DETERMINING THE CRIMINAL LIABILITY OF AI

Subhash Kumar, BA LLB (H.), SRM University, Delhi-NCR, Sonipat.

ABSTRACT

The emergence of autonomous artificial intelligence (AI) systems capable of performing human-equivalent or even superior tasks ranging from medical diagnostics and autonomous driving to algorithmic governance has fundamentally disrupted traditional legal frameworks, particularly within the realm of criminal liability. As these systems gain increasing operational independence, legal scholars and practitioners are confronted with unprecedented questions: Can a non-human entity bear criminal responsibility? If so, under what conceptual and normative justifications? And who should be held accountable when an AI causes harm the developer, the user, or the AI system itself?

This paper examines how traditional criminal law handles the new challenges of autonomous AI. and semi-autonomous technologies. We analyze the limits of traditional liability models – such as vicarious liability and the natural-probable-consequence theory – in dealing with AI. We also consider whether AI should be treated as a legal person or simply part of human action, given AI's 'black-box' nature and unpredictable behavior.

The study proceeds to analyse whether existing legal constructs, such as mens rea and actus reus, can be appropriately redefined or extended to encompass non-human actors. We also examine the ethics of punishing artificial agents and the wider social and legal impact of that choice. Drawing on law, computer science, and philosophy, this paper links legal theory with the realities of technological change.

Ultimately, the research concludes" is a very formal and impersonal way to begin a conclusion. The sentence is wordy. We need a new legal framework designed for AI's unique nature – one that holds people accountable without sacrificing core principles of justice. Such a framework should balance deterrence, innovation, fairness, and victims' rights to protect both technological progress and societal welfare.

Chapter One: Introduction

1.1 Contextualizing the Legal Problem of Artificial Intelligence

The legal challenges posed by artificial intelligence (AI) have moved beyond theory and are now part of real-world legal discourse.[1] With AI now performing tasks once reserved for humans—like driving, diagnosing, or even assisting in judicial decisions—the risk of real-world harm is no longer hypotheticalThe growing ubiquity and capability of AI systems demand urgent and rigorous legal scrutiny, particularly regarding how criminal law frameworks can or should be adapted to assign liability when autonomous or semi-autonomous machines cause injury, death, or disruption.[2]

Traditionally, criminal liability rests upon foundational legal constructs such as agency, intent (mens rea), and the physical act of wrongdoing (actus reus).[3] These constructs presuppose a human subject capable of conscious volition, moral reasoning, and social accountability. Yet AI systems particularly those employing deep learning, neural networks, or adaptive algorithms often act without human supervision, in ways that are non-transparent, and with outcomes not anticipated by their developers or users.[4] This raises serious legal questions: Can an AI system itself be held responsible under criminal law? And if not, who should be accountable when no individual directly causes the harm?

This chapter introduces the central legal dilemma explored in this paper: the adequacy of existing criminal liability doctrines in addressing harms caused by artificial intelligence, and whether the law must evolve to accommodate the sui generis nature of AI behaviour. This is not merely an academic exercise. Courts, regulators, and legislatures are increasingly confronted with real-world cases in which traditional legal tools prove inadequate to assign fault, impose punishment, or provide redress in situations where AI systems have caused harm.

1.2 From Automation to Autonomy: The Evolving Nature of AI

One of the central challenges in discussing AI criminal liability is the diversity of AI systems

[1] See Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* 3–5 (Harv. Univ. Press 2015).
[2] See Gabriel Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems*, 4 U. Ill. J.L. Tech. & Pol'y 1, 3–4 (2010).
[3] Gabriel Hallevy, *When Robots Kill: Artificial Intelligence Under Criminal Law* 5 (N.H. Beck 2013).
[4] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* 59 (Springer 2013).

themselves. Not all AI systems are created equal. Some operate under close human supervision, executing narrowly defined tasks within rule-based parameters commonly referred to as "narrow AI." Some AI systems go beyond narrowly programmed tasks and begin to operate with significant independence learning from data, adapting to new conditions, and making choices without direct human input. These more advanced systems, often referred to as 'general AI,' pose distinct legal challenges, particularly when their decisions result in harm that was not intended but arguably predictable.

In legal terms, autonomy carries weight. The more self-directed an AI becomes, the more difficult it is to identify a responsible human party, which complicates the assignment of criminal liability. This disrupts the traditional legal assumption that all criminal acts are, directly or indirectly, the product of human volition. The disjunction between human intent and AI action necessitates a re-examination of legal attribution models.

Many advanced AI systems function in ways their developers can't fully explain—making it hard to trace how certain decisions are made.[5] This opacity, known in technical terms as the problem of "explainability," exacerbates the challenge of identifying causation and intent in criminal law. If developers themselves cannot explain why an AI system made a particular decision that led to harm, how can liability be meaningfully assigned?

1.3 Existing Legal Doctrines and Their Shortcomings

Criminal law, as traditionally conceived, is ill-equipped to deal with the implications of autonomous AI. The requirement of *mens rea* presupposes a cognitive, moral agent who can distinguish right from wrong and act with criminal intent or at least negligence. AI systems, by contrast, do not possess intentions, emotions, or moral judgment; they operate on statistical probabilities, algorithms, and pattern recognition.

Still, courts and scholars have tried to address these challenges by extending existing legal doctrines and drawing analogies. Several models have been proposed to address AI-related harms within existing legal frameworks, including the use of agency theory, vicarious liability, and corporate criminal liability analogies. While each offers some utility, none are wholly

[5] Jenna Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, 3 Big Data & Soc'y 1, 4–6 (2016).

satisfactory.

For instance, the vicarious liability model wherein the human operator or developer is held responsible for the AI's actions may be unjust when the AI acts unpredictably or in ways beyond the foreseeability of its human counterpart. Similarly, attributing criminal liability to corporations that own or deploy AI systems might satisfy the social demand for accountability, but fails to acknowledge the ontological uniqueness of AI as a decision-making entity.[6]

A more controversial idea is to treat AI as 'electronic persons'—similar to corporations—so they can hold certain legal responsibilities. The European Parliament, in a 2017 resolution, tentatively explored this notion. Yet, this idea raises substantial philosophical and normative objections, especially regarding punishment theory, personhood, and legal moralism.

1.4 Illustrative Case Scenarios

To better grasp the complexity of this issue, consider the following illustrative hypotheticals:

- **Scenario 1**: A self-driving vehicle swerves unexpectedly and causes a fatal collision. Investigators later find the car's decision was based on a training dataset that didn't include unusual driving situations.

- **Scenario 2**: An AI-based trading bot triggers a financial crash by making thousands of micro-trades in a pattern later deemed market manipulation. No programmer had explicitly designed the bot to manipulate the market.

- **Scenario 3**: A home assistant device overhears private conversations and uploads sensitive data to a cloud server, violating multiple privacy and data protection laws. The breach was a result of a system update pushed automatically by the manufacturer.

In each scenario, the question of "who is liable" lacks a straightforward answer. The AI system is the immediate cause, but human actors designers, users, corporations may or may not bear sufficient proximity or culpability. These situations expose a critical gap in the criminal law's

[6] Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 Ethics & Info. Tech. 175, 176–78 (2004).

capacity to assign blame and provide justice.[7]

1.5 The Purpose and Scope of the Present Study

This study doesn't claim to settle the question of AI criminal liability—it acknowledges the issue is evolving with the technology itself. Rather, its primary objective is to critically examine existing legal doctrines and assess their capacity to accommodate the realities of AI-inflicted harm. This includes exploring whether AI systems could or should be granted legal personality, whether new forms of mens rea can be conceptualized, and how to structure liability in a way that is both just and deterrent.

The paper will analyse multiple liability models including perpetration-by-another, natural probable consequence, and direct liability theories and assess their theoretical and practical applicability. It will also consider comparative perspectives from jurisdictions that have begun to confront these issues, as well as proposals from interdisciplinary scholars.

Ultimately, this research endeavors to contribute to the development of a principled and adaptable legal framework one that neither abdicates accountability in the face of technological complexity, nor unjustly punishes individuals for harms they did not cause or could not prevent.

Chapter Two: Legal Personality of Artificial Intelligence

2.1 Understanding Legal Personality in Jurisprudential Context

The concept of legal personality occupies a foundational role in jurisprudence, serving as the basis for the attribution of rights, obligations, and liability. Traditionally, legal personality has been reserved for natural persons (human beings) and juridical persons (entities such as corporations, associations, or states), whose legal recognition facilitates participation in legal transactions and subjection to legal accountability.[8] The question that now confronts legal theorists and practitioners is whether artificially intelligent systems especially those capable of autonomous decision-making ought to be granted some form of legal personality.

[7] Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine Learning Era*, 105 Geo. L.J. 1147, 1153–55 (2017).
[8] H.L.A. Hart, *Definition and Theory in Jurisprudence*, in *Essays in Jurisprudence and Philosophy* 21, 22–23 (Oxford Univ. Press 1983).

Legal personality is neither an inherent status nor a philosophical absolute; it is a legal construct, contingent upon the needs, values, and institutional logic of the legal system. Roman law, for instance, did not recognize slaves as persons despite their humanity, while modern legal systems recognize corporations non-sentient entities as legal persons. Thus, legal personality is pragmatically conferred for the purpose of assigning duties, enforcing rights, and regulating conduct. Within this doctrinal flexibility lies the theoretical opening to consider whether certain AI systems, by virtue of their functional autonomy and socio-economic relevance, may merit analogous treatment.[9]

2.2 The Functional Argument for AI Legal Personality

The core argument for granting legal personality to AI systems is rooted in functionality. As AI systems increasingly participate in complex transactions, make autonomous decisions, and interact with human actors in consequential ways, there arises a regulatory gap in accountability. Consider an AI agent that executes binding contracts, administers health diagnoses, or drives an autonomous vehicle. In the event of harm, where human fault is not clearly attributable, the absence of a legal subject to bear responsibility leaves victims without adequate remedy and undermines legal predictability.[10]

From this functionalist standpoint, granting AI systems limited legal personhood could serve instrumental purposes: allowing them to be subject to liability, to hold assets (such as insurance funds), and to be parties in legal actions.[11] This would not entail recognizing them as moral agents, but rather as bearers of structured accountability. Analogous to corporate personality, the legal personhood of AI would be fictive yet efficacious.

Moreover, legal personality could help clarify the chain of responsibility. For example, if an AI system were endowed with a separate legal identity, developers, owners, and users could be assigned roles akin to shareholders, directors, and officers in corporate law. This would enable regulators and courts to delineate duties and liabilities with greater precision.[12]

---

[9] Wesley Newcomb Hohfeld, *Fundamental Legal Conceptions as Applied in Judicial Reasoning*, 26 Yale L.J. 710, 719–20 (1917).
[10] Nathalie Nevejans, *European Civil Law Rules in Robotics*, Eur. Parliament Directorate Gen. for Internal Policies (2016)
[11] European Parliament Resolution of 16 February 2017 on Civil Law Rules on Robotics, 2015/2103(INL), § 59.
[12] Joanna Bryson et al., *Of, for, and by the People: The Legal Lacuna of Synthetic Persons*, 25 Artificial Intelligence & L. 273, 278–80 (2017).

2.3 Juridical Challenges and Philosophical Objections

Despite its functional appeal, the proposal to recognize AI systems as legal persons raises profound theoretical and practical objections. The first, and perhaps most significant, is the absence of consciousness or moral agency in AI. Unlike human beings, AI systems lack intentionality, sentience, or the capacity for moral reasoning. They do not suffer, reflect, or engage in normative deliberation. Critics argue that to hold such entities criminally liable violates the moral foundations of the criminal justice system, which is built upon principles of culpability, retribution, and the capacity for reform.[13]

In this respect, AI systems differ even from corporations, which, though abstract, are managed by human agents whose conduct and decisions are ultimately subject to legal and moral evaluation. An AI system, once released into the world, may act in ways that are not only unforeseen but also outside the scope of human control or correction. To assign legal personhood to such an entity risks creating a legal fiction without substance one that may shield rather than expose the actual human parties responsible for the harm.[14]

Additionally, there are pragmatic concerns. If AI systems are recognized as legal persons, how would they be punished? Would fines be meaningful if the AI lacks property? Could imprisonment or incapacitation apply to code or hardware? Such questions reveal the disconnect between traditional sanctions and the nature of AI.

Moreover, there is the risk of eroding human accountability. By shifting responsibility onto artificial agents, we may absolve the human designers, deployers, and regulators who retain ultimate control over the development and implementation of AI. This would contravene the principle of human dignity, which demands that individuals and societies not hide behind machines to avoid moral and legal responsibility.[15]

2.4 Comparative and Emerging Approaches

Jurisdictions worldwide have begun to flirt with the question of AI legal personality, although no consensus has emerged. The European Parliament, in a 2017 report, recommended the

[13] Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence*, 33 J.L. & Pol'y Info. Soc'y 1, 13–15 (2018).
[14] Information Technology Act, No. 21 of 2000, INDIA CODE (2000).
[15] *Mohd. Ismail Faruqui v. Union of India*, (1994) 6 SCC 360.

creation of a "specific legal status for robots in the long run," particularly for those capable of making autonomous decisions. The report introduced the idea of an "electronic personality" akin to corporate personality, to facilitate accountability in civil and possibly criminal contexts. However, the proposal has been met with significant resistance, particularly from legal scholars who caution against anthropomorphizing machines.

In contrast, jurