# ARTIFICIAL INTELLIGENCE AND CRIMINAL LIABILITY: RETHINKING MENS REA IN THE AGE OF AUTONOMOUS SYSTEMS

Advocate Ritika Sharma, LL.M., Fairfield Institute of Management and Technology, Guru Gobind Singh Indraprastha University (GGSIPU)

**ABSTRACT**

As Artificial Intelligence (AI) continues its exponential evolution, it challenges traditional legal doctrines built around human agency, culpability, and intent. The cornerstone of criminal liability mens rea requires a conscious, culpable mental state, but AI systems operate through sophisticated algorithms, neural networks, and deep learning models, entirely devoid of human consciousness or subjective intent. This article explores the conceptual and legal complexities surrounding the attribution of criminal liability for harms caused by highly autonomous systems. It critically examines why existing legal frameworks, particularly in the Indian context, are insufficient and explores proposed models, such as Synthetic Mens Rea and Vicarious Liability. Drawing from comparative international perspectives, this study seeks to provide a balanced understanding of how criminal jurisprudence must adapt in the digital age by advocating for a hybrid framework that bridges the growing accountability gap between human creators and autonomous agents.

**Keywords:** Artificial Intelligence, Criminal Liability, Mens Rea, Autonomous Systems, Legal Reform, Legal Personhood

## I. INTRODUCTION

Technological advancement has always served as a catalyst for the evolution of law, consistently reshaping how societies define rights, duties, and accountability. From the industrial revolution to the digital era, each wave of innovation has compelled legal systems to adapt to reconsider old doctrines and to accommodate new realities. However, the contemporary rise of Artificial Intelligence (AI) marks a turning point far more profound than any prior technological transformation. AI no longer functions merely as a tool extending human capability; rather, it is emerging as an independent actor capable of learning, reasoning, and making decisions with minimal or even no direct human oversight. This qualitative shift challenges the very foundations of criminal jurisprudence, which has, for centuries, been deeply anthropocentric built upon assumptions about human behaviour, intention, and moral culpability.

Unlike traditional tools that operate strictly within the boundaries of human control, AI systems possess the capacity to act unpredictably, evolving through self-learning mechanisms such as machine learning and neural networks. They can now perform complex, high-stakes functions from medical diagnosis and autonomous driving to predictive policing and automated financial trading often without any real-time human command. When these systems malfunction or cause harm, the legal question becomes both urgent and deeply philosophical: who should bear the blame? Should it be the programmer, the user, the manufacturer, or the AI itself? Such questions are not merely theoretical curiosities; they directly impact notions of justice, fairness, and accountability within criminal law.

The challenge is intensified by the fact that criminal responsibility traditionally hinges on the doctrine of *mens rea* the "guilty mind." The principle assumes that culpability arises from a conscious, intentional, or at least reckless human state of mind accompanying the wrongful act (*actus reus*). Yet, AI, by its very nature, lacks consciousness, emotion, or subjective intent. It operates through algorithms and data patterns, not through moral reasoning or awareness of wrongdoing. Consequently, applying traditional concepts of *mens rea* to autonomous systems appears increasingly inadequate. The legal system, therefore, faces a conceptual dilemma: how can it assign liability for harm caused by an entity that does not possess a "mind" in the human sense, but nonetheless acts in ways that mirror or even surpass human decision-making?

This article argues that the existing human-centric doctrines of criminal law are rapidly

becoming obsolete in the face of rising machine autonomy. To preserve the integrity of justice in the age of intelligent systems, criminal jurisprudence must undergo a profound re-examination — one that moves beyond anthropomorphic assumptions and rethinks fundamental notions of intent, control, and culpability in the context of artificial agency.

## II. BACKGROUND: THE FOUNDATION OF MENS REA

Mens rea, or the "guilty mind," together with actus reus, the "guilty act," forms the twin foundation of criminal liability in most common law jurisdictions, including India. These two principles operate in tandem to ensure that punishment is not merely for causing harm but for doing so with a culpable state of mind. In essence, criminal law has always sought to distinguish between the morally blameworthy and the morally blameless by emphasizing the mental element behind every prohibited act. The doctrine of mens rea reflects society's belief that true guilt lies not only in the act itself but in the intention, awareness, or recklessness that accompanies it.

Mens rea is, therefore, the moral and psychological component of crime a measure of an individual's internal state at the time of committing an offense. It encompasses varying degrees of culpability, ranging from specific intent, as in cases of premeditated murder, to recklessness and negligence, where harm results from a disregard of foreseeable risks. This doctrine is deeply rooted in the maxim *actus non facit reum nisi mens sit rea* "the act does not make one guilty unless the mind is also guilty." The maxim ensures that criminal responsibility cannot be imposed merely on the basis of outcomes or consequences; it must be grounded in the actor's conscious or negligent mental state. Thus, under traditional jurisprudence, intent, knowledge, foresight, and control are indispensable for establishing guilt.

However, the emergence of Artificial Intelligence (AI) fundamentally disrupts this equilibrium. AI systems whether in the form of autonomous vehicles, high-frequency trading bots, predictive policing algorithms, or medical diagnostic system operate on principles that are devoid of consciousness, moral understanding, or intentionality. These systems process vast amounts of data, identify patterns, and make decisions based on algorithms and probabilistic models, not moral reasoning or awareness. Consequently, when such systems act autonomously and cause harm for example, when a self-driving car causes a fatal accident or an algorithm unfairly discriminates in a criminal sentencing decision the traditional framework of mens rea becomes conceptually inadequate.

The difficulty lies in the displacement of human agency. While human beings design, program, and deploy AI systems, there often exists a significant causal and temporal gap between human action (programming or input) and the AI's eventual output (the harmful act). As AI systems continuously learn and adapt beyond their initial programming, this gap widens, creating what legal scholars call the "responsibility vacuum." In such situations, attributing blame becomes profoundly complex. The programmer may not have foreseen the specific harm, the operator may not have had control, and the AI itself lacks a mind capable of forming intent or malice. Thus, the traditional mens rea construct built upon the presumption of human consciousness fails to map neatly onto autonomous, algorithm-driven entities.

This growing disconnect between human intention and machine behaviour forces criminal law to confront a fundamental dilemma: can liability exist without intent, or must our legal frameworks evolve to recognize new forms of culpability applicable to artificial agents? As AI continues to assume greater autonomy in decision-making, the long-standing human-centric conception of mens rea may need to be reinterpreted, expanded, or even replaced to ensure that justice and accountability remain meaningful in an era where non-human entities increasingly act with real-world consequences.

## III. THE LIABILITY GAP: AI'S CHALLENGE TO CULPABILITY

Recent incidents involving algorithmic bias, automated trading fraud, and fatal accidents caused by autonomous vehicles clearly illustrate the potential of AI systems to cause criminal harm. These scenarios expose a critical *liability gap* or *accountability gap* in current law.

### A. The Problem of Causation

The doctrine of causation lies at the heart of criminal liability, serving as the critical bridge between an accused's conduct and the resulting harm. It ensures that criminal responsibility is not attributed arbitrarily but only where the defendant's act or omission can be shown to have caused the prohibited consequence in a legally relevant manner. Traditionally, causation in criminal law involves two interrelated components: factual causation and legal (proximate) causation. Factual causation is often established through the *"but-for" test* that is, whether the harm would not have occurred *but for* the defendant's conduct. Legal causation, on the other hand, considers whether the outcome was a reasonably foreseeable consequence of that conduct. This framework functions effectively in human-driven scenarios, where intent,

knowledge, and foreseeability can be attributed to identifiable individuals. However, in the context of Artificial Intelligence (AI), this clear chain of reasoning becomes profoundly complicated.

AI systems, particularly those that employ machine learning (ML) and deep learning (DL) architectures, operate in ways that often transcend their initial programming. These systems continuously learn, evolve, and modify their decision-making parameters by processing massive volumes of real-world data. This dynamic and self-updating capability, while central to the power of AI, also introduces what is commonly referred to as the "black box" problem a condition in which the internal reasoning or decision-making process of an AI becomes opaque, even to its own developers. When an AI system acts in an unexpected or harmful manner, the causal chain linking human action (e.g., coding, training, or deployment) to machine output (the harmful act) becomes fragmented and uncertain.

For instance, consider a self-driving vehicle that causes a fatal accident after misinterpreting road conditions or sensor data. The immediate cause of harm lies in the vehicle's autonomous decision-making system an algorithmic process far removed from the programmer's direct input. Similarly, in the context of predictive policing or automated sentencing algorithms, when biased outcomes emerge from machine learning models trained on historical data, determining who or what is causally responsible becomes almost impossible. Did the fault lie with the data provider, the programmer, the user, or the system's own evolving logic? The answer is neither clear nor easily provable within existing legal doctrines.

This causal indeterminacy strikes at the core of criminal jurisprudence. The "but-for" test, which demands a direct and foreseeable connection between human conduct and criminal outcome, is often undermined by the autonomous and adaptive nature of AI systems. Moreover, even if factual causation could be stretched to include the original human act of programming or deployment, the foreseeability requirement of legal causation falters when the system's subsequent behavior diverges from any predictable trajectory. The result is a profound accountability gap a legal vacuum where harm is real but blame cannot be convincingly assigned under traditional causation tests.

Thus, the problem of causation in AI-driven acts is not merely procedural; it is conceptual and structural. It exposes the limitations of criminal law's reliance on linear, human-centered models of cause and effect. To maintain coherence and fairness in criminal adjudication, legal

systems must grapple with whether existing causation doctrines can be reinterpreted or redefined to accommodate the complex, non-linear, and probabilistic causality characteristic of autonomous technologies. Without such evolution, the law risks becoming increasingly disconnected from the realities of technological agency in the modern world.

**B. The Absence of Subjective Intent**

Perhaps the most profound and conceptually challenging issue in assigning criminal liability to Artificial Intelligence (AI) lies in the imputation of a mental state, or *mens rea.* Mens rea, which literally translates to the "guilty mind," functions as the moral and psychological foundation of criminal culpability. It is not merely a procedural requirement but an ethical cornerstone that differentiates accidental harm from culpable wrongdoing. Through mens rea, the law assesses the degree of moral blameworthiness attached to a prohibited act, ensuring that punishment corresponds to the actor's mental state be it intention, knowledge, recklessness, or negligence. Without this subjective component, criminal law risks degenerating into a regime of strict liability, punishing outcomes without regard to moral fault.

However, imputing such a mental state to autonomous systems presents an unprecedented doctrinal challenge. Unlike humans, AI systems do not possess consciousness, emotions, or subjective awareness. They operate according to mathematical models, probabilistic algorithms, and programmed utility functions, rather than motives, desires, or moral reasoning. Therefore, the traditional hierarchy of mental states intent, knowledge, recklessness, and negligence becomes exceedingly difficult, if not impossible, to apply within the AI context.

At the highest end of culpability, intention and knowledge imply deliberate awareness and volition. A human being who intentionally kills or defrauds does so with cognitive understanding of the nature and consequences of the act, as well as its legal and moral implications. In contrast, even the most sophisticated AI cannot "intend" to commit murder or "know" that its actions are unlawful in any human sense. Its decision-making processes are instrumental rather than moral guided by optimization algorithms, utility-based reasoning, and data-driven pattern recognition. For example, a trading algorithm that manipulates markets to maximize profit does not "intend" to commit fraud; it merely executes the most efficient strategy within its coded parameters. The absence of subjective cognition makes it conceptually incoherent to ascribe true intent or knowledge to such systems.

At the lower end of the culpability spectrum, the notions of recklessness and negligence pose equally thorny questions. Human recklessness involves a conscious disregard of an unreasonable risk, reflecting both awareness and choice. Negligence, while less subjective, still assumes that the actor could have reasonably foreseen and avoided the harmful consequence. An AI system, however, lacks such foresight in the human sense. While one might describe an AI as "negligent" when it fails to apply an optimal decision rule or misclassifies critical data, this is merely an objective characterization — a metaphorical borrowing of human legal terminology. The machine's "failure" does not stem from a mental lapse or disregard of duty but from an algorithmic limitation, data bias, or unforeseen computational error. Consequently, imputing human-like culpability to AI under existing definitions of recklessness or negligence distorts the moral and psychological foundations of these concepts.

This mens rea dilemma underscores the anthropocentric limits of current criminal law frameworks. As AI systems continue to act autonomously and influence real-world outcomes, the law must confront whether it is conceptually sustainable to continue applying human mental state doctrines to non-human agents. Scholars increasingly argue for the development of alternative liability models such as *constructive, proxy, or vicarious intent* or even for the recognition of "electronic personhood" in certain contexts. Without such evolution, the doctrine of mens rea risks becoming functionally obsolete in addressing crimes emerging from autonomous technological behaviour.

## IV. LEGAL ANALYSIS: MODELS FOR ATTRIBUTION

In the absence of a legally recognized personhood for Artificial Intelligence (AI), legal scholars and jurists have sought alternative frameworks to address the pressing issue of accountability and attribution of liability. The fundamental difficulty lies in reconciling the traditional requirement of a guilty mind (*mens rea*) with the non-human, autonomous, and non-conscious nature of AI systems. To bridge this gap, a range of theoretical models have been proposed each attempting to allocate responsibility among the human actors involved in the design, deployment, and supervision of AI. Among these, one of the most prominent and widely discussed approaches is the Producer or Programmer Liability model, often described as the "Original Sin" Theory.

### A. Producer or Programmer Liability (The 'Original Sin' Theory)

The Producer or Programmer Liability model is premised on the notion that any criminal act committed by an AI system can ultimately be traced back to human agency specifically, to the designers, manufacturers, or programmers who created and trained the system. Under this theory, the AI's seemingly autonomous behaviour is viewed as the extended manifestation of the programmer's original code, intention, or error. The "Original Sin" metaphor reflects the belief that the "fault" of the machine lies not in its independent conduct but in the foundational act of programming, where biases, vulnerabilities, or risky operational parameters were first introduced. In essence, the AI's harmful act (*actus reus*) is interpreted as a derivative consequence of the human actor's initial moral and technical choices.

Proponents of this model argue that programmers occupy a unique position of control and foreseeability, as they are responsible for determining the architecture, decision-making protocols, and boundaries within which AI operates. Accordingly, when a machine performs an unlawful act for instance, when a self-driving car violates traffic norms leading to a fatality, or an algorithmic trading bot manipulates financial markets the liability should rest with those who conceived or engineered the system. From this perspective, criminal culpability is not displaced onto the AI but remains rooted in human intent and human error, maintaining the anthropocentric foundation of criminal jurisprudence.

However, this model faces significant doctrinal and practical limitations. The central weakness of the "Original Sin" theory emerges when an AI system evolves or learns beyond the scope of its original programming. Modern machine learning systems continuously adapt based on data inputs, environmental feedback, and probabilistic inference processes that can lead to outcomes completely unforeseen by their creators. In such cases, attributing liability to the programmer becomes too remote and attenuated to satisfy the legal requirements of both *causation* and *mens rea.* The programmer's original intent may have been lawful, and yet the AI's self-generated decision may culminate in unlawful harm. The traditional "but-for" and foreseeability tests thus collapse under the weight of algorithmic autonomy.

Moreover, imposing criminal liability on programmers or developers for unforeseeable, emergent AI behaviour carries serious policy implications. It risks stifling innovation by deterring researchers and companies from advancing AI technologies for fear of disproportionate liability. Such a punitive approach could discourage experimentation, slow

technological progress, and undermine the very social and economic benefits that AI promises. Legal scholars also caution that expanding the scope of programmer liability could blur the line between negligence and strict liability, resulting in unfair punishment for individuals who neither intended nor could reasonably predict the machine's harmful conduct.

Consequently, while the Producer or Programmer Liability model provides a convenient means to anchor accountability within human agency, it fails to adequately address the autonomy and unpredictability of modern AI systems. Its reliance on backward attribution to human creators is conceptually fragile in an era where machines can self-modify, evolve, and act in ways that even their designers cannot fully comprehend. The challenge, therefore, is to devise new hybrid or adaptive frameworks that balance accountability with fairness ensuring that responsibility is neither unfairly imposed on innocent human agents nor allowed to dissipate in the technological void.

**B. Operator or User Liability (The 'Temptation' Theory)**

The Operator or User Liability model, often described as the 'Temptation' Theory, shifts the focus of criminal accountability from the producer or programmer to the individual who deploys or operates the AI system in the real world. The rationale is straightforward: even if the machine acts autonomously, the human operator remains the ultimate decision-maker in activating, supervising, or relying upon its functions. Accordingly, liability is premised on the operator's duty of care and their potential criminal negligence in the use or oversight of the AI system.

This model finds support in existing legal doctrines that hold users accountable for their instruments or tools, much like how the law attributes responsibility to a driver for the actions of a vehicle under their control. For instance, if an individual operates a semi-autonomous vehicle and fails to take corrective control when the system malfunctions, resulting in injury or death, the human operator may be held liable for gross negligence or criminal omission. Similarly, a company employee who deploys an AI-enabled surveillance or trading system without adequate safeguards could be liable if the system's actions violate privacy norms, financial regulations, or criminal statutes.

The 'Temptation' aspect of this theory arises from the human tendency to over-rely on or abdicate responsibility to intelligent systems, especially when these systems appear highly

accurate or self-sufficient. The law, therefore, seeks to discourage "automation bias" — the blind trust placed in machine recommendations — by reinforcing that ultimate accountability rests with the human operator. Under this framework, criminal negligence may be established if it is proven that the user failed to exercise reasonable supervision, failed to intervene when warning signs were evident, or activated the AI in an inherently risky context.

However, this model's strength diminishes significantly in the context of fully autonomous systems — those that function in closed-loop environments with minimal or no real-time human oversight. In such cases, the operator may neither have the capacity nor the technical means to intervene once the AI has been deployed. For example, a fully autonomous drone operating under machine learning protocols may alter its flight path and cause collateral damage without any direct human command. Here, assigning criminal responsibility to the operator would be unjust and conceptually strained, as the human's ability to foresee or control the outcome is minimal.

Furthermore, the complexity and opacity of advanced AI systems often prevent users from understanding the inner workings or decision-making logic of the technology they employ. This undermines the traditional criminal law requirement that liability be grounded in foreseeability and voluntary control. While the Operator Liability model remains viable for semi-autonomous systems such as assisted-driving vehicles, diagnostic tools, or AI decision-support mechanisms it becomes increasingly untenable as AI evolves toward self-learning, unsupervised, and adaptive models.

In essence, the "Temptation" theory underscores the need for context-sensitive regulation that distinguishes between degrees of human involvement. Blanket attribution of liability to users may serve as a deterrent, but in cases of genuine machine autonomy, it risks collapsing into an unfair presumption of guilt without meaningful fault.

### C. Corporate Criminal Liability

In contrast to individual attribution models, many legal systems particularly in the United States, the United Kingdom, and the European Union have adopted the approach of addressing AI-related harm through Corporate Criminal Liability doctrines. This model recognizes that AI systems are typically developed, owned, and deployed by corporations, and that these entities should bear responsibility for harms resulting from the actions of their automated agents.

Corporate criminal liability operates through established doctrines such as the doctrine of identification and vicarious liability. Under the doctrine of identification, a corporation is held liable for criminal acts committed by its senior officials or controlling minds. Under vicarious liability, the corporation can be held responsible for acts committed by employees or agents within the course of their employment, even if the corporation itself lacks intent. When applied to AI, the system is viewed as a functional "agent" of the corporation, performing tasks that serve the company's economic or operational interests.

For example, if an AI-driven financial algorithm manipulates markets or engages in fraudulent transactions, liability may attach to the corporation that designed, implemented, or benefited from the system's conduct. Similarly, if an autonomous vehicle developed by a car manufacturer causes death due to design flaws in its AI system, the corporation can be held criminally liable under doctrines akin to product liability or corporate negligence.

This model is particularly attractive from a policy standpoint, as it allows for the imposition of financial penalties, regulatory sanctions, and compliance obligations, without the need to prove the presence of individual mens rea. It ensures that victims receive compensation and that corporations are incentivized to maintain ethical and safe AI practices. Moreover, corporations are better positioned than individuals to absorb financial losses and to implement systemic safeguards to prevent future harm.

Nonetheless, the corporate model remains an imperfect solution. It primarily serves an instrumental rather than moral function. While it ensures regulatory control and deterrence, it sidesteps the core moral inquiry of criminal law namely, the evaluation of personal guilt and culpability. Corporations are legal fictions, and AI systems are non-sentient entities; thus, the imposition of moral blame in such cases is largely symbolic. Furthermore, the diffusion of responsibility across corporate hierarchies can lead to accountability dilution, where no individual within the corporation is directly answerable for the harm.

Therefore, while corporate criminal liability offers a pragmatic mechanism for managing AI-related risks, it fails to confront the philosophical challenge posed by autonomous decision-making systems the erosion of the traditional nexus between human intent, control, and culpability.

### D. Synthetic Mens Rea

A more radical and intellectually provocative proposal emerging from modern jurisprudence is the concept of "Synthetic Mens Rea." This model seeks to reimagine culpability not as a reflection of human consciousness but as a functional and objective assessment of the AI system's behaviour, structure, and risk potential. Under this theory, liability would be grounded in quantifiable data, such as the system's design logic, algorithmic architecture, data sources, and operational outcomes, rather than on the subjective mental state traditionally required for human offenders.

Proponents of this approach argue that AI's "mental state" can be reconstructed synthetically by analysing its decision pathways i.e., whether the AI's conduct was the result of foreseeable programming parameters or whether it demonstrated patterns of "reckless" disregard for embedded safety norms. In effect, the AI's culpability is derived from its predictive and operational design, enabling the law to assign a form of constructive intent based on objective functionality rather than consciousness. For example, if an AI system repeatedly exhibits harmful tendencies that its programmer fails to correct, the system could be said to possess a "synthetic" intent to cause harm, imputing liability to its creator or owner.

While theoretically innovative, the concept of Synthetic Mens Rea raises significant philosophical and doctrinal challenges. By constructing a legal fiction of intent for non-sentient entities, it risks undermining the moral foundation of criminal law, which rests on the notion of conscious choice and moral agency. Intent, as traditionally understood, presupposes awareness, volition, and the capacity for moral reasoning qualities that machines do not and, arguably, cannot possess. To ascribe mens rea to a machine is to anthropomorphize technology, treating algorithmic behaviour as moral decision-making.

Furthermore, such an approach may lead to practical inconsistencies. If AI systems are deemed capable of possessing synthetic mens rea, they could, by extension, be viewed as legal persons, thereby necessitating rights, defences, and procedural safeguards akin to those enjoyed by humans a concept that legal systems are ill-equipped to operationalize. Critics therefore contend that while Synthetic Mens Rea provides a useful analytical lens, it should function as a tool for assessing human negligence or institutional fault, rather than as a substitute for genuine intention.

In sum, the synthetic Mens Rea framework reflects the growing recognition that traditional concepts of culpability must evolve to meet the realities of intelligent automation. However, it also highlights the inherent tension between legal pragmatism and moral coherence — between adapting the law to new technologies and preserving the ethical foundations upon which criminal justice is built.

## V. COMPARATIVE AND REFORM PERSPECTIVES

The rapid advancement of Artificial Intelligence (AI) and its increasing role in decision-making have compelled legal systems around the world to confront the complex issue of accountability and criminal liability for autonomous systems. Different jurisdictions have begun exploring distinct frameworks to address these challenges, reflecting variations in legal philosophy, institutional readiness, and public policy priorities. While some regions, notably within the European Union, have taken bold steps toward recognizing a form of legal status for autonomous entities, others, such as India, continue to rely on traditional human-centric criminal law doctrines that presuppose consciousness, intent, and volition. A comparative analysis of these approaches offers valuable insight into possible reform trajectories and their implications for rethinking *mens rea* in the age of autonomous systems.

### A. The European Proposal: 'Electronic Personhood'

The European Parliament has been one of the most proactive bodies globally in grappling with the legal implications of Artificial Intelligence. In a 2017 resolution on civil law rules for robotics, the European Parliament proposed the idea of granting "electronic personhood" to certain highly autonomous and self-learning AI systems. This concept is not equivalent to full human legal personhood but instead envisions a limited, functional legal status that would allow such systems to bear rights and obligations within a restricted legal framework. The primary motivation behind this proposal was to fill the accountability gap that arises when harm is caused by autonomous AI for which no natural or corporate person can be directly held responsible.

Under this proposal, highly autonomous AI systems those capable of operating independently and learning from their environments would be recognized as electronic persons for the purposes of civil liability, primarily in the domains of tort and contract law. The key objective is to ensure that victims of AI-related harm have an identifiable entity against which to claim

compensation, even when the causal chain to human fault is broken. For example, if a self-learning robotic system causes physical injury or property damage due to an emergent behaviour beyond its programming, the system itself could, in theory, be held liable as an electronic person with a dedicated insurance or compensation fund.

However, while the European Parliament's proposal represents a progressive step toward addressing the realities of AI autonomy, its extension into criminal law remains deeply controversial and widely resisted across jurisdictions. The fundamental reason lies in the moral and philosophical underpinnings of criminal responsibility. Recognizing an AI system as a criminal "person" would necessarily imply that it possesses a degree of moral awareness, intentionality, and the capacity for guilt or reform attributes that current AI systems fundamentally lack. Criminal liability, unlike civil liability, is premised on moral blameworthiness, deterrence, and retribution principles that lose their meaning when applied to a non-sentient, purely computational entity.

Furthermore, granting AI legal personhood in the criminal context risks creating unintended normative distortions. If AI systems were treated as criminal persons, it could lead to the absolution of human or corporate actors who design, deploy, or profit from them, shifting responsibility away from those capable of moral reasoning. The fear is that AI personhood may become a convenient "liability shield" allowing corporations to deflect accountability by attributing fault to a non-human entity. Thus, the consensus among European legal scholars remains that accountability must trace back to human or corporate actors, ensuring that responsibility continues to rest within the domain of human agency and control.

Nonetheless, the European discourse on electronic personhood has had a significant impact on global legal thought. It highlights a growing recognition that traditional doctrines of causation, agency, and intention are increasingly inadequate for regulating autonomous systems. While criminal law may not yet be ready to embrace electronic personhood, the European initiative provides an important template for hybrid liability frameworks, where civil, administrative, and quasi-criminal mechanisms coexist to fill emerging gaps in accountability. In this sense, Europe's approach serves as a laboratory of ideas, testing the limits of legal imagination in the age of artificial intelligence.

**B. Indian Jurisprudence: The Indian Penal Code, 1860**

In contrast to Europe's experimental stance, Indian jurisprudence remains firmly grounded in human-centric principles of criminal liability. The primary source of substantive criminal law in India, the Indian Penal Code (IPC), 1860, was drafted in an era when the concept of machine autonomy was inconceivable. Unsurprisingly, its structure and terminology are premised entirely on the assumption that offenders are human beings or legal entities composed of humans. Section 11 of the IPC defines a "person" to include "any Company or Association or Body of Persons, whether incorporated or not." This inclusion allows for corporate criminal liability, thereby extending the scope of criminal responsibility beyond natural persons. However, this statutory definition does not anticipate or encompass non-human, non-corporate autonomous systems such as AI agents or robots.

Indian criminal law rests on intent-based culpability, with mental state requirements embedded across various provisions of the IPC. Terms such as *"voluntarily," "dishonestly," "fraudulently," "intentionally,"* and *"knowingly"* are central to determining the degree of criminal guilt. Each of these expressions presupposes conscious awareness and volition, attributes that machines inherently lack. For instance, Section 39 defines "voluntarily" as causing an effect "by means whereby he intended to cause it." Similarly, Sections 24 and 25 define *"dishonestly"* and *"fraudulently"* in reference to the intention to cause wrongful gain or loss. The doctrinal structure of the IPC thus ties criminal liability inseparably to subjective human mental states.

This framework creates an immediate and substantial legal vacuum in addressing AI-related harms. Suppose an autonomous vehicle operating under a self-learning algorithm causes a fatal accident due to an unforeseeable system error. The current criminal law apparatus would struggle to identify an offender capable of forming *mens rea*. The programmer's role may be too remote to satisfy causation, and the operator may lack real-time control. Moreover, the AI itself cannot be prosecuted, as it does not fall within the statutory definition of a "person" under Section 11. Consequently, Indian criminal law, in its present form, is structurally ill-equipped to assign culpability in such scenarios.

To address this emerging gap, Indian jurisprudence will require significant legislative and conceptual reform. One possible approach is to introduce specific statutory amendments that explicitly recognize AI-induced harms and establish graded liability frameworks for designers, operators, and corporations. Such amendments could clarify the standards for foreseeability,

control, and risk assessment in AI contexts. Another possibility is the creation of a specialized statute or code, akin to data protection or cybercrime laws, that delineates offences involving autonomous systems. This could include provisions for "AI negligence," "algorithmic recklessness," or failure to supervise autonomous agents, thereby bridging the gap between human accountability and machine autonomy.

Additionally, India could explore administrative or quasi-criminal models, where accountability is enforced through regulatory agencies rather than traditional criminal courts. This would align with global trends emphasizing preventive regulation, transparency obligations, and ethical compliance standards for AI development. In this way, the Indian legal system can uphold the spirit of criminal justice ensuring accountability, deterrence, and fairness without stretching traditional doctrines beyond their conceptual limits.

In conclusion, while European legal thought is experimenting with electronic personhood as a mechanism to address the accountability gap, Indian jurisprudence remains anchored in classical notions of human intent and corporate responsibility. Both approaches reveal the tension between technological innovation and legal adaptation. For India, the way forward lies not in granting AI moral or legal personhood but in restructuring existing legal principles to reflect the new realities of technological agency, ensuring that law remains both responsive and just in the age of intelligent machines.

## VI. ETHICAL AND POLICY DILEMMAS

The attribution of criminal liability to Artificial Intelligence (AI) raises some of the most complex and unsettled ethical and policy dilemmas in modern jurisprudence. The fundamental challenge lies in reconciling technological autonomy with moral accountability that is, how to hold someone or something responsible for harm caused by a machine that can make decisions independent of direct human command. The dilemma operates at two extremes: on one side lies the danger of overextending liability to AI itself, thereby diluting human moral responsibility; on the other, the risk of ignoring the autonomous nature of AI systems, thereby allowing harmful acts to go unpunished and victims uncompensated. Striking an equilibrium between these poles is not merely a legal exercise but an ethical and policy imperative that will shape the future relationship between law, technology, and human society.

At the heart of the ethical debate is the question of moral agency. Criminal liability has

historically rested upon the presumption that moral beings endowed with free will, rationality, and awareness — are capable of distinguishing right from wrong and making conscious choices. AI systems, however, lack all such faculties. They operate on the basis of mathematical logic, probabilistic reasoning, and algorithmic optimization. They do not possess consciousness, intention, empathy, or moral understanding. Consequently, assigning criminal culpability to an AI system raises profound philosophical concerns. It risks creating a legal fiction of guilt devoid of genuine moral content, thereby undermining the foundational purpose of criminal law: to ascribe blame only where there is conscious wrongdoing.

Conversely, ignoring AI's growing functional autonomy risks enabling a new kind of impunity. Modern AI systems are capable of making complex decisions in areas such as autonomous driving, financial trading, law enforcement, medical diagnostics, and military applications. Their actions can produce real-world harm physical injury, economic loss, reputational damage, or even death. If no clear framework exists to allocate responsibility for such outcomes, victims may be left without recourse, and harmful behaviour could proliferate unchecked. The law thus faces a dual moral hazard: either absolving humans by shifting blame to machines, or permitting harm to go unpunished because the law cannot conceptualize non-human accountability.

A key ethical concern emerging from this tension is the potential dilution of human accountability. If AI systems were granted any form of legal or quasi-legal personhood, there exists a genuine risk that individuals and corporations could use this status to deflect responsibility for unlawful acts. A company might argue that an AI algorithm made an "independent" decision to engage in price manipulation or discrimination, thereby insulating its human managers or developers from blame. This could lead to the fragmentation of moral responsibility, where culpability is diffused across technical systems, data inputs, and algorithmic processes making it increasingly difficult to identify a responsible human actor. In such a scenario, the moral purpose of criminal law to hold human beings accountable for their choices would be seriously undermined.

On the other hand, treating AI systems as mere instruments under complete human control is becoming increasingly unrealistic. As AI continues to evolve, its capacity for self-learning and adaptive behaviour means that its decisions cannot always be anticipated or explained by its creators. This "black box" problem where even the designers cannot fully understand how the

AI arrived at a particular decision challenges the traditional assumption that humans retain total control. In criminal contexts, this opacity poses severe evidentiary and ethical problems: if the causal chain and reasoning process are indeterminate, can we fairly punish a human actor who neither foresaw nor could have prevented the harm?

Policymakers, therefore, must navigate these conflicting ethical imperatives with caution and foresight. The overarching policy goal should be to maintain human accountability while promoting responsible technological innovation. This requires a multi-layered approach that integrates legal reform, ethical governance, and technical regulation. Instead of focusing solely on post hoc criminal punishment, emphasis must shift toward ex ante accountability ensuring that AI systems are designed, deployed, and monitored in ways that minimize risk and preserve traceability.

A central tenet of this approach is the development of auditable, transparent, and explainable AI systems. The concept of "explainable AI" (XAI) has gained prominence as both a technical and ethical standard. It requires that AI decision-making processes be comprehensible, reconstruct able, and subject to human review. Transparency mechanisms such as algorithmic documentation, bias testing, and traceability protocols enable regulators and courts to determine how and why a particular outcome occurred. This, in turn, ensures that responsibility can be traced back to identifiable human agents whether they are designers, data providers, deployers, or corporate managers. In the absence of such explainability, assigning mens rea or even establishing causation becomes practically impossible.

Ethically, this aligns with the principle of "human-in-the-loop accountability." Under this model, every autonomous or semi-autonomous system must be embedded within a governance structure that guarantees a human actor's ability to supervise, intervene, and assume responsibility for the system's actions. In practice, this might involve mandatory audit trails, certification requirements for high-risk AI, or compulsory reporting obligations for algorithmic decision-making. The policy objective is not to suppress AI innovation but to institutionalize accountability ensuring that human oversight remains an integral component of AI governance.

Furthermore, policymakers must grapple with distributional and justice-related concerns. The deployment of AI technologies can amplify existing social inequalities and biases, particularly when algorithms are trained on skewed or discriminatory data. The ethical implications here extend beyond individual criminal liability to broader questions of collective and systemic

accountability. If an AI-driven predictive policing system disproportionately targets marginalized communities, or if automated hiring algorithms perpetuate gender or caste biases, who should bear the moral and legal responsibility? The state? The corporation? The programmer? Addressing these issues requires an integrated framework that combines ethical AI design principles with enforceable regulatory oversight to prevent algorithmic harm at both individual and societal levels.

Ultimately, the ethical and policy response to AI-related criminal liability must rest on a few guiding principles:

1. Preservation of human accountability: No AI system should function as a moral shield that absolves humans of responsibility.

2. Transparency and traceability: Every AI decision must be explainable and attributable to human design or oversight.

3. Proportional regulation: The degree of legal and ethical scrutiny should correspond to the level of autonomy and risk inherent in the system.

4. Global harmonization: Given the transnational nature of AI deployment, international cooperation is essential to avoid jurisdictional loopholes and regulatory arbitrage.

In conclusion, the attribution of criminal liability to AI demands not only legal innovation but also ethical reimagination. The challenge for contemporary policymakers is to foster an environment where technological progress and moral responsibility evolve together, rather than in opposition. The development of transparent, auditable, and accountable AI systems where every decision can be traced to human intent, design, or neglect remains the most effective safeguard against both technological impunity and moral evasion. The law must therefore evolve not just to punish harm but to prevent it through foresight, ethics, and responsible governance in the age of autonomous intelligence.

## VII. CONCLUSION

The rise of Artificial Intelligence (AI) represents not merely a technological revolution but an ontological challenge to the very foundations of criminal law. For centuries, legal systems across the world have anchored criminal responsibility in a distinctly human framework one

that assumes consciousness, free will, intention, and moral reasoning as the essential prerequisites of culpability. AI, however, disrupts this paradigm by introducing non-human agents capable of autonomous decision-making, often beyond human comprehension or control. The result is a profound doctrinal crisis: the traditional concepts of *mens rea* (guilty mind), *actus reus* (guilty act), and legal personhood are strained to their limits when confronted with entities that can act, learn, and evolve without possessing moral intent or subjective awareness.

This research has demonstrated that the human-centric structure of criminal law is increasingly ill-equipped to respond to harms caused by autonomous systems. The classical model of liability, which predicates guilt on intention, knowledge, recklessness, or negligence, falters in the face of machine learning and self-evolving algorithms. The problem of causation further compounds this difficulty, as AI often operates through complex and opaque processes that break the linear chain of human control. When a deep learning system produces an unforeseen and harmful outcome, it becomes almost impossible to trace that act back to a human *mens rea* in the traditional sense. This disconnection exposes a "responsibility vacuum" a gap where legal blame cannot be fairly or coherently assigned.

The temptation to resolve this vacuum by granting AI a form of legal or moral personhood—as proposed in certain European debates under the concept of "electronic personhood"—is, however, fraught with conceptual and ethical dangers. While such an approach may serve limited civil or regulatory functions, extending full criminal liability to AI systems remains fundamentally unsound. Criminal law is, at its core, a moral institution it presupposes consciousness, intent, and the capacity for guilt, remorse, and rehabilitation. To ascribe these qualities to an algorithmic system would be to anthropomorphize technology, thereby eroding the moral coherence upon which the entire edifice of criminal justice stands. Machines cannot "intend," "know," or "desire" in any meaningful sense; they execute code, process data, and optimize outputs according to human-designed parameters. Thus, to punish a machine would be both morally hollow and legally meaningless.

Yet, the opposite extreme ignoring AI's independent agency is equally untenable. As autonomous systems increasingly mediate critical aspects of human life healthcare, transportation, finance, law enforcement the potential for harm arising from their decisions grows exponentially. A legal regime that fails to recognize the unique nature of AI risks creating

zones of technological impunity, where neither humans nor machines are held accountable for serious harm. The task, therefore, is to design a framework that preserves the moral core of criminal law while accommodating the practical realities of autonomous behaviour.

The most pragmatic and ethically defensible solution lies in the development of a hybrid model of accountability. Such a model would integrate human responsibility and technological accountability within a unified legal structure. On the human side, doctrines of negligence, recklessness, vicarious liability, and corporate criminal liability must be expanded to cover the full spectrum of AI-related harms. Programmers, manufacturers, and deployers should bear legal responsibility for the design, testing, and supervision of AI systems, particularly where failures of oversight or foreseeability contribute to harm. Corporations, as collective entities benefiting from AI deployment, should also be held criminally accountable under identification and vicarious liability principles, ensuring that economic profit does not come at the expense of ethical or public safety obligations.

On the technological side, the law must begin to recognize functional forms of accountability rooted in technical governance rather than moral agency. This entails the establishment of mandatory standards for transparency, auditability, and explainability in AI design and operation. Every autonomous system should be built with embedded mechanisms that record, trace, and explain decision-making processes thereby enabling forensic reconstruction in the event of harm. The concept of "explainable AI" (XAI) must transition from a research ideal into a legal requirement. Regulators should mandate algorithmic auditing, ethical certification, and real-time monitoring frameworks to ensure that AI decisions remain interpretable and attributable to human oversight.

In addition, adopting graded liability frameworks could provide a proportionate response to different levels of autonomy. For low-autonomy or semi-automated systems, traditional human-based liability rules may suffice. For high-autonomy systems operating in closed-loop environments, the law could impose strict or no-fault liability on manufacturers and operators, coupled with mandatory insurance or compensation schemes to ensure that victims are not left remediless. Such approaches would maintain fairness while acknowledging the unpredictable nature of advanced AI systems.

Crucially, the evolution of criminal law in the digital age must remain anchored in ethical reasoning and social justice. As technology reshapes human interaction, the law must not

abdicate its role as the guardian of moral responsibility. Rather than reacting defensively to innovation, the legal system should evolve proactively, ensuring that accountability and justice continue to align with the values of human dignity and fairness. This requires interdisciplinary collaboration—between jurists, technologists, ethicists, and policymakers to craft legal doctrines that are both technically informed and morally grounded.

Ultimately, the future of criminal law in the era of Artificial Intelligence will depend on its ability to balance progress with principle. Law must not hinder technological advancement, but neither should it allow innovation to erode the accountability that underpins the social contract. The goal is not to punish machines, but to ensure that humans whether as designers, deployers, or beneficiaries remain answerable for the consequences of artificial decisions. In doing so, criminal law will reaffirm its enduring purpose: to safeguard justice, assign responsibility, and preserve the moral order even as the definition of agency evolves.

As we move deeper into the age of intelligent systems, the question is not whether the law can adapt it must. The true challenge lies in how it will adapt without losing sight of its ethical compass. A future-ready criminal jurisprudence must, therefore, evolve in tandem with technological progress, ensuring that accountability for harm human or algorithmic remains clear, traceable, and just. Only then can the rule of law retain its relevance and moral authority in an age where the line between human intent and machine autonomy continues to blur.

## ENDNOTES

[1] **Hallevy, G.**, "The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control" (**2010**), *Akron Intellectual Property Journal* (4(2)), 171–178.

[2] Pagallo, U. (**2013**), *The Laws of Robots: Crimes, Contracts, and Torts*. Springer, p. 95.

[3] **Yotova, R.**, "Artificial Intelligence and Legal Personality: Rethinking Accountability" (**2022**), *International Journal of Law and Technology* (14(1)), 30–35.

[4] European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

[5] Indian Penal Code, 1860, S. 11 (defining 'person').

## REFERENCES

**Hallevy, G.**, "The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control" (**2010**), *Akron Intellectual Property Journal* (4(2)), 171–205.

**Yotova, R.**, "Artificial Intelligence and Legal Personality: Rethinking Accountability" (**2022**), *International Journal of Law and Technology* (14(1)), 22–45.

European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

Indian Penal Code, 1860.

Pagallo, U. (2013). *The Laws of Robots: Crimes, Contracts, and Torts*. Springer.