Page: 2044-2063 *URL: www.lijdlr.com*

LAWFOYER INTERNATIONAL JOURNAL OF DOCTRINAL LEGAL RESEARCH

ISSN: 2583-7753

LawFoyer International Journal of Doctrinal Legal Research

[ISSN: 2583-7753]

Volume 4 | Issue 1

2026

DOI: https://doi.org/10.70183/lijdlr.2026.v04.89

RETHINKING MENS REA & CRIMINAL LIABILITY IN THE AGE OF ARTIFICIAL INTELLIGENCE

Swati Kumari[1]

I. ABSTRACT

Artificial Intelligence has moved beyond being a mere technological aid and now performs functions that involve independent decision-making, often with serious real-world consequences. This shift raises difficult questions for penal law, particularly in relation to the requirement of mens rea. While harm caused by AI systems can usually satisfy the element of actus reus, identifying a guilty mind becomes difficult when the actor is a non-human system lacking consciousness or intent. This paper examines whether existing principles of criminal liability are capable of addressing harms caused by AI, or whether their application reveals a structural problem. It analyses the problem of legal personhood in intelligent systems and evaluates different approaches to liability, including perpetration through another, natural and probable consequences, and direct liability of AI. using real incidents involving autonomous vehicles and AI-driven decision making, the paper argues that attributing criminal responsibility directly to AI risks weakening the moral basis of criminal law. Instead, it supports a framework that places responsibility on human actors involved in the design, deployment, and supervision of AI systems, while emphasising the need for preventive regulation to address emerging risks.

II. KEYWORDS

Artificial Intelligence, Criminal Liability, Mens Rea, Actus Reus, Autonomous Systems, Human Accountability.

III. INTRODUCTION

Artificial intelligence is no longer a futuristic concept; with the rapid advancement of exponential technologies, it has transitioned from a conceptual possibility to an

[1] Student, 4th year student at Bharati Vidyapeeth (deemed to be university), New Law College, Pune (India). Email: swatikri1908@gmail.com

 (ISSN: 2583-7753)

operational reality, permeating nearly every sphere of human life.[2] Today, computers and robots are replacing tasks once performed by humans. For a considerable period, computers used to work merely as a tool, just like screwdrivers or telephones, executing predefined instructions without any independent judgement, but their growing sophistication led to systems that no longer simply "think" for humans but operate as thinking machines, known as artificial intelligence.[3]

AI refers to the capacity of a system to perform a task that would ordinarily require human intelligence, using algorithms to analyse and synthesise the vast amount of data, adapt to changing circumstances, and make predictions that facilitate or even fully automate decision-making without constant human interventions.[4] For example, self- driving cars collect data by identifying patterns, building models from those patterns and using them to make predictions or decisions.[5]

These capabilities enable self-driving cars to respond to real-time traffic conditions, recognise obstacles and anticipate the behaviour of other road users with minimal or no human control. Such autonomy shows that AI is now an active decision-maker rather than just a passive tool. This shift could lead to serious legal and safety issues. This raises an urgent question: can AI cause harm, and if it can, who will be held legally responsible?

In 2016, a man in Florida was killed while driving a Tesla that was operating in autopilot mode, a system that relies on computer vision technology for vehicle detection.[6] A similar incident took place in March 2018, where a woman was killed by

[2]Ankit Kumar Padhy & Amit Kumar Padhy, Criminal Liability of the Artificial Intelligence Entities, Manupatra Articles (July 2019), https://docs.manupatra.in/newsline/articles/Upload/4e5c9c80-320b-4433-9f87-f56059a5345c.pdf

[3]Hallevy, Gabriel, The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control, 4 Akron Intell. Prop. J. 171, 171 (2010).

[4]Preethiya, T., Priyanga Subbiah, T. Pandiarajan, Karthikeyan Subramanian, Prince Chelladurai & C. Selvalakshmi, Autonomous System and AI, 1 (eds., 2024)

[5]Gless, S., Silverman, E. & Weigend, T., If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability, 19 New Crim. L. Rev. 412, 412 (2016)

[6] National Transportation Safety Board, *Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida*, NTSB Highway Accident Report HWY16FH018 (2017).

an autonomous car operated by Uber in Tempe, Arizona.[7] In another tragic incident, a man from the USA killed himself and his mother, after a conversation with ChatGPT, which made him believe that his mother might be spying on him and might attempt to kill him by poisoning.[8]

These incidents underscore the growing tension between technological autonomy and traditional principles of existing criminal liability, as the existing legal framework struggles to determine whether responsibility should rest with the manufacturer, the software developer, the operator, or the human user of the system. Harm caused by an autonomous system often does not have a clear human actor who carried out the prohibited act and had the necessary guilty mindset.

