Jurnal Hukum Siber
dan Regulasi Teknologi

SAH PUBLISHER

e-ISSN 3123-8114
Vol. 1, No. 2, May 2026

Artificial Intelligence and Criminal Liability: The Incompatibility of Autonomous Decision-Making Systems with the Mens Rea Requirement

Sonali Kumari[1]*, Usha Shree Parija[2]
[1,2]KIIT Law School, Bhubaneswar, India
Email: [1]Sk1756016@gmail.com,[2] 2282138@kls.ac.in

Abstract. The increasing deployment of Artificial Intelligence (AI) in autonomous decision-making systems has generated complex challenges for criminal law. Traditional criminal liability is grounded in human agency, requiring the coexistence of actus reus and mens rea. However, AI systems now operate with varying degrees of autonomy, unpredictability, and opacity, raising the fundamental question: who should bear criminal responsibility when AI causes harm? This paper examines the doctrinal compatibility of AI-related harm with existing criminal law principles, particularly within the framework of the Bharatiya Nyaya Sanhita, 2023 (BNS). Using a qualitative doctrinal and analytical legal research methodology based exclusively on statutory interpretation, judicial precedent, and contemporary scholarly literature - without reliance on empirical or experimental data - the study evaluates whether AI can be treated as a subject of criminal liability or whether responsibility must remain exclusively human-centered. The findings demonstrate that AI lacks moral agency and legal personhood, thereby preventing direct criminal attribution. Instead, liability must be distributed among developers, deployers, corporations, and regulatory authorities through a layered accountability model structured around developer responsibility, corporate governance liability, operator oversight obligations, and negligence-based attribution under existing statutory provisions. The novelty of this paper lies in integrating responsibility-gap theory with a concrete doctrinal analysis of the Bharatiya Nyaya Sanhita, 2023. Unlike prior scholarship that primarily emphasizes ethical governance or regulatory policy frameworks, this study situates AI accountability within the punitive and blame-based structure of substantive criminal law, thereby offering a legally operational model for addressing AI-related harms.

Keywords: Artificial Intelligence, Criminal Liability, Responsibility Gap, Explainability, Bharatiya Nyaya Sanhita

INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) technologies has significantly transformed contemporary society, particularly in domains such as healthcare, finance, transportation, surveillance, and automated decision-making. AI systems increasingly operate with substantial autonomy, producing outcomes without continuous human supervision. While such developments enhance efficiency and innovation, they also generate profound challenges for criminal law. The central question is: who should bear criminal responsibility when an autonomous AI system causes harm?

Traditional criminal liability is grounded in human agency, requiring the concurrence of actus reus (guilty act) and mens rea (guilty mind). Criminal jurisprudence presupposes moral

Received Feb 2026 / Revised Month 2026 / Accepted May 2026
*Corresponding author.
Email addresses: Sk1756016@gmail.coml (Kumari)

blameworthiness, intention, knowledge, or negligence. AI systems, however, operate without consciousness, volition, or moral awareness, thereby unsettling classical frameworks of culpability.

The issue of responsibility attribution in AI has been widely debated in contemporary scholarship. Coeckelbergh emphasizes relational responsibility and human-centered attribution [1], while Matthias introduces the concept of the "responsibility gap," highlighting the difficulty of assigning blame where autonomous systems act unpredictably [2]. Floridi and Taddeo situate AI harms within broader data ethics frameworks [3], and Mittelstadt et al. map structural accountability concerns such as opacity and bias [4]. Although this scholarship contributes significantly to ethical and regulatory discourse, limited attention has been devoted to doctrinal criminal law analysis within specific statutory frameworks.

In the Indian context, the Bharatiya Nyaya Sanhita, 2023 (BNS) retains the foundational structure of criminal culpability grounded in intention, knowledge, and negligence. Indian jurisprudence has long recognized corporate criminal liability through attribution doctrines. In Standard Chartered Bank v Directorate of Enforcement (2005) and Iridium India Telecom Ltd v Motorola Inc (2011), the Supreme Court affirmed that corporations - though artificial entities - may possess mens rea through attribution to controlling human agents. These decisions demonstrate that Indian criminal law already accommodates non-natural actors through derivative liability. However, AI systems differ from corporations in that they lack legal personality and identifiable directing minds, raising a distinct doctrinal challenge.

