15] GDPR Article 22(1)-(3)
[16] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1
[17] AI Bill 2025, cl. 8, 10, 11

Institute. This article contends that the AI Safety Institute should be established by statute rather than through executive order or ministerial discretion, and should possess three distinct technical functions that differentiate it from the AI (Ethics and Accountability) Bill.[18]

One of these functions involves operating a national AI incident registry, which is a publicly accessible database where harms caused by AI systems are reported, classified, investigated, and published. This registry serves two primary purposes, first, it establishes an evidentiary foundation for understanding the scale and nature of AI-caused harm in India, which is currently lacking; second, it exerts reputational pressure on organizations whose AI systems are the subject of multiple incident reports, thereby reinforcing their legal obligations. Another function of the AI Safety Institute is developing and publishing domain-specific hallucination and bias benchmarks in Indian languages. The AI Safety Institute should utilize the benchmark development methodology provided by the NIST AI RMF to create benchmarks for judicial risk assessment AI in Hindi, Bengali, Tamil, Telugu, and other major languages; for credit scoring AI tested against the demographic diversity of India's borrowing population; and for healthcare diagnostic AI benchmarked against disease profiles relevant to India's epidemiological landscape. This is substantial work that necessitates sustained public investment and cannot be achieved through voluntary industry initiatives.

The third function which is crucial for providing the AI Safety Institute with its legal authority is issuing binding technical standards that serve as safe harbour criteria. Organizations that comply with the AI Safety Institute's published technical standards for a given application category are presumed to meet the statutory risk management obligations for high-risk AI. Conversely, those who do not comply with the standards must demonstrate, through their own conformity evidence, that their system provides equivalent levels of safety, fairness, and accuracy. This design borrows the safe harbour principle from Indian data protection law and applies it to AI safety, creating positive incentives for compliance while maintaining regulatory flexibility for technical innovation.[19]

This model combines the technical infrastructure logic of NIST with the binding legal character of the EU approach, allowing the AI Safety Institute to begin as a standard approach and

[18] MeitY AI Guidelines, principle VII ("Safety, Security and Sustainability")
[19] Office of the Principal Scientific Adviser to the Government of India, *Techno-Legal Framework for AI Governance in India* (Jan. 2026)
https://www.pib.gov.in/PressReleasePage.aspx?PRID=2217839&reg=3&lang=1 (last visited June 17, 2026)

registry operator and potentially evolve into full conformity assessment authority without immediately replicating India's notified-body conformity assessment infrastructure.

V.CONCLUSION

The question this article grapples with pertains fundamentally to the kind of society India wants aspires to become, one that is increasingly a significant global AI power. The issue of who is obligated, who is protected, and who bears the cost of failure, whether through voluntary governance or binding regulation is inherently a choice about societal values. By opting to govern through voluntary measures rather than binding regulations is not a neutral technical preference, it signifies a decision to retain the burden of failure where it currently resides, on those least equipped to bear it.

India's credit markets are employing algorithms to make decisions about tens of millions of individuals who have never interacted with a bank. Similarly, India's court system is encouraged to use these algorithms for case management and translation in a system already overwhelmed by millions of pending cases. In rural and semi-urban areas, where the doctor-patient ratio often makes diagnostic services commercially attractive, India's hospitals are deploying tools whose accuracy and reliability have not been independently verified. In these contexts, when the algorithmic system fails to make correct predictions, discriminates, or produces confident but erroneous outputs, and then the individual who bears the cost is not the developer who created the system or the deployer who profited from it. Instead, it falls on the undertrial in a district jail whose bail was assessed by a biased tool, the first-generation entrepreneur whose creditworthy business was invisible to a model trained on different data, or the patient in an underrepresented government hospital who trusted a diagnosis that was statistically probable but medically wrong.

A tiered legislative framework does not eliminate the risks associated with algorithmic decision-making. While it may not be a realistic legislative ambition to completely eliminate these risks, a tiered framework does redistribute the burden of these risks towards those best positioned to manage them such as the developers who design the systems, the deployers who profit from them and the regulators who set the standards. This shift away from individual users, who lack the power to inspect, audit, or correct the systems that impact their lives, is not merely pragmatic rather it is constitutionally required. A constitutional order committed to substantive equality under Articles 14, 15, and 16, and to individual dignity and informational

autonomy under Article 21, cannot be coherently interpreted to permit a framework that allows the most powerful actors in the AI economy to externalize the costs of their systems' failures onto the most vulnerable people in the country.

Whether India's legislature will act with the urgency demanded by the current moment remains uncertain. However, what can be said is that the cost of waiting will be borne by those who can least afford it.