Criminal law relies on blaming the person through the two requirements of actus reus and mens rea. This situation makes it especially challenging. The main difficulty lies in proving mens rea, even though both parts are needed to establish a crime. It is especially challenging to determine the intent of a non- human entity, since an AI system lacks consciousness.[9]

A. Research Objectives

This paper aims to examine the compatibility of traditional principles of criminal liability, particularly actus reus and mens rea, with harms caused by artificial intelligence systems. It seeks to analyse whether existing legal doctrines can adequately address AI-mediated harm or whether they reveal structural limitations. The paper further aims to evaluate competing models of liability attribution and to identify a coherent framework that preserves the moral foundations of criminal law while addressing emerging technological risks.

[7] Troy Griggs & Daisuke Wakabayashi, *How a Self-Driving Uber Killed a Pedestrian in Arizona*, N.Y. Times (Mar. 20, 2018).
[8] Julie Jargon & Sam Kessler, A Troubled Man, His Chatbot and a Murder-Suicide in Old Greenwich, The Wall Street Journal (August 29, 2025, 8:00 AM)*, https://www.wsj.com/tech/ai/chatgpt-ai-stein-erik-soelberg-murder-suicide-6b67dbfb
[9] **N**ora Osmani, "The Complexity of Criminal Liability of AI Systems," Masaryk University Journal of Law and Technology (Vol. 14, No. 1) 53, 53–82 (2020)

B. Research Questions

This study is guided by the following core questions:

1. Whether existing principles of criminal liability are adequate to address harm caused by artificial intelligence systems?
2. How mens rea can be interpreted or attributed in cases involving autonomous AI decision-making?
3. Whether artificial intelligence systems should be granted legal personhood for the purpose of criminal liability?
4. Which model of liability attribution most effectively balances accountability, fairness, and the evolving nature of technology?

C. Research Methodology

This paper adopts a doctrinal and analytical research methodology, relying on the examination of existing legal principles, case law, and scholarly literature relating to criminal liability and artificial intelligence. It further incorporates a comparative analytical approach by evaluating different theoretical models of liability attribution, including perpetration through another, natural and probable consequences, and direct liability models. The analysis is supplemented by illustrative real-world incidents involving AI systems, with the objective of assessing how established legal doctrines operate in technologically mediated contexts.

IV. ARTIFICIAL INTELLIGENCE: MEANING AND APPLICATION

For the general public, "artificial intelligence" often conjures images of machines designed to replicate human abilities, especially cognitive functions. Such machines are commonly imagined as having human personality, thought process and decision-making capacity.[10] AI theorist Eliezer Yudkowsky[11] has observed that the greatest danger of artificial intelligence is the belief that it is fully understood. As a result, it is important to define it clearly. This first requires understanding the concept of general

[10] K. R. Walsh, S. Mahesh & C. C. Trumbach, "Autonomy in AI Systems: Rationalizing the Fears," J. Technol. Stud. Vol. 47, No. 1, 38, 38–47 (2021).
[11]Stephan De Spiegeleire, Matthijs Maas & Tim Sweijs, What Is Artificial Intelligence? HCSS Report (The Hague Centre for Strategic Studies) 25 (2017)

intelligence. The ability of an intelligent being to see, process, and respond to its environment is reflected in its five basic characteristics. The first one is communication.[12]

An intelligent entity is generally capable of communication,[13]which means it can understand the information and respond to it in a meaningful way. The second attribute is internal knowledge.[14]it means that the entity should know its own state, abilities and limitations and act accordingly after analysing it. For example, a human knows when they are tired or injured and adjusts their action accordingly. The third is external knowledge.[15] An intelligent entity is expected to understand the outside world, learn from it and use that knowledge to guide its actions.

The fourth attribute is goal- oriented behaviour[16], which means it can plan and take actions to achieve a specific objective. The last one is creativity.[17] It is the ability to adapt and find alternative solutions when an initial approach fails. When combined, these traits provide a helpful view of intelligence. Many of these traits can be found in modern artificial intelligence systems to different extents.