Existing literature often frames AI either as a potential legal person or merely as an instrument. Both positions are conceptually incomplete. Treating AI as a legal person conflicts with criminal law's emphasis on moral agency, while treating AI purely as a passive tool ignores its capacity for autonomous and self-learning behavior. Consequently, a doctrinal gap persists: although AI-related harms are increasing, criminal law remains structured around human cognition and intention.

The current state of the art in AI regulation further confirms this gap. Contemporary instruments such as the EU AI Act regulate AI through risk management, documentation, transparency, and human oversight duties, but they do not resolve whether an autonomous AI system can itself possess criminal culpability. This shows that AI governance has advanced significantly at the administrative and regulatory levels, while substantive criminal law continues to depend on human-centered concepts of intention, knowledge, negligence, and blameworthiness.

This research seeks to bridge the gap between AI responsibility scholarship and substantive criminal law doctrine under the Bharatiya Nyaya Sanhita, 2023. The objectives of this study are:

(1) To analyze whether AI systems can satisfy the traditional elements of criminal liability; (2) To examine the applicability of the Bharatiya Nyaya Sanhita, 2023 to AI-related harms; (3) To evaluate the responsibility gap in the context of criminal culpability; and (4) To propose a structured layered accountability framework that ensures criminal responsibility remains enforceable despite technological autonomy.

The novelty of this paper lies in its integration of responsibility-gap theory with a doctrinal analysis grounded specifically in the Bharatiya Nyaya Sanhita, 2023. Unlike prior scholarship

that focuses primarily on ethical governance or regulatory reform, this study situates AI accountability within the punitive and blame-based architecture of criminal jurisprudence, offering a legally operational model consistent with established attribution principles in Indian law.

METHODS

This research adopts a doctrinal and analytical legal research methodology to examine the criminal liability of Artificial Intelligence (AI) systems and to determine who should be held responsible under contemporary criminal law frameworks. The study primarily relies on statutory interpretation, case-based doctrinal analysis, and comparative evaluation of existing scholarly literature. The objective of this methodological approach is to systematically evaluate whether traditional elements of criminal liability - actus reus and mens rea - can be satisfied in AI-related harms and to assess how responsibility may be attributed within the framework of the Bharatiya Nyaya Sanhita, 2023 (BNS).

The doctrinal method is appropriate because criminal liability is fundamentally grounded in statutory interpretation and judicial reasoning. As Turner [9] and Coeckelbergh [1] suggest, AI accountability questions require integration of technological realities with existing legal principles. Therefore, this study analyzes relevant provisions of the BNS 2023 [12], especially those relating to intention, knowledge, negligence, abetment, conspiracy, and corporate criminal liability. These statutory provisions are examined to determine whether AI systems themselves can satisfy legal requirements or whether liability must remain human-centered.

A. Research Design

The qualitative normative legal research technique used in this study is based on doctrinal analysis and statutory interpretation. Rather than relying on empirical data, the research systematically evaluates legal principles and theoretical frameworks to assess the criminal liability of AI systems. The methodology is structured into three analytical stages:

Conceptual Analysis of Criminal Liability

The foundational doctrines of criminal law are examined, including actus reus, mens rea, intention, recklessness, negligence, and vicarious liability. These principles serve as the normative benchmark to determine whether AI systems can qualify as legal subjects capable of criminal blameworthiness. The analysis evaluates whether AI satisfies the requirements of moral agency, voluntariness, and culpable mental state.

Responsibility Gap Assessment

The study applies Matthias's "responsibility gap" theory [2] and Coeckelbergh's responsibility attribution framework [1] to assess whether autonomous AI behavior creates a vacuum in accountability. This stage critically examines the implications of distributed agency and diminished human control in AI-mediated decision-making.

Statutory Application under BNS 2023

The Bharatiya Nyaya Sanhita, 2023 [12] is analyzed to determine whether AI-related harm can be addressed through existing provisions such as:

Criminal negligence

Abetment and conspiracy

Corporate criminal liability

Knowledge- and intention-based offenses

The focus remains on interpretative sufficiency rather than legislative reform, ensuring doctrinal consistency and analytical precision.

B. Data Sources

This research is grounded exclusively in authoritative legal and academic sources, ensuring doctrinal integrity and analytical rigor. The data sources are categorized into primary and secondary materials to maintain structural clarity and scholarly reliability.

Primary Legal Sources

Primary sources form the statutory and regulatory foundation of the analysis. These include:

Bharatiya Nyaya Sanhita, 2023 (BNS 2023) [12], which provides the core criminal law framework for evaluating liability in the Indian context.

European Commission's Ethics Guidelines for Trustworthy AI [11], which offer internationally recognized principles on accountability, transparency, and governance relevant to AI deployment.

Indian Supreme Court decisions on corporate criminal liability and negligence.

Foundational comparative precedents from the United States and the United Kingdom addressing corporate criminal attribution.

These sources are examined through interpretative legal analysis to assess their applicability to AI-related harm.

Comparative Case Selection Rationale

The selection of U.S. and U.K. precedents is grounded in their foundational role in shaping modern corporate criminal liability doctrine within common law systems. Indian criminal jurisprudence has historically evolved within the common law tradition, and Supreme Court decisions concerning corporate liability have drawn upon comparative principles developed in these jurisdictions. Accordingly, cases such as New York Central & Hudson River Railroad Co. v. United States [13] and Tesco Supermarkets Ltd v Nattrass [14] are examined because they establish influential attribution models relevant to analyzing AI-related criminal responsibility

Secondary Academic Sources (Peer-Reviewed Literature)

To support doctrinal evaluation and theoretical development, the study relies on contemporary peer-reviewed scholarship, including:

Responsibility gap theory [2];

AI ethics and accountability debates [3], [4]; responsible AI governance frameworks [5];

Explainability and transparency research [6]

Empirical bias studies in AI systems [7]; and regulatory and governance scholarship [8], [9], [10].

More than 80% of the cited materials are peer-reviewed journal articles published within the last decade, ensuring contemporary relevance and compliance with academic publication standards.

C. Analytical Method

The analytical framework of this study is structured around four doctrinal tests designed to systematically evaluate criminal liability in AI-related harm scenarios. These tests ensure that liability analysis remains grounded in established principles of criminal jurisprudence rather than speculative technological assumptions.

Actus Reus Test

The first test examines whether AI systems can legally satisfy the requirement of a "guilty act." Since AI lacks an independent legal personality, the analysis focuses on attribution. It evaluates whether AI-generated conduct can be traced to identifiable human or institutional actors, including developers, programmers, operators, or corporate entities. The central question is whether AI actions may be treated as derivative human conduct within established criminal law doctrine.

Mens Rea Test

The second test addresses the mental element of crime. Criminal liability traditionally requires intention, knowledge, recklessness, or negligence. This stage, therefore, assesses whether subjective culpability can be transferred from human agents to AI-related outcomes. Particular emphasis is placed on foreseeability of harm and whether negligence standards can apply where risks were identifiable yet disregarded.

Corporate Liability Model Application

The third test applies principles of corporate criminal liability. Since corporations operate through structured processes and human agents, the study evaluates whether similar attribution doctrines - such as identification theory and vicarious liability - can be extended to AI systems deployed within corporate environments.

Distributed Responsibility Mapping

Finally, responsibility is mapped across multiple stakeholders to prevent oversimplified attribution. This structured approach ensures a balanced and legally coherent assessment of accountability.

D. Reproducibility and Logical Structure

To ensure doctrinal clarity, analytical transparency, and methodological coherence, the research adopts a structured and transparent analytical approach. First, it clearly identifies the statutory provisions under examination, particularly relevant sections of the Bharatiya Nyaya Sanhita,

2023, relating to negligence, abetment, conspiracy, corporate liability, and intention-based offences. Each provision is examined within its textual and interpretative framework to avoid arbitrary or selective reasoning.

Second, the study applies uniform doctrinal tests across carefully constructed hypothetical AI harm scenarios. The same analytical standards - actus reus, mens rea, attribution principles, and corporate liability doctrines - are consistently employed in each scenario to maintain internal consistency and interpretative coherence.

Third, consistent interpretative standards are maintained throughout the analysis. The research adheres to established principles of statutory interpretation, including purposive construction and contextual reading, without extending beyond doctrinal boundaries recognized in criminal jurisprudence.

Rather than emphasizing scientific reproducibility in an experimental sense, this doctrinal study prioritizes logical consistency, interpretative transparency, and reasoned justification - hallmarks of normative legal research.

E. Limitations

This research does not include empirical experimentation, algorithmic simulations, or technical system testing. Its focus is normative and doctrinal. A limitation of this approach is that it assumes AI-generated harm can be assessed through legal reasoning rather than technical modeling. However, for purposes of criminal liability analysis, doctrinal precision and conceptual clarity are paramount.

The methodology thus establishes a structured legal framework capable of systematically addressing the research questions outlined in the Introduction while maintaining fidelity to established criminal law principles.

RESULTS AND DISCUSSION

The central research question of this study is: Can Artificial Intelligence systems be held criminally liable, and if not, who should bear responsibility under contemporary criminal law, particularly under the Bharatiya Nyaya Sanhita, 2023?

Applying the doctrinal framework developed in the Methods section, the results indicate that AI systems cannot independently satisfy the legal requirements of criminal liability. However, criminal accountability can be structured through derivative human liability, corporate liability models, and negligence-based attribution mechanisms.

The findings are organized into eight doctrinal findings, beginning with the incompatibility of AI systems with mens rea and ending with the implications of the Bharatiya Nyaya Sanhita, 2023 for future AI-related criminal harm.

1. AI Systems Cannot Satisfy the Mens Rea Requirement

The first major finding confirms that AI systems cannot fulfill the subjective element (mens rea) required under criminal law.

Under classical criminal jurisprudence, criminal liability requires:

Intention

Knowledge

Recklessness

Negligence

AI systems operate through algorithmic processing and statistical optimization. They lack consciousness, moral awareness, and intentionality in the legal sense. As Coeckelbergh argues, AI may simulate agency but does not possess moral agency [1]. Similarly, Matthias identifies a "responsibility gap" when autonomous systems act unpredictably without human control [2].

Under the Bharatiya Nyaya Sanhita, 2023 [12], intention and knowledge remain foundational to serious criminal offenses. These mental states require:

Cognitive awareness

Volitional decision-making

Capacity to understand wrongdoing

AI systems cannot demonstrate these mental capacities. Even machine learning systems that adapt autonomously do not "intend" outcomes - they compute outputs based on data patterns. Therefore, AI cannot be treated as a criminal subject under existing Indian criminal law. The Supreme Court of India in Iridium India Telecom Ltd v Motorola Inc (2011) [16] affirmed that corporations may possess mens rea through attribution to their controlling human agents. Similarly, in Standard Chartered Bank v Directorate of Enforcement (2005) [17], the Court clarified that corporations may be prosecuted for criminal offences even where punishment includes imprisonment, reinforcing that artificial entities are not immune from criminal sanction. However, in both cases, liability was derivative and dependent upon human agency. AI systems, lacking any human cognitive substrate of their own, cannot satisfy this requirement independently. This finding aligns with Turner's regulatory analysis [9], which emphasizes that criminal law remains anthropocentric.

2. The Responsibility Gap Exists but Is Not Absolute

The second result confirms the existence of what scholars describe as a "responsibility gap," but it does not support the claim that this gap produces a complete legal vacuum. Matthias [2] argues that highly autonomous AI systems create situations in which no single human actor fully controls the outcome, the harm may be partially unforeseeable, and developers cannot predict the long-term learning trajectory of adaptive systems. These characteristics appear to weaken traditional criminal law doctrines that rely on identifiable intention and direct causal control. However, this concern must be assessed in light of negligence-based liability principles under the Bharatiya Nyaya Sanhita, 2023 (BNS 2023) [12]. Criminal liability does not always require proof of mens rea in the form of direct intention. Many offenses are grounded in negligence, where liability depends upon foreseeability of harm, breach of duty of care, and

failure to take reasonable safeguards. The principles governing criminal negligence articulated in Jacob Mathew v State of Punjab (2005)[18] are particularly instructive in this regard. The Supreme Court of India held that criminal negligence requires a gross deviation from reasonable standards of care and a high degree of foreseeability of harm. This standard is directly applicable to AI deployment scenarios. Where developers or corporations recklessly ignore foreseeable technological risks, liability may arise not from AI autonomy itself but from human failure to exercise due care.Thus, even when an AI system operates autonomously, liability may arise if developers failed to conduct adequate risk testing, if deployers ignored documented warnings, or if corporations neglected proper supervision and compliance mechanisms. As Mittelstadt et al. emphasize, algorithmic accountability should be grounded in institutional governance rather than in narrow individual blame [4]. Therefore, the responsibility gap shifts the focus from intention-based crimes to systemic negligence and structural failure, rather than eliminating liability.

3. Corporate Criminal Liability as the Doctrinal Model for AI Harm

The third finding demonstrates that the corporate criminal liability model provides the most coherent doctrinal framework for addressing AI-related harm. Although corporations lack biological consciousness, they are nevertheless recognized as legal persons capable of bearing criminal responsibility. This recognition does not arise from independent moral agency but from attribution principles grounded in human conduct and institutional control.

The foundational recognition of corporate criminal liability in New York Central & Hudson River Railroad Co. v. United States [13] established that corporations may be held criminally liable for acts committed by their employees within the scope of employment and for corporate benefit. The Court did not attribute consciousness to the corporation; rather, it relied upon derivative attribution from human agents acting on its behalf. This doctrinal development marked a decisive shift in criminal jurisprudence by confirming that artificial legal entities may be subject to penal sanction through structured attribution.

The identification doctrine, further refined in Tesco Supermarkets Ltd v Nattrass [14], clarified that corporate liability depends upon locating the "directing mind and will" within the organizational hierarchy. Criminal responsibility attaches only when the wrongful act can be traced to senior managerial authority whose mental state is legally attributable to the corporation. This doctrine reinforces a central principle of criminal law: liability requires an identifiable human cognitive agent whose intention or knowledge can be recognized in law.

Indian jurisprudence similarly affirms this position. In Standard Chartered Bank v Directorate of Enforcement, the Supreme Court of India held that corporations may be prosecuted for criminal offences even where statutory punishment includes imprisonment. Subsequently, in Iridium India Telecom Ltd v Motorola Inc [16], the Court explicitly recognized that corporations can possess mens rea through attribution to their controlling human agents. These decisions confirm that corporate criminal liability operates through derivative human intent rather than independent institutional consciousness.

Applying this analogy to AI systems yields a clear doctrinal conclusion. AI systems, like corporations, lack biological consciousness. However, unlike corporations, they do not possess legal personality, nor can they serve as a "directing mind" within an organizational structure. Their outputs are generated through algorithmic processing shaped by human design, training

data, deployment decisions, and governance frameworks. Accordingly, AI systems cannot be treated as autonomous legal offenders.

Instead, liability must attach to the corporations that design, deploy, supervise, or profit from AI systems, particularly where governance deficiencies, risk management failures, or negligent oversight can be established. This approach aligns with Dignum's Responsible AI framework [5] and Taddeo and Floridi's distributed responsibility theory [8], both of which emphasize structural accountability rather than artificial personhood.

Under the Bharatiya Nyaya Sanhita, 2023 [12], corporations may be prosecuted for offences committed through their agents. AI systems may therefore be conceptualized as technological instruments operating within corporate decision-making structures. Derivative corporate liability - grounded in attribution and institutional fault - thus emerges as the most stable and legally defensible model for criminal accountability in AI-related harm.

4. Explainability and Transparency as Preconditions for Criminal Attribution

The fourth finding concerns explainability as a foundational precondition for criminal attribution. At its core lies the concept of epistemic responsibility - the obligation of human actors to understand, supervise, and justify decisions made with or through AI systems. As Miller [6] argues, explanation is not merely a technical feature but a socially embedded process essential to accountability. Legal legitimacy depends upon intelligible and reasoned justification. When AI systems operate as opaque "black boxes," human operators may rely on outputs without comprehending the underlying decision-making logic, thereby weakening responsible agency and complicating attribution of fault. From a criminal law perspective, opacity produces two interconnected implications.

(i) Supervisory Negligence and Duty of Care
First, the deployment of non-explainable AI systems may constitute supervisory negligence. Criminal liability based on negligence requires a breach of duty coupled with foreseeable harm. Where corporations deploy opaque AI systems in high-risk sectors - such as healthcare, autonomous transportation, finance, or predictive policing - without ensuring traceability, auditability, and effective human oversight, they may breach established standards of reasonable care. Failure to implement transparency safeguards may amount to reckless or negligent deployment when risks are foreseeable. In this context, opacity becomes a legally relevant governance deficiency rather than a neutral technological characteristic.

(ii) Evidentiary and Causation Challenges
Second, opacity generates serious evidentiary complications. Criminal prosecution requires proof beyond reasonable doubt, including clear establishment of causation and attribution. If algorithmic decision-making pathways cannot be reconstructed, courts may struggle to determine whether harm resulted from defective design, negligent supervision, biased training data, or independent misuse. The European Commission's Ethics Guidelines for Trustworthy AI [11] emphasize transparency, traceability, and auditability precisely because these elements facilitate legal accountability and procedural fairness.

Judicial engagement with these concerns is evident in State v Loomis (2016) [15], where the Wisconsin Supreme Court examined the use of a proprietary risk-assessment algorithm

(COMPAS) in criminal sentencing. Although the Court permitted its continued use, it acknowledged limitations arising from algorithmic opacity and cautioned against exclusive reliance on such systems. The decision underscored due process concerns when legally consequential determinations are influenced by systems whose internal logic cannot be fully scrutinized. While Loomis did not impose criminal liability, it illustrates judicial recognition that algorithmic opacity raises structural challenges for fairness, accountability, and evidentiary transparency.

(iii) Doctrinal Implication for Criminal Liability

Accordingly, non-explainable AI raises not only ethical concerns but also concrete criminal law risks. Where traceability is absent, supervisory actors may be unable to discharge their duty of care. In such circumstances, failure to ensure auditability may itself constitute negligent conduct. Transparency, therefore, operates not merely as a governance ideal but as a doctrinal safeguard necessary for maintaining coherent criminal attribution in technologically mediated environments. Explainability, in this sense, becomes a structural condition for preserving the integrity of criminal justice systems. Without it, both responsibility attribution and evidentiary proof are significantly destabilized.

5. Bias and Data-Based Harm Can Attract Criminal Liability

The fifth major finding concerns the criminal implications of discriminatory AI outcomes. Empirical research by Caliskan, Bryson, and Narayanan [7] demonstrates that AI systems trained on large language corpora can reproduce and even amplify societal biases embedded in data. Such biases are not random errors but structural reflections of historical patterns present in digital texts and datasets. When these systems are deployed in sensitive domains - such as credit scoring, hiring, policing, or facial recognition - the resulting discriminatory outcomes can cause tangible harm.

If AI systems deny loans disproportionately to certain communities, misidentify individuals from specific racial groups, or generate harmful profiling decisions, these outcomes may satisfy elements of criminal misconduct under certain circumstances. The key legal inquiry is not whether the AI "intended" discrimination, but whether human actors exercised due diligence. Liability becomes more plausible where bias was foreseeable, where testing and validation were inadequate, or where known risks were ignored prior to deployment.

Under the Bharatiya Nyaya Sanhita, 2023 [12], offenses involving wrongful harm, cheating, or discriminatory conduct may be implicated depending on factual context. Floridi and Taddeo [3] argue that algorithmic harms are ethically significant because they result from systemic design choices rather than neutral automation. This reinforces the principle that biased outcomes reflect governance failures. Accordingly, criminal liability may arise when developers knowingly deploy biased systems, corporations fail to conduct impact assessments, or harm is foreseeable and preventable. Algorithmic discrimination thus shifts the focus from machine behavior to institutional accountability and risk management.

6. Individual vs Institutional Liability: Who Should Be Responsible?

Applying established criminal law doctrines, responsibility for AI-related harm should be structured hierarchically, reflecting degrees of control, knowledge, and institutional authority.

First, developers may incur liability where they intentionally design harmful functionalities, recklessly ignore documented technical risks, or fail to comply with accepted professional standards of testing, validation, and risk mitigation. Where foreseeable harm arises from negligent coding, biased training data, or inadequate system safeguards, liability may attach on the basis of professional negligence or reckless disregard.

Second, deploying corporations bear the most significant burden of accountability. Corporate criminal liability becomes relevant when governance structures are defective, compliance systems are absent, risk assessments are ignored, or commercial interests are prioritized over safety and public welfare. This framework is particularly applicable in high-impact sectors such as autonomous vehicle deployment, AI-driven medical diagnostics, and financial trading algorithms, where organizational decisions determine operational risk exposure. Third, operators may incur liability where they knowingly misuse AI systems or override established safety protocols without justification. Regulatory authorities generally face limited exposure; however, in exceptional cases involving gross systemic negligence, questions of institutional accountability may arise.

Structured Layered Responsibility Model (Schematic Summary)
To clarify the hierarchical allocation of accountability, the proposed layered responsibility model may be summarized as follows:

Table 1. Structured layered responsibility model for AI-related criminal harm

Actor	Basis of Liability	Doctrinal Foundation under BNS 2023
Developers	Negligent design, failure to test foreseeable risks, reckless coding practices	Criminal negligence; breach of duty of care
Deploying Corporations	Governance failure, inadequate supervision, and compliance deficiencies	Corporate criminal liability; attribution principles
Operators	Misuse of AI systems; reckless override of safeguards	Knowledge-based or negligence-based offences
Regulatory Authorities (Exceptional Cases)	Gross systemic oversight failure	Institutional negligence (exceptional threshold)

This structured framework reinforces that criminal liability remains human-centered while accommodating distributed technological agency through differentiated duties for developers, deploying corporations, operators, and, in exceptional cases, regulatory authorities.

7. Why AI Should Not Be Granted Criminal Personhood

Some scholars have proposed granting artificial intelligence systems limited legal personhood in order to resolve attribution difficulties. However, doctrinal analysis demonstrates that such an approach is conceptually unsound within criminal jurisprudence. Criminal liability presupposes moral blameworthiness, a capacity for intent, and the ability to comprehend wrongdoing. AI systems, regardless of sophistication, do not possess consciousness, moral awareness, or volitional autonomy.

Furthermore, traditional theories of punishment - retribution, deterrence, and reform - cannot meaningfully operate when applied to AI. Retribution requires moral culpability; deterrence

presupposes rational fear of sanction; and reform depends on the possibility of moral improvement. None of these conditions are satisfied by algorithmic systems.

Sanctions imposed on AI would ultimately require translation into consequences for human actors or corporations, rendering AI personhood redundant. Accordingly, recognizing AI as a criminal person would distort foundational principles of criminal law rather than resolve accountability concerns.

8. Implications for Indian Criminal Law Under BNS 2023

The Bharatiya Nyaya Sanhita, 2023 (BNS 2023) provides a sufficiently adaptable doctrinal structure to address harms arising from artificial intelligence without requiring a complete legislative overhaul. Existing criminal law mechanisms - particularly negligence-based offenses, corporate criminal liability, abetment, conspiracy, and duty of care standards - are capable of accommodating AI-related misconduct. Where harm results from reckless system design, inadequate supervision, or failure to implement safeguards, liability can be established through principles already embedded in criminal jurisprudence. The framework does not demand that AI be treated as a legal person; instead, it allows attribution to human actors and corporate entities responsible for development and deployment.

Nevertheless, practical enforcement challenges remain significant. AI systems often involve complex evidentiary trails, making causation and fault difficult to prove beyond reasonable doubt. Additionally, cross-border AI development and deployment complicate jurisdictional control and regulatory oversight. Future reform may therefore focus not on redefining criminal subjects, but on strengthening preventive governance. This could include mandatory AI risk assessments, statutory audit obligations, and explicit criminal sanctions for reckless or grossly negligent deployment of high-risk AI systems.

Overall Discussion Synthesis

The findings of this study directly respond to the research objectives articulated in the Introduction. First, the analysis establishes that artificial intelligence systems cannot independently bear criminal liability because they lack moral agency, intentionality, and legal personality. Second, responsibility for AI-related harm remains fundamentally human-centered, attaching to developers, deployers, operators, and corporate entities. Third, the corporate criminal liability model emerges as the most coherent doctrinal framework for addressing AI harms, as it already accommodates non-natural actors acting through structured decision-making systems. Fourth, explainability and traceability are shown to be indispensable for legal accountability, particularly in satisfying evidentiary standards in criminal proceedings. Finally, the study demonstrates that the Bharatiya Nyaya Sanhita, 2023 is sufficiently flexible to address AI-related harm through principled statutory interpretation rather than radical legislative restructuring.

These conclusions align with the relational accountability framework advanced by Coeckelbergh [1], the responsibility-gap analysis discussed by Gunkel [10], and regulatory approaches examined by Turner [9]. However, this study extends existing scholarship by grounding these theoretical debates specifically within the doctrinal architecture of Indian criminal law. Its novelty lies in integrating responsibility-gap theory with statutory interpretation, systematically applying corporate liability analogies, and demonstrating the doctrinal adequacy of BNS 2023 for emerging AI-related criminal harms.

CONCLUSION

This study examined whether Artificial Intelligence systems can be held criminally liable and, if not, who should bear responsibility under contemporary criminal law, particularly within the framework of the Bharatiya Nyaya Sanhita, 2023. Through a doctrinal and comparative analysis, the research demonstrates that AI systems cannot independently satisfy the essential elements of criminal liability, especially the requirement of mens rea. Criminal law remains fundamentally anthropocentric, grounded in moral blameworthiness, intention, knowledge, and volition - qualities that AI systems do not possess.

The findings confirm that although AI may operate autonomously and produce harmful consequences, the so-called "responsibility gap" does not eliminate legal accountability. Instead, it shifts the focus of attribution from direct intention to derivative and systemic responsibility. The corporate criminal liability model provides the most coherent and legally sustainable framework for addressing AI-related harm. Just as corporations are held accountable for actions conducted through organizational structures, AI systems should be understood as technological instruments operating within human-controlled institutional environments. Consequently, developers, deploying corporations, and supervising operators must bear criminal responsibility where negligence, recklessness, or governance failure can be established.

The research further establishes that transparency and explainability are not merely ethical aspirations but legal necessities. Without traceability and auditability, criminal attribution becomes evidentially difficult, undermining justice and accountability. Therefore, responsible AI governance requires institutional oversight, risk assessment, and enforceable standards of care. Importantly, the Bharatiya Nyaya Sanhita, 2023, is doctrinally flexible enough to address AI-related criminal harm through existing provisions on negligence, corporate liability, and abetment. However, practical enforcement will require technical expertise, regulatory clarity, and proactive compliance mechanisms.

In conclusion, AI should not be granted criminal personhood. Instead, criminal liability must remain human-centered and institutionally structured. Future legal reform should focus on strengthening governance frameworks, mandating transparency obligations, and ensuring that technological innovation progresses within the boundaries of criminal accountability and substantive justice.

REFERENCES

[1] M. Coeckelbergh, "Artificial intelligence, responsibility attribution, and a relational justification of explainability," Science and Engineering Ethics, vol. 26, no. 4, pp. 2051 - 2068, 2020.

[2] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," Ethics and Information Technology, vol. 6, no. 3, pp. 175 - 183, 2004.

[3] L. Floridi and M. Taddeo, "What is data ethics?" Philosophical Transactions of the Royal Society A, vol. 374, no. 2083, 2016.

[4] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," Big Data & Society, vol. 3, no. 2, pp. 1 - 21, 2016.

[5] V. Dignum, "Responsible artificial intelligence: Designing AI for human values," IT Professional, vol. 19, no. 3, pp. 54 - 58, 2017.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1 - 38, 2019.

[7] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," Science, vol. 356, no. 6334, pp. 183 - 186, 2017.
[8] M. Taddeo and L. Floridi, "How AI can be a force for good," Science, vol. 361, no. 6404, pp. 751 - 752, 2018.
[9] J. Turner, Robot Rules: Regulating Artificial Intelligence. Cham, Switzerland: Palgrave Macmillan, 2018.
[10] D. J. Gunkel, "Mind the gap: Responsible robotics and the problem of responsibility," Ethics and Information Technology, vol. 20, no. 2, pp. 87 - 99, 2018.
[11] European Commission High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, Brussels, 2019.
[12] Parliament of India, Bharatiya Nyaya Sanhita, 2023, Act No. 45 of 2023, New Delhi, India, 2023. Available: https://www.indiacode.nic.in/handle/123456789/20062
[13] New York Central & Hudson River R.R. Co. v. United States, 212 U.S. 481 (1909).
[14] Tesco Supermarkets Ltd v Nattrass [1972] AC 153 (HL).
[15] State v Loomis, 881 N.W.2d 749 (Wis. 2016).
[16] Iridium India Telecom Ltd v Motorola Inc (2011) 1 SCC 74.
[17] Standard Chartered Bank v Directorate of Enforcement (2005) 4 SCC 530.
[18] Jacob Mathew v State of Punjab (2005) 6 SCC 1.
[19] European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng