INDIAN JOURNAL OF LEGAL REVIEW

VOLUME 5 AND ISSUE 7 OF 2025

INSTITUTE OF LEGAL EDUCATION

INDIAN JOURNAL OF LEGAL REVIEW

APIS – 3920 – 0001 | ISSN - 2583-2344

(Open Access Journal)

Journal's Home Page – https://ijlr.iledu.in/

Journal's Editorial Page - https://ijlr.iledu.in/editorial-board/

Volume 5 and Issue 7 of 2025 (Access Full Issue on - https://ijlr.iledu.in/volume-5-and-issue-7-of-2025/)

Publisher

Prasanna S,

Chairman of Institute of Legal Education

No. 08, Arul Nagar, Seera Thoppu,

Maudhanda Kurichi, Srirangam,

Tiruchirappalli – 620102

Phone : +91 94896 71437 - info@iledu.in / Chairman@iledu.in

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 5 AND ISSUE 7 OF 2025

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

ARTIFICIAL INTELLIGENCE, LEGAL PERSONHOOD, AND DETERMINATION OF CRIMINAL LIABILITY

AUTHOR - MS. SETIKA PRIYAM* & DR. KUNVAR DUSHYANT SINGH**

* STUDENT AT AMITY LAW SCHOOL, AUUP, LUCKNOW

** ASSISTANT PROFESSOR AT AMITY LAW SCHOOL, AUUP, LUCKNOW

BEST CITATION - MS. SETIKA PRIYAM & DR. KUNVAR DUSHYANT SINGH, ARTIFICIAL INTELLIGENCE, LEGAL PERSONHOOD, AND DETERMINATION OF CRIMINAL LIABILITY, INDIAN JOURNAL OF LEGAL REVIEW (IJLR), 5 (7) OF 2025, PG. 212-222, APIS – 3920 – 0001 & ISSN - 2583-2344

ABSTRACT

The broad adoption of artificial intelligence (AI) across vital domains ranging from autonomous vehicles and financial markets to healthcare diagnostics and legal analytics has exposed significant gaps in our legal systems when AI-driven errors or malfunctions cause harm. Autonomous systems often involve multiple stakeholder hardware suppliers, software developers, sensor manufacturers, and corporate overseers making it difficult to pinpoint who is responsible for a system's failure. The 2018 Uber autonomous-vehicle crash in Tempe, Arizona, where a pedestrian was misclassified repeatedly by the AI's perception module and the emergency braking function was disabled, underscores this challenge: with safety overrides turned off and state oversight minimal, liability became entangled among engineers, operators, and corporate policy not the machine alone.

Traditional criminal law doctrines rest on *actus reus* (the guilty act) and *mens rea* (the guilty mind), both premised on human agency and intent. AI entities, however, can execute complex decision-making without consciousness or moral awareness, creating a "responsibility gap" under current frameworks. To bridge this gap, scholars like Gabriel Hallevy have proposed three liability models—perpetration-via-another (holding programmers or users accountable), the natural-probable-consequence model (liability for foreseeable harms), and direct liability (attributing responsibility to AI itself if it meets legal thresholds for *actus reus* and an analogue of *mens rea*). Each model offers insight but struggles with AI's semi-autonomous nature and opacity.

This paper argues against prematurely conferring legal personhood on AI an approach that risks absolving human actors and diluting accountability. Instead, it advocates for a human-centric policy framework that combines clear oversight duties, mandated explainability measures, and calibrated negligence or strict-liability standards for high-risk AI applications. Such reforms are especially urgent in jurisdictions like India, where AI governance remains nascent. By anchoring liability in human oversight and regulatory clarity rather than on machines themselves, we can ensure that accountability evolves in step with AI's growing capabilities, safeguarding both innovation and public safety.

Keywords: Artificial Intelligence, Criminal Liability, Legal Personhood, Actus Reus, Mens Rea, Vicarious Liability, AI Regulation

1. INTRODUCTION

Artificial intelligence has woven itself into almost every facet of modern life from the way we navigate cities and manage our finances to how diagnoses are made in hospitals and contracts are drafted in boardrooms. These AI-driven systems promise efficiency and insight, yet their growing autonomy raises an uncomfortable question: when an AI causes harm, who takes the fall? The very premise of criminal law linking a guilty act (*actus reus*) with a guilty mind (*mens rea*) relies on human intention and control, neither of which neatly applies to self-learning algorithms and autonomous machines.

Traditional liability models strain at the seams when an accident involves no distracted driver or negligent technician, but code that "learned" to misidentify a pedestrian or a drone that "decided" to breach no-fly zones. Real-world incidents, such as the 2018 Uber autonomous-vehicle crash in Tempe, Arizona, illustrate this challenge vividly: a system designed to brake failed to recognize a pedestrian first as an unknown object, then as a car and finally as a bicycle and, critical safety overrides disabled, the sole human operator was engrossed in entertainment on her phone .

Meanwhile, AI's "black box" nature and rapid evolution confound regulators everywhere, but nowhere more acutely than in jurisdictions still crafting their first AI rules. In India, for instance, there is yet no cohesive policy framework to determine when liability should attach to developers, hardware suppliers, corporate owners or, controversially, the AI itself. This paper examines whether AI could ever be treated like a "legal person" with its own rights and duties, or whether existing doctrines such as vicarious liability, the "directing mind" theory borrowed from corporate law, and principles of strict liability can and should be adapted to plug the responsibility gap.

By weaving together Gabriel Hallevy's tripartite liability models (perpetration-via-another, natural-probable-consequence, and direct liability), comparative case studies, and a qualitative analysis of both Indian and international jurisprudence, this study argues for a human-centric approach: AI should remain a tool, not a bearer of moral blame. Only through clear policy on oversight duties, mandated explainability, and calibrated negligence standards can we ensure that, as AI continues to reshape our world, accountability remains firmly within human hands.

2. THEORETICAL FOUNDATIONS

2.1 Legal personhood and AI

Imagine for a moment how we treat corporations in law: they sign contracts, own property, even sue and be sued yet they are not "people" in any biological sense. Instead, the law grants them a fictional personality, a tool to bundle rights and responsibilities under a single legal umbrella. Could AI ever fit into that same mold? Scholars like Kurki and Pietrzy kowski have asked whether sophisticated software agents deserve a similar technical personhood, separating "personality" from humanity and examining whether AI could carry rights and duties in its own name.

Under Hohfeld's framework, every right entails a corresponding duty, so granting personhood to AI would automatically shift some legal burdens away from the human actors who build, program, or deploy it.

Building on this, Kelsen's "technical personification" theory describes legal personhood as merely a juridical device a convenient shorthand for organizing complex webs of accountability. If a self-driving car were a legal person, for instance, claims arising from a crash could proceed directly against the AI "entity" rather than through the manufacturer, software developer, or owner.

Yet there are strong reasons for caution. First, AI systems lack consciousness, moral awareness, and the ability to form intentions in any meaningful sense. Personhood, even in its legalistic form, traditionally presupposes some minimal capacity for agency an attribute that today's algorithms simply do not possess. Second, prematurely granting AI a standalone legal status risks diluting human responsibility, allowing engineers or corporations to hide behind their machines.

At this early stage of development when AI is still learning, iterating, and often opaque it may be more prudent to treat AI as a complex agent whose actions reflect back on its human creators and operators. By analogizing AI to a corporation without extending full personhood, we preserve the flexibility of existing doctrines (like vicarious liability or the "directing mind" theory) while ensuring that those who design and deploy these systems remain squarely in the legal spotlight.

2.2 Challenges with Criminal Law Application

At the heart of criminal law lie two indispensable pillars: *actus reus*, the physical commission of a prohibited act, and *mens rea*, the mental element or "guilty mind" behind it. These concepts presume a sentient actor someone who can choose, intend, and understand the wrongdoing. AI systems, however, upend these assumptions. When an autonomous vehicle or an algorithmic trading bot malfunctions, no human driver or trader is directly at the wheel; instead, layers of code and data drive the outcome. This raises the fundamental question: can we meaningfully attribute either *actus reus* or *mens rea* to a non-sentient machine?

2.2.1 Actus-Reus

For an act to qualify as *actus reus*, it must be voluntary and attributable to a legal person. While AI clearly performs observable actions braking too late, misclassifying a pedestrian, or executing an unauthorized transaction its "voluntariness" is a proxy for the designer's or operator's inputs. In practice, courts would have to trace every harmful outcome back through a chain of software updates, sensor inputs, and human overrides to identify which link actually "acted." This maze of responsibility complicates traditional causation analysis and challenges prosecutors to pinpoint who truly committed the guilty act.

2.2.2 Mens-Rea

Mens rea demands intentionality or, at minimum, willful blindness. AI, by contrast, lacks consciousness or moral awareness; it does not "intend" to harm, nor can it appreciate the wrongfulness of an act. Even the most advanced machine-learning model remains a statistical optimizer, not a moral agent. As such, insisting on a mental state for AI is akin to expecting a wind-up toy to form criminal intent an exercise in futility that leaves a yawning "responsibility gap" in the law.

2.2.3 Opacity and Foreseeability

Another hurdle is the so-called "black box" nature of many AI systems. When an AI-driven decision cannot be fully explained after the fact, it becomes nearly impossible to demonstrate that a designer or operator should have foreseen a specific harm. Traditional negligence or strict-liability regimes rely on clear standards of foreseeability and duty of care; with AI, those standards blur, as even experts may struggle to predict how complex models will behave in novel circumstances.

2.2.4 Multiplicity of Stakeholders

Finally, AI incidents rarely involve a single individual. Hardware manufacturers, data scientists, software engineers, service providers, and end-users may all play a role. Pinning criminal liability on just one actor risks overlooking the interconnected ecosystem that made the harm possible. Without tailored doctrines or statutory guidance, prosecutors face the daunting task of untangling who among several candidates deserves blame and whether any human at all truly "controlled" the outcome. Together, these challenges underscore why simply shoehorning AI into existing criminal law is ineffective. Rather than stretching *actus reus* and *mens rea* to their breaking point, jurisdictions will need innovative legal constructs blending human accountability, enhanced explainability requirements, and calibrated liability standards to ensure victims can obtain justice without unduly stifling technological progress.

3. LIABILITY MODELS FOR AI CONDUCT

3.1 Perpetration-via-Another Model

Under the **perpetration-via-another** model, the AI system is treated as an "innocent agent" that merely executes the intentions of a human actor. This approach, first articulated by Solum, imagines AI as comparable to a mechanical agent: while the machine carries out the physical act (*actus reus*), the criminal intent (*mens rea*) is attributed to the person who programmed, deployed, or directed it. Gabriel Hallevy expands on this by arguing that whenever an AI-driven action results in unlawful harm, responsibility should rest with the human "perpetrator-via-another" be it the software developer who wrote the faulty code or the operator who failed to supervise the system adequately.

This model works best for relatively simple AI tools whose behavior remains largely predictable and under human control for example, a logistics robot following a predefined route or a rule-based chatbot that issues erroneous legal advice. In such cases, foreseeability is clear: the human actor could have anticipated the harmful outcome and taken steps to prevent it. However, as AI systems gain adaptive learning capabilities and begin to operate in more dynamic environments, the perpetration-via-another model may struggle to account for genuinely autonomous deviations from programming, leading to what Hallevy terms a "semi-innocent agent" scenario that the model cannot neatly address.

3.2Natural-Probable-Consequence Model

The **natural-probable-consequence** model shifts the focus from direct control to **foreseeability** of harm. Under this approach, developers, programmers, or operators can be held liable for adverse outcomes that a "reasonable" person in their position should have anticipated—even if the AI system appeared to act on its own. In practical terms, if an autonomous drone or algorithmic trading bot causes damage in a way that seasoned designers or users could foresee as a natural consequence of its design or deployment, liability attaches to those human actors.

This model blends traditional negligence principles with accomplice-style liability:

- **Negligence** applies where the AI's harmful deviations were foreseeable risks that the human actor failed to guard against.
- **Accomplice liability** arises if the AI's actions, though independent, fall within the scope of risks the actor willingly enabled or failed to prevent.

For example, if a developer trains a facial-recognition system without accounting for known bias in the training data, and the system later misidentifies individuals resulting in wrongful arrests the developer could be held responsible under this model, since the possibility of misidentification was a predictable consequence of flawed data practices.

However, as AI models grow more complex and self-learning, determining what was truly "foreseeable" becomes challenging.

Black-box algorithms may hide emergent behaviors that even expert programmers could not predict, risking either over-extension of liability or an "all-or-nothing" outcome where no one is blamed because no one could have foreseen the specific failure. Addressing this requires clear industry standards for risk assessment, mandatory safety audits, and explainability mandates, so that foreseeability can be assessed against objective benchmarks rather than hindsight alone.

3.3Direct Liability Model

The **Direct Liability Model** pushes the envelope by treating the AI system itself as the "perpetrator" of a crime, rather than merely the instrument of a human actor. Under this approach, an AI entity is held directly responsible for unlawful conduct when it independently satisfies both elements of a crime: the physical act (*actus reus*) and a functional equivalent of the mental element (*mens rea*).

3.3.1 **Autonomous Actus Reus**

Here, the machine's own behavior whether it be an unsupervised drone strike, a self-navigating vehicle collision, or an algorithmic trading flash crash is treated as the criminal "act." Rather than tracing the fault back through layers of code or hardware, courts would consider the AI's decision-making process itself as the wrongful deed.

3.3.2 **Constructed Mens Rea**

To bridge the gap between human intent and machine operation, proponents have suggested technical proxies for *mens rea*. These might include evidence that the AI was programmed to "know" certain probabilities of harm, or that it was designed to bypass safety checks effectively "willfully disregarding" risk. In effect, the AI's internal objectives or error-handling rules stand in for conscious intent.

3.3.3 **Advantages and Aspirations**

- **Accountability for Truly Autonomous Harms**: When an AI system deviates in unforeseeable ways beyond the programmer's or operator's reasonable foresight—the direct model ensures some entity is held to account.
- **Symmetry with Corporate Personhood**: Just as corporations can be prosecuted for crimes they "commit," so too could AI be recognized as a non-human legal actor capable of bearing liability.

3.4 Practical and Ethical Hurdles

- **Absence of Genuine Agency**: AI lacks consciousness and moral understanding, making any ascribed *mens rea* a legal fiction rather than a reflection of real intent.
- **Risk of Diluting Human Responsibility**: By redirecting blame to the machine, engineers, corporations, and operators may escape scrutiny, undermining the very accountability the model seeks to reinforce.
- **Legal and Technological Readiness**: Implementing direct liability would require upheaval of criminal codes—creating new statutes that define AI "persons," set out AI-specific intent standards, and establish sentencing regimes for machines.

In sum, while the Direct Liability Model offers a bold conceptual framework one that recognizes the growing autonomy of AI it remains more of a theoretical exercise than a pragmatic solution. For the foreseeable future, grounding liability in human and corporate actors, supported by robust oversight and clear regulatory duties, provides a more reliable path to justice and public safety.

4. ANALYSIS OF *ACTUS REUS* AND *MENS REA* IN AI-DRIVEN CRIMES

When we talk about crime, *actus reus* the external, physical component of wrongdoing is the bedrock of guilt. Yet in AI-driven incidents, pinpointing who "pulled the trigger" becomes a tangled affair. Autonomous systems from delivery drones that deviate mid-flight to high-frequency trading bots that spark flash crashes commit observable acts. However, those acts emerge from layers of sensor inputs, machine-learning models, and software updates rather than a single human hand. In Saudi's systematic review, *actus reus* is described as encompassing both positive acts and omissions, highlighting how the absence of a proper safety override or human fallback can itself give rise to criminal liability. Thus, courts must unravel a web of code, data, and design choices to trace which link in the chain truly "acted" in a legally cognizable sense.

Equally challenging is *mens rea*, the mental element of crime that demands intent, knowledge, or recklessness. AI systems, however sophisticated, lack consciousness or moral insight; they do not "intend" harm, nor can they appreciate the wrongfulness of their outputs. As Hallevy points out, attributing *mens rea* to an algorithm verges on legal fiction, since AI cannot form beliefs or desires in the human sense. Even when an AI model is explicitly programmed to optimize for profit despite known risks it remains a cold optimizer rather than a moral agent. Consequently, insisting on traditional *mens rea* thresholds risks leaving truly autonomous harms unpunished, widening the so-called "responsibility gap."

The opacity of many AI architectures compounds these difficulties. "Black-box" models may behave unpredictably in novel contexts, rendering harm unforeseeable even to their creators. Under negligence or strict-liability regimes, foreseeability is pivotal to assigning fault but with inscrutable AI decisions, neither prosecutors nor defendants can reliably demonstrate what a "reasonable" designer should have foreseen. This threatens to either over-extend liability to every AI stakeholder or to collapse it entirely when no human actor can credibly claim control.

In sum, AI-driven crimes dismantle the neat pairing of *actus reus* and *mens rea* on which criminal law depends. To bridge this divide, legal systems must develop tailored mechanisms such as mandated explainability, human-in-the-loop safeguards, and calibrated negligence standards so that the physical act and mental fault underlying AI harms can once again map onto accountable human decision-makers.

5. CASE STUDIES AND PRACTICAL COMPLEXITIES

Real-world incidents involving artificial intelligence have brought abstract legal theories into sharp focus, exposing the limitations of existing liability frameworks when AI systems cause harm or engage in criminal-like conduct. Two illustrative cases the **2018 Uber autonomous vehicle crash in Arizona** and the **Random Darknet Shopper experiment** highlight the complex interplay between software autonomy, human oversight, and legal accountability.

- **Uber's 2018 Fatal Crash: Diffused Responsibility and Systemic Oversight Failures**

In March 2018, a self-driving Uber test vehicle struck and killed a pedestrian in Tempe, Arizona. The fatal crash was not only the first of its kind involving a fully autonomous vehicle but also a chilling example of how distributed responsibility can paralyze accountability. Investigations revealed that the AI system had difficulty classifying the pedestrian it initially identified her as an unknown object, then as a vehicle, and finally as a bicycle. The system decided to initiate emergency braking but that feature had been **disabled** to avoid erratic driving behaviors during testing. Responsibility for applying the brakes had shifted to a human safety operator inside the vehicle, who, at the time of the crash, was found to be **distracted by a video on her phone**.

Beyond the driver, a tangled web of actors including software engineers, sensor manufacturers, and Uber's corporate policy-makers played roles in designing and implementing the system. With minimal regulatory oversight and unclear policy on operator training and system redundancy, liability could not be cleanly attributed to any single entity. The case exemplifies how AI incidents often involve **multiple stakeholders**, making it difficult to apply traditional doctrines like sole criminal liability or negligence.

- **Random Darknet Shopper: Criminal Acts Without Human Intent?**

In another unsettling case, a Swiss art group created the **Random Darknet Shopper**, an AI-powered bot programmed to autonomously browse darknet markets and make purchases using cryptocurrency. The bot ended up buying a range of items, including a set of counterfeit clothing, a Hungarian passport scan, and a small quantity of ecstasy. While the creators argued that the bot was part of an art installation meant to provoke dialogue around digital legality and surveillance, the police confiscated the items and opened an investigation into possible violations of narcotics and counterfeit laws.

Here, the central legal dilemma was not whether the items were illegal (they were), but **who—or what was legally responsible**. The bot had been intentionally set up to operate without human intervention, raising questions of **authorship, intent, and criminal responsibility**. Was the AI merely a tool executing pre-programmed logic? Or had the creators indirectly abetted criminal behavior by enabling the bot's unsupervised operation?

The case challenges the traditional requirement of *mens rea* and the assumption that criminal conduct must originate from a conscious, culpable mind. It also spotlights the legal grey areas in experimental or semi-autonomous AI deployments where the system is explicitly designed to act unpredictably.

Conclusion of Case Analysis

Both cases illustrate the **practical complexities** in assigning liability for AI-driven actions. They demonstrate that criminal and civil accountability cannot be framed solely in terms of traditional actors or doctrinal models. Rather, what is needed is a layered, context-sensitive framework that accounts for the technological sophistication of AI, its operational independence, and the multifaceted roles of human designers, deployers, and regulators. These examples underscore the urgent need for **clear legal guidelines and a structured policy framework** that ensures responsibility is not diffused to the point of nonexistence.

6. REGULATORY AND JURISDICTIONAL RESPONSES

As artificial intelligence becomes increasingly embedded in daily life, governments and legal institutions around the world are grappling with how to regulate its development, deployment, and potential harms. Despite growing concern over the legal accountability of AI-driven actions—particularly those that cause injury, loss, or rights violations here is a striking lack of consistent and comprehensive legal frameworks capable of assigning liability, especially for fully autonomous systems. A comparative look at regulatory efforts in India, the European Union, and the United States reveals both progress and gaps in legal preparedness.

6.1 India: An Urgent Need for a Structured Liability Framework

India, a rapidly growing hub for AI innovation and digital transformation, currently lacks a formal legal framework dedicated to AI liability. While general principles from tort law, consumer protection, and corporate liability may offer some tools for redress, they fall short in addressing the unique challenges posed by AI systems particularly those with autonomous decision-making capabilities.

Existing Indian laws, such as the Information Technology Act, 2000, are not equipped to handle questions of criminal or civil liability when the harm is caused by non-human agents like AI. Moreover, there is no legislative clarity on key concepts such as AI explainability, safety standards, human oversight obligations, or allocation of responsibility among developers, operators, and corporate owners. As a result, incidents involving AI-driven errors or misconduct risk slipping through the legal cracks, especially when blame is diffused across complex technical teams and automated systems. There have been calls from academic, industry, and policy circles for the Indian government to draft an AI policy framework that includes enforceable standards for risk assessment, algorithmic transparency, liability thresholds, and accountability mechanisms. In the absence of these, India's legal system remains reactive rather than proactive addressing harms after they occur rather than preventing them through regulation.

6.2European Union: The AI Act and a Risk-Based Regulatory Model

The European Union has taken a more structured and forward-looking approach to AI regulation. In 2021, the European Commission introduced the EU Artificial Intelligence Act, the first comprehensive attempt to regulate AI systems based on their risk profiles. The Act categorizes AI applications into four tiers unacceptable risk, high risk, limited risk, and minimal risk and assigns legal duties to developers and users based on the system's potential for harm. Under the Act, high-risk AI systems (e.g., those used in critical infrastructure, education, or law enforcement) must comply with strict requirements such as human oversight, robustness, cybersecurity, and transparency. Failure to meet these requirements can lead to penalties and restricted access to the European market.

However, while the AI Act addresses regulatory standards, it does not comprehensively resolve questions of liability, especially in the case of fully autonomous systems that make decisions independent of human input. The European Parliament has recommended exploring the notion of electronic personhood for advanced AI systems, though this remains controversial and has not been adopted as law.

6.3 United States: Fragmented and Sector-Specific Approaches

In the United States, regulatory efforts are more fragmented and largely sector-specific. The National AI Initiative Act of 2020 focuses primarily on promoting AI research and development, with limited provisions on legal responsibility or ethical safeguards. Responsibility for AI-related incidents is generally managed at the federal or state level under existing laws such as product liability, data privacy, and cybersecurity statutes.

Agencies like the Federal Trade Commission (FTC) and the Food and Drug Administration (FDA) have issued guidelines for AI systems in consumer protection and medical device contexts, respectively. Yet these guidelines are non-binding and lack the enforceability of statutory regulation. The absence of a centralized AI liability framework makes it difficult to address harms arising from cross-border or multi-stakeholder systems, leaving victims with limited legal recourse.

Conclusion of Comparative Analysis

Each jurisdiction offers valuable insights: the EU's structured risk-based approach, the U.S.'s sectoral flexibility, and India's growing recognition of the need for reform. However, none has yet fully addressed the challenges posed by fully autonomous AI systems particularly when they operate without direct human involvement and cause harm in unpredictable ways. The global nature of AI development demands not only national regulatory frameworks but also international cooperation on standards, enforcement, and liability. Without a coordinated response, AI regulation risks becoming patchwork and ineffective, enabling regulatory arbitrage and leaving both users and victims unprotected.

Going forward, lawmakers must focus on building adaptive, technology-neutral legal mechanisms that ensure accountability without stifling innovation. These mechanisms should include enforceable obligations for human oversight, algorithmic transparency, and tiered liability structures that reflect the degree of autonomy and risk associated with AI systems. Only then can the legal system begin to effectively manage the promises and perils of artificial intelligence.

7. ETHICAL AND PHILOSOPHICAL CONSIDERATIONS

The question of whether artificial intelligence should be held criminally liable—or even granted legal personhood—cannot be fully addressed without exploring its deeper ethical and philosophical dimensions. While AI systems may mimic human decision-making, their resemblance to human reasoning is functional, not moral. They lack consciousness, emotional understanding, and the ability to comprehend right from wrong. As such, they also lack moral agency, a foundational principle in attributing criminal responsibility.

Criminal justice, at its core, is built upon concepts such as blameworthiness, deterrence, rehabilitation, and retribution. These principles presuppose that the subject of punishment understands the nature and consequences of their actions. AI, no matter how advanced, does not possess intentionality, guilt, or remorse elements that underpin moral responsibility and justify punishment. To punish an AI system is, therefore, to engage in a symbolic exercise devoid of the rehabilitative or deterrent effect that punishment holds over human actors. It risks turning the justice system into a theatre of appearances, while the real agents developers, corporations, or operators remain untouched.

Philosophers such as John Searle and Daniel Dennett have long argued that even the most intelligent machines do not "understand" in the human sense; they process input based on code, not conscience. From this standpoint, holding AI criminally liable would be ethically incoherent. It would also risk undermining human responsibility by creating legal loopholes through which corporations and individuals could deflect accountability. If AI were granted legal personhood, it might be used as a scapegoat a buffer between law enforcement and the actual wrongdoers who designed, trained, or deployed the system negligently or maliciously.

There is also a danger of overextending legal personhood. While corporations are considered legal persons, they are made up of humans who benefit from and act through them. By contrast, AI lacks any community, interest, or self-perception. Granting AI the status of a legal person without any of the social or ethical grounding that justifies such status in humans or corporations could cheapen the concept and lead to unpredictable consequences.

Furthermore, ethical AI development emphasizes responsibility by design ensuring that humans remain in control, that systems are transparent, and that harmful outcomes can be traced back to a human decision or oversight failure. Shifting blame onto the AI risks disincentivizing responsible development practices, encouraging developers to offload liability onto systems they claim to no longer fully control.

From a utilitarian perspective, it is also inefficient. Creating legal pathways to "punish" AI does not prevent future harm, educate developers, or create incentives for better design. Instead, holding accountable those who make choices about how AI behaves those who set its goals, train its data, and determine when and where it is used is far more ethical and effective.

In conclusion, while the growing autonomy of AI systems has prompted discussions about extending legal personhood and liability to machines, ethical reasoning firmly supports the view that AI should remain a tool, not a bearer of blame. It is the human agents behind designers, programmers, corporations, and regulators who must be held responsible. Doing so preserves the moral integrity of the legal system while ensuring that accountability remains where it belongs: with those capable of choice, intention, and ethical reasoning.

8. CONCLUSION AND SUGGESTION

The rise of artificial intelligence presents one of the most profound challenges to the modern legal system. As AI systems become increasingly autonomous, unpredictable, and embedded in critical infrastructure, the traditional doctrines of criminal liability built on human intention, foreseeability, and agency are showing their limits. The cases, theories, and comparative regulations discussed in this paper reveal that while AI can perform functions once reserved for human decision-makers, it does so without consciousness, moral judgment, or the capacity for legal accountability in any meaningful sense.

Given these limitations, granting legal personhood to AI at this stage would be premature and potentially harmful. Such a move risks diluting human responsibility, creating loopholes for powerful corporations to deflect blame, and undermining the ethical foundation of criminal law. Instead, the law must evolve to clarify and reinforce the responsibilities of those who create, deploy, and profit from AI technologies.

To manage AI's legal implications effectively, the following policy recommendations are essential:

- **Mandatory Audit Trails and Explainability Requirements:** AI systems, especially those used in high-risk applications like healthcare, transportation, and public safety, must be built with mechanisms that allow for post-incident analysis. These audit trails should document how decisions were made and which data inputs were used. Transparency and explainability are not just technical features they are legal imperatives that enable courts and regulators to trace accountability when something goes wrong.
- **Defined Roles and Oversight Duties:** Legislatures should mandate clear delineation of responsibilities among developers, deployers, owners, and users of AI systems. This includes legally binding duties to monitor, test, and update AI systems regularly, as well as specific requirements for human oversight in decision-making processes that carry legal or ethical weight. Ambiguity in responsibility often leads to gaps in liability clarity is key to preventing that.

- **Civil Liability Extensions with Criminal Negligence Thresholds**: While AI may not possess intent or awareness, the people and entities behind it do. Legal reforms should introduce hybrid models that extend civil liability into the criminal domain when negligence in design, deployment, or oversight leads to serious harm. For example, a developer who knowingly fails to address foreseeable risks in an AI's design could be subject to criminal penalties under a "gross negligence" standard.

Ultimately, human accountability must remain at the heart of AI regulation. Until and unless AI systems evolve to possess consistent, interpretable, and demonstrable moral agency a hypothetical that remains far from reality blame, responsibility, and legal liability must continue to rest with those who act through, control, or benefit from these systems. By adapting the law thoughtfully, and reinforcing ethical and regulatory guardrails, we can ensure that AI serves society without undermining justice, safety, or human dignity.

References

1. Priyanka Majumdar, Dr. Bindu Ronald & Dr. Rupal Rautdesai, Artificial Intelligence, Legal Personhood and Determination of Criminal Liability, 6 J. CRITICAL REVIEWS 323 (2019).
2. Alaa Saud, Criminal Liability About the Use of Artificial Intelligence, 1 INT'L J.L. MGMT. & HUM. (2020), https://www.ijlmh.com/wp-content/uploads/Criminal-Liability-about-the-Use-of-Artificial-Intelligence.pdf.
3. Gabriel Hallevy, The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control, 4 AKRON INTELL. PROP. J. 171 (2016).
4. John McCarthy, What Is Artificial Intelligence?, STAN. U. (2007), http://jmc.stanford.edu/articles/whatisai/whatisai.pdf.
5. NITI Aayog, National Strategy for Artificial Intelligence, GOV'T OF INDIA (2018), https://niti.gov.in/sites/default/files/2022-11/NationalStrategy-for-AI.pdf.
6. Lawrence B. Solum, Legal Personhood for Artificial Intelligences, 70 N.C. L. REV. 1231 (1992).
7. United States v. Andrews, 681 F. Supp. 2d 591 (E.D. Va. 2010).
8. State v. Kaiser, 672 N.W.2d 122 (Minn. Ct. App. 2003).
9. Visa Kurki & Tomasz Pietrzykowski, Legal Personhood: Animals, Artificial Intelligence and the Unborn, 25 ETHICAL THEORY & MORAL PRAC. 215 (2017).
10. Hans kelsen, general theory of law and state (1945).
11. Nicola lacey & celia wells, reconstructing criminal law 53 (2d ed. 1998).
12. ISAAC ASIMOV, I, ROBOT 40 (1950).
13. Daisuke Wakabayashi, Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam, N.Y. TIMES (Mar. 19, 2018), https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.
14. Olivia Solon, Why Uber's Self-Driving Car Killed a Pedestrian, THE GUARDIAN (Mar. 22, 2018), https://www.theguardian.com/technology/2018/mar/21/uber-self-driving-car-death-arizona-tempe.
15. Bryan Menegus, Random Darknet Shopper Bought Ecstasy, But Its Creators Won't Be Prosecuted, GIZMODO (Jan. 13, 2015), https://gizmodo.com/random-darknet-shopper-bought-ecstasy-but-its-creators-1680003775.