With AI's practical features, it can be used in many real-life scenarios. It has been used in different fields ranging from healthcare, law, stock trading, remote sensing, scientific discovery and many more.[18] As the global deployment of artificial intelligence reaches unprecedented heights, India has emerged as an active and significant participant in this technological transformation. This proactive stance is reflected in the AI for India 2030 initiative- co-hosted by the Ministry of Electronics and Information Technology, and the other key governmental and industry

[12]HALLEVY, supra note 2, at 175
[13] Id. at 175
[14] Id. at 176
[15] Id. at 176
[16] Id. at 176
[17] Id. at 176
[18] Shukla, S. & Vijay, J., Applicability of Artificial Intelligence in Different Fields of Life, International Journal of Scientific Engineering and Research (Int. J. Sci. Eng. Res.) Vol. 1, 28, 30 (2013).

stakeholders, which seeks to institutionalise the integration of AI within India's socio-economic framework.[19]

The Indian government approved the IndiaAI mission in 2024, allocating Rs. 10,300 crores over five years to boost the country's AI capabilities. This decision shows a strong commitment to the policy. In addition, AI has improved the healthcare industry significantly. As noted by NITI Aayog, AI has immense potential to strengthen India's healthcare delivery system.[20]

V. INDIAN LEGAL FRAMEWORK AND AI-MEDIATED HARM

The question of criminal liability for harm caused by artificial intelligence must also be examined within the framework of existing Indian law, particularly the Bharatiya Nyaya Sanhita, 2023, the Information Technology Act, 2000, and the Digital Personal Data Protection Act, 2023. At present, these statutes do not explicitly address liability arising from autonomous AI systems, thereby requiring reliance on general principles.

Under the Bharatiya Nyaya Sanhita, provisions relating to death caused by negligence and rash or negligent acts may potentially be invoked in cases involving AI systems, particularly where harm results from defective design, inadequate supervision, or improper deployment. However, these provisions presuppose human conduct and fault, making their application to autonomous systems indirect and often strained. Similarly, doctrines such as criminal conspiracy or common intention may be difficult to extend to AI systems, as they depend on shared mental states that cannot be attributed to non-human entities.

The Information Technology Act, 2000, while addressing issues of electronic governance and cyber offences, does not comprehensively regulate algorithmic decision-making or assign liability for harms caused by autonomous technologies. Its provisions may apply in limited circumstances involving data misuse or system

[19] Purushottam Kaushik, Harsh Sharma & Ayushi Sarna, AI for India 2030: A Blueprint for Inclusive Growth and Global Leadership, World Economic Forum (January 22, 2025), https://www.weforum.org/stories/2025/01/ai-for-india-2030-blueprint-inclusive-growth-global-leadership/

[20] Karan Kashyap, How AI Is Impacting India's Healthcare Industry, Forbes (February 09, 2025, 10:29 AM EST), https://www.forbes.com/sites/krnkashyap/2025/02/09/how-ai-is-impacting-indias-healthcare-industry/

tampering, but they do not address broader questions of accountability in AI-driven harm. The Digital Personal Data Protection Act, 2023, introduces obligations relating to data processing and consent, which may indirectly impact AI systems reliant on personal data; however, its focus remains on informational privacy rather than physical or criminal harm.

This fragmented legal position reveals a significant lacuna in Indian law, where existing statutes are not adequately equipped to deal with the complexities of AI-mediated harm. The absence of clear statutory guidance creates uncertainty in attributing liability and underscores the need for a more coherent legal framework that specifically addresses the role of artificial intelligence within the domain of criminal responsibility.

VI. THE COMPATIBILITY OF ACTUS REUS & MENS REA WITH ARTIFICIAL INTELLIGENCE

To determine criminal liability, two basic requirements must be fulfilled: the physical element- actus reus and the mental element- mens rea.[21]Actus reus refers to the conduct that constitutes the offence, whether by an act or by a failure to act.[22]In simpler terms, it is the physical behaviour that the law prohibits. This could mean deliberately breaking the law, such as harming someone or failing to take legal action. An omission only becomes a crime when someone has a legal duty to act and does not. For example, if a lifeguard does not save a drowning person, he could face criminal charges. An act must be voluntary—that is, a person must have control over their actions—in order to qualify as a criminal offence. Actions that are done without awareness or control are involuntary, and these actions do not satisfy the requirement of actus reus.[23]

The second element is mens rea or guilty mind. This term refers to the mental state or intention of a person when they commit the crime. A basic principle of criminal law is that a person is guilty of a crime only when a wrongful act occurs alongside a guilty

[21] PADHY & PADHY, supra note 1, at 17
[22] Akshaj Garg, Concurrence of Actus Reus and Mens Rea: A Descriptive Analysis, Indian Journal of Legal Review (Ind. J. Legal Rev.) Vol. 4, No. 2, 1312, 1313 (2024)
[23]Id. at 1313

mind. This idea is expressed in the maxim 'actus non facit reum nisi mens sit rea', which means an act does not make a person criminally liable, unless it is done with the culpable state of mind.[24] As emphasised by H.L.A. Hart, the doctrine of mens rea is central to criminal responsibility because criminal law is concerned not merely with preventing harm but with holding individuals accountable for actions they had the capacity to control and choose.[25]

It is well known that for a crime to happen, both actus reus and mens rea must be present. Since you can identify a crime by the occurrence of a harmful act, actus reus is relatively easy to determine. For instance, an autonomous AI-driven vehicle may cause injury to the pedestrian, a surgical robot governed by AI may commit an error resulting in serious harm to a patient, oran AI driven autonomous weapon may be responsible for the killing of humans.[26]

In each of these examples, the actus reus element is evident, as it is clearly identifiable that the action performed by the AI system directly resulted in harm to a human being. However, determining mens rea is far more challenging because it relates to what is going on inside a person's mind. What someone was thinking when the crime happened cannot be directly seen, measured, or proven. Because of this, courts do not observe mens rea directly; instead, they infer it from external evidence, such as a person's actions, words, behaviour before and after the act and surrounding circumstances.[27] As a result, courts rely on indirect indicators to determine intent, concluding how a person acted, what they said and the context in which the act occurred. This method works fairly well for human offenders. Their behaviour can show conscious choice and intention.

[24]Law Faculty, University of Delhi, Law of Crimes-I: Bharatiya Nyaya Sanhita, 2023, lawfaculty.du.ac.in (2024), https://lawfaculty.du.ac.in/userfiles/downloads/LLBCM/Law%20of%20Crimes-I%20BNS%202024.pdf

[25]Robert A. Wasserstrom, H. L. A. Hart and the Doctrines of Mens Rea and Criminal Responsibility, 35 U. Chi. L. Rev. 92, 95 (1967).

[26]Raajdwip Vardhan, Between Code and Culpability: Deciphering the Possibility of AI Mens Rea for Criminal Liability Through Juristic Personhood for AI, 4 Panj. Univ. L. Mag. (MagLaw) 68, 76 (2025).

[27]Id. at 77

However, when we apply this approach to AI, it becomes problematic. AI systems do not have intention or belief, and they lack legal personhood. Although such a system may act autonomously and perform complex tasks, the law does not recognise them as a person capable of bearing rights and duties.[28] Instead, AI is treated as a tool used by individuals and organisations that possess legal capacity.[29]

VII. THE PROBLEM OF PERSONHOOD IN INTELLIGENT SYSTEMS

Humans have long defined themselves by distinguishing their abilities from those of other beings, particularly in terms of intelligence, awareness and the ability to learn. Law reflects this distinction by giving rights and duties to humans and recognising them as "persons", setting them apart from all other creatures.[30]According to Black's Law Dictionary, a person, in legal theory, refers to any being whom the law recognises as capable of holding rights and duties.[31]

The concept of legal personality is no longer confined to natural persons, as recognised in a landmark decision of Salomon vs Salomon & Co. ltd.[32], where it was held that a company is a separate legal entity with an identity separate of its members and shareholders and becomes an artificial person upon incorporation, a position that has also been affirmed in the Indian context in Chiranjit Lal Chowdhuri v Union of India and Others, AIR 1951 SC 41 (SC), where the Supreme Court observed that the fundamental rights guaranteed under the Constitution are available not only to individual citizens but also corporate bodies.[33]

In the 1990s, some scholars discussed whether AI should be treated as a legal person and how responsibility for its actions could be decided, but this discussion remained

[28] David C. Vladeck, Machines Without Principals: Liability Rules and Artificial Intelligence, 89 Wash. L. Rev. 117, 121 (2014)
[29]VLADECK, supra note 26, at 122.
[30] Katherine B. Forrest, The Ethics and Challenges of Legal Personhood for AI, 133 Yale L.J. Forum 1175, 1175 (2024)
[31]Visa A. J. Kurki, A Theory of Legal Personhood pg. 89 (Oxford Legal Philosophy Series, 1st ed. 2019)
[32] Salomon v Salomon & Co Ltd, (1896) 11 WLUK 76 (1897) (UK)
[33]Anushka Rao & Vanshika Jain, The Conundrum of Legal Personhood in the Realm of Artificial Intelligence, Manupatra (October 9, 2024), https://articles.manupatra.com/article-details/The-Conundrum-of-Legal-Personhood-in-the-Realm-of-Artificial-Intelligence

mostly theoretical because the technology was not advanced enough at the time.[34] Today, however, the rapid development in machine learning has transformed this debate into a practical legal concern, as AI systems increasingly operate with limited human control and are capable of causing real-world harm. AI lacks consciousness and true agency, which makes it difficult to treat it as a legal person and also complicates the clear attribution of responsibility for its actions to either the system itself or solely to its designers, developers, or users.[35]

The question of whether an entity should be granted juristic personality is therefore inseparable from the question of whether it can meaningfully bear rights and duties,[36]particularly within the framework of criminal law, which is concerned with blame, responsibility and punishment. Proponents of extending personhood to artificial intelligence argue that such recognition would promote accountability and responsibility, provide greater clarity and predictability in assigning liability, and allow AI systems to be represented within legal and ethical decision-making processes.[37]

Critics, however, caution that granting personhood to AI risks anthropomorphising technology and may lead to a problematic shifting of responsibility away from human actors, especially in cases involving unintended or unforeseen consequences.[38] The issue of according legal identity to AI systems is further complicated by the diversity of technologies and the degree of autonomy they exhibit, all of which point towards a gradual shift that challenges the adequacy of traditional liability regimes; until technological developments compel the judiciary to reconsider existing legal frameworks, litigation is likely to continue on the assumption that responsibility rests

[34]Andreas Nanos, Criminal Liability of Artificial Intelligence, Charles Univ. Fac. L. Rsch. Paper No. 2023/III/3, 1, 5 (2023)
[35]Pranjal Chaturvedi & Dr. Suhasini, AI and Personhood: Navigating Legal Rights and Responsibilities in the Age of Intelligent Machines, LiveLaw (February 27, 2025), https://www.livelaw.in/lawschool/articles/ai-personhood-human-consciousness-chatbot-information-technology-act-european-union-ks-puttaswamy-national-strategy-for-artificial-intelligence-chatgpt-285068
[36]Lawrence B. Solum, Legal Personhood for Artificial Intelligence, SSRN Elec. J. 1231, 1239 (2007)
[37]NANOS, supra note 33, at 5
[38]Id. at 5

primarily with human principals, including developers, manufacturers, and owners who choose to deploy such systems.[39]

VIII. CRIMINAL RESPONSIBILITY & ATTRIBUTION OF LIABILITY

Criminal law has traditionally proceeded on the assumption that liability must be attributed to an identifiable actor who commits a prohibited act with the requisite mental state. However, when AI systems cause harm, this assumption is placed under strain, as the conduct giving rise to the offence is often executed by a technological intermediary rather than directly by a human actor. To reconcile AI-mediated harm with the established doctrine of criminal responsibility, it becomes necessary to adopt a structured analytical framework capable of locating culpability within existing principles of liability attribution. On this basis, three distinct models may be employed to assess criminal responsibility in cases involving AI:

A. The Perpetration via Another Liability Model

This is the first model, under which artificial intelligence is treated as an innocent agent or instrument.[40] Under this model, the AI system is originally designed and programmed to perform lawful or beneficial functions, but owing to inappropriate use, manipulation or contextual deployment by human actors, it departs from its intended purpose and becomes the means through which the wrongful act is ultimately committed.[41]

When AI systems are used by humans, responsibility of any harm they cause should lie with those who deploy or control them, much like the principle that a person is liable for the acts of an agent on theri behalf; since AI lacks intention, its use should be judged by objective standards such as negligence, strict liability, or heightened duties in fiduciary roles, and human actors should not be allowed to lower their duty of care or escape liability simply by replacing a human agent with an AI system.[42]

[39]Id. at 5
[40] HALLEVY, supra note 2, at 179
[41]Sadaf Fahim & G. S. Bajpai, AI and Criminal Liability, 1 Indian J. Artificial Intel. & L. 64,68 (2020).
[42]Ian Ayres & Jack M. Balkin, The Law of AI Is the Law of Risky Agents without Intentions, SSRN Elec. J. 1, 2 (2024)

In this context, just as a principal may be liable for failing to properly supervise or train an agent, humans may similarly be held legally responsible for negligently designing, training, or deploying inadequately regulated AI systems that cause harm, with liability assessed through a risk-utility approach akin to that applied in defective design cases.[43]

However, the perpetration via another liability model is inappropriate where an AI system independently commits an offence based on its own learning or functions as a semi- innocent agent, but it may apply when a programmer or user merely uses the AI instrumentally without relying on its advanced capabilities, in which case criminal liability rests entirely on the human actor and not on the AI.[44]

B. The Natural-Probable-Consequence Liability Model

The Natural-Probable-Consequence doctrine[45] holds a person criminally liable when they intentionally assist in one offence, but their assistance foreseeably results in the commission of a different offence that they did not specifically intend.[46] In State v. Kaiser [47], the court held that the defendant may be held criminally liable for consequences that are the natural and probable result of their conduct, even if the precise harm was not specially intended, reinforcing the principle that foreseeability rather than direct intent can ground criminal liability.

This reasoning can be extended to artificial intelligence systems, as under this model, an AI system is initially programmed to perform lawful or beneficial tasks, but due to inappropriate use, contextual deployment, or operational conditions, it departs from its intended purpose and causes harm.[48] A commonly cited illustration is provided by legal scholar Hallevy, who refers to an incident in a Japanese motorcycle factory where

[43]Id. at 2
[44]HALLEVY, supra note 2, at 181
[45] Id. at 181
[46]Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities*, 4 Akron Intell. Prop. J. 171, 176 (2010).
[47]State v. Kaiser, (1996) 260 Kan. 235 (1996) (U.S.)
[48]FAHIM & BAJPAI, supra note 40, at 69

an artificially intelligent robot killed a human worker.[49] The robot, while performing its assigned task, mistakenly perceived the employee as a threat to the successful completion of its operation, and, acting on this erroneous assessment, used its hydraulic arm to push the worker into a nearby operating machine, killing him instantly before resuming its duties.[50] This example demonstrates how an AI system, despite lacking intention, can independently cause serious harm when its decision-making processes malfunction or misclassify human presence.

Another illustration of this liability model arises in the context of an AI system designed to function as an automatic pilot and programmed to prioritise the successful completion of a flight mission[51]. After the human pilot activates the autopilot during the flight, the pilot later detects an approaching storm and attempts to abort the mission and return to base. The AI system interprets the pilot's attempt to override the mission as an ejection mechanism, ultimately resulting in the pilot's death. Although the programmer had no intention to cause harm, let alone to kill the human pilot, the death nevertheless occurred as a direct result of the AI system's action, which was carried out strictly in accordance with its programmed mission objectives.

In both illustrations, the harmful outcome was not the result of any direct human intention to cause death, but rather a foreseeable consequence of deploying AI systems with autonomous decision-making authority in environments involving close human interaction. This places the cases squarely within the natural-probable-consequence liability model, under which criminal or civil responsibility is attributed to human actors where the resulting harm could reasonably have been anticipated at the stage of design, programming, or deployment of the AI system.

However, the application of this model produces different legal outcomes depending on the facts: in cases of negligent programming or use without criminal intent, liability

[49] When an AI Finally Kills Someone, Who Will Be Responsible?, MIT Technology Review (March 12, 2018), https://www.technologyreview.com/2018/03/12/144746/when-an-ai-finally-kills-someone-who-will-be-responsible/
[50] HALLEVY, supra note 2, at 172
[51] Id. at 180

is based on foreseeability and lack of due care, whereas in cases where programmers or users intentionally employ an AI system to commit one offence but the AI deviates and commits another, liability may still attach if the resulting offence was a natural and probable consequence of their conduct.[52]

Hallevy goes further and contoversially argues that where AI does not merely work as an innocent tool, but instead exercises a degree of independent decision-making, criminal liability should not rest solely on the programmer or user under the natural–probable-consequence doctrine; rather, the AI entity itself may also be considered directly responsible for the specific offence it commits, in addition to any liability attributed to the human actors.[53]

C. Direct Liability Model

Kingston explicitly recognises a direct liability model of artificial intelligence in which an AI system itself may be treated as the subject of criminal liability.[54]Drawing on Hallevy's framework, he explains that actus reus can be attributed to an AI system where it either performs an action that results in a criminal offence or fails to act when there is a legal duty to do so. While acknowledging the difficulty of attributing mens rea to AI, Kingston notes that this obstacle does not arise in the case of strict liability offences, where no intention is required.[55] Accordingly, he suggests that AI systems may be held criminally liable for such offences, using the example of a self-driving car exceeding the speed limit, where liability could attach directly to the AI program operating the vehicle. This approach treats the AI system as the immediate perpetrator of the offence, without the need to establish intermediary human intent.

Some scholars take a more radical position by arguing that AI should be treated as directly criminally liable.[56] Using the example of social media algorithms that select and recommend content, they suggest that AI systems make autonomous decisions based on user data and behavioural patterns, without direct human control. When

[52]Id. at 180
[53]Id. at 180
[54]John Kingston, Artificial Intelligence and Legal Liability, arXiv:1802.07782 [cs], 1, 4 (2018)
[55]Id. at 4
[56]Danila Kirpichnikov, Albert Pavlyuk, Yulia Grebneva & Hilary Okagbue, Criminal Liability of the Artificial Intelligence, 159 E3S Web Conf. 04025, 04025–04034 (2020)

users instruct the system to "see fewer such publications," the algorithm changes its future behaviour, which is said to resemble how criminal sanctions influence and correct human conduct. From this perspective, AI systems are capable of modifying their behaviour, preventing future harmful outcomes, and responding to regulatory pressure. On this basis, these scholars argue that AI should not be viewed merely as a tool, used by humans, but as an autonomous entity capable of being directly subject to criminal responsibility.[57]

This possibility raises several unresolved legal questions.[58] If an AI system is treated as directly liable, it becomes necessary to consider whether traditional criminal defences could meaningfully apply to it. For instance, it is unclear whether a malfunctioning AI system could invoke a defence analogous to human insanity, or whether an AI affected by an electronic virus could claim a defence comparable to coercion or intoxication.[59] the direct liability model also raises the further question of attribution of responsibility, particularly whether liability should rest solely with the AI system, or whether it should continue to extend to programmers, designers, or operators who created, deployed, or failed to adequately secure the system. These unresolved issues highlight the significant doctrinal and practical challenges that accompany any attempt to impose direct criminal liability on AI.

D. Comparative Perspective: The European Union Artificial Intelligence Act

A significant development in the regulation of artificial intelligence is the enactment of the EU AI Act, which entered into force on 1 August 2024 and represents the first comprehensive binding legal framework governing AI systems. The Act adopts a risk-based classification model, categorising AI systems into unacceptable risk, high-risk, limited risk, and minimal risk. Notably, Annex III of the Act identifies high-risk AI systems, including those deployed in critical sectors such as transportation, healthcare, and law enforcement, which are subject to stringent regulatory obligations. Under this framework, providers, deployers, and distributors of high-risk AI systems

[57]Id. at 7
[58]FAHIM & BAJPAI, supra note 40, at 74
[59]Id. at 74

are required to comply with detailed obligations relating to risk management, data governance, transparency, human oversight, and post-market monitoring.

Although the Act primarily operates within a regulatory and administrative law framework, it implicitly allocates responsibility by imposing compliance duties on identifiable human and corporate actors involved in the lifecycle of AI systems. This structured allocation of obligations serves as a functional substitute for traditional fault-based liability by emphasising ex ante risk prevention and accountability.

In contrast, the Indian legal framework currently lacks a dedicated statute addressing artificial intelligence, relying instead on general principles under criminal law and statutes such as the Information Technology Act, 2000. This creates uncertainty in attributing liability for AI-mediated harm, particularly in cases involving autonomous decision-making. The EU approach demonstrates how a preventive, risk-based regulatory model can complement traditional liability doctrines by clearly identifying responsible actors and imposing proactive compliance requirements. Incorporating similar regulatory strategies within the Indian context could strengthen accountability mechanisms while preserving the foundational principles of criminal responsibility.

IX. PUNISHMENT OF ARTIFICIAL INTELLIGENCE: SCOPE AND LIMITATIONS

Punishment presents one of the most complex challenges in responding to harm caused by artificial intelligence, as traditional criminal sanctions presuppose a human offender capable of understanding legal commands and moral condemnation. Scholars such as Lemley and Casey highlight those conventional punitive measures, including fines, injunctions, and punitive damages, are difficult to apply to AI systems because non-human entities cannot internalise blame, fear punishment, or consciously alter their behaviour in response to legal censure.[60]

They therefore argue that the law may need to rethink remedies for AI by shifting the focus from moral blameworthiness toward functional control and the prevention of

[60]Mark A. Lemley & Bryan Casey, Remedies for Robots, 86 U. Chi. L. Rev. 1311, 1342 (2019).

harm.[61] On this basis, the authors suggest that certain sanctions directed at AI, such as permanently disabling or withdrawing a harmful system from operation, sometimes described as a "robot death penalty", may operate as forms of specific deterrence. However, such measures function more as preventive or regulatory controls than as genuine criminal punishment, and their implementation would likely face resistance within common law systems that tie punishment closely to moral agency and human culpability, thereby underscoring the need for significant conceptual reconfiguration before traditional penal objectives can be meaningfully applied to AI.

X. SUGGESTIONS AND RECOMMENDATIONS

In order to address the limitations of existing criminal law frameworks in dealing with artificial intelligence, it is proposed that legislative intervention be undertaken to explicitly recognise and regulate AI-mediated harm. Specific amendments may be introduced within the Bharatiya Nyaya Sanhita, 2023 to incorporate provisions clarifying liability in cases involving autonomous systems, particularly by recognising negligence, recklessness, and failure of oversight in the design and deployment of such technologies. Similarly, the Information Technology Act, 2000 may be expanded to include obligations relating to algorithmic accountability, transparency, and auditability.

A structured framework for liability allocation should be developed, distinguishing between developers, deployers, and operators of AI systems. Developers should bear responsibility for defects in design and training data, deployers for improper implementation and contextual misuse, and operators for failures in supervision or intervention. This layered approach ensures that liability is assigned based on the degree of control and foreseeability associated with each actor.

Further, a regulatory regime for pre-deployment certification of high-risk AI systems should be established, particularly in sectors such as autonomous transportation, healthcare, and defence. Such a regime should mandate safety testing, risk

[61]Id. at 1389

assessment, and compliance with minimum technical and ethical standards before public deployment. Additionally, periodic audits and post-deployment monitoring mechanisms should be instituted to ensure continued compliance and to mitigate evolving risks. These measures would enable a shift from reactive liability to preventive governance, while preserving the foundational principles of criminal accountability.

XI. CONCLUSION

The increasing autonomy of AI challenges the core principles of criminal law, which rest on human intention, control, and moral responsibility. While harm caused by AI systems can often satisfy the requirement of actus reus, attributing mens rea remains inherently difficult, as AI lacks consciousness and the capacity to form intent. This disconnect exposes the limits of applying traditional criminal liability frameworks to technologically mediated harm.

As this paper argues, the most coherent legal response is to locate responsibility with human actors involved in the design, deployment, and supervision of AI systems, rather than attributing criminal personhood to the technology itself. Liability models based on foreseeability and negligence allow criminal law to retain its moral foundation while addressing the risks posed by autonomous systems. Until AI can meaningfully bear rights and duties, criminal law must continue to prioritise human accountability, supported by regulatory safeguards aimed at preventing harm and ensuring responsible innovation.

XII. BIBLIOGRAPHY

A. Books and Monographs

1. H.L.A. Hart, *Punishment and Responsibility: Essays in the Philosophy of Law* (2nd edn, Oxford University Press 2008).
2. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th edn, Pearson 2021).

B. Journal Articles

1. Gabriel Hallevy, 'The Criminal Liability of Artificial Intelligence Entities' (2010) 4 Akron Intellectual Property Journal 171.
2. David C. Vladeck, 'Machines Without Principals: Liability Rules and Artificial Intelligence' (2014) 89 Washington Law Review 117.
3. Mark A. Lemley and Bryan Casey, 'Remedies for Robots' (2019) 86 University of Chicago Law Review 1311.

C. Cases

1. *Salomon v A Salomon & Co Ltd* [1897] AC 22 (HL).
2. *Chiranjit Lal Chowdhuri v Union of India and Others* AIR 1951 SC 41 (SC).

D. Legislation and Statutory Instruments

1. Bharatiya Nyaya Sanhita, 2023 (India).
2. Information Technology Act, 2000 (India).
3. Digital Personal Data Protection Act, 2023 (India).
4. Regulation (EU) 2024/1689 (Artificial Intelligence Act) (European Union).

E. Reports and Institutional Materials

1. National Transportation Safety Board, *Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida* (Highway Accident Report HWY16FH018, 2017).
2. NITI Aayog, *National Strategy for Artificial Intelligence* (Government of India, 2018).

F. Newspaper Articles

1. Troy Griggs and Daisuke Wakabayashi, 'How a Self-Driving Uber Killed a Pedestrian in Arizona' *The New York Times* (20 March 2018).

G. Online and Policy Sources

1. Ministry of Electronics and Information Technology, Government of India, 'IndiaAI Mission' (2024).
2. European Commission, 'Artificial Intelligence Act' (2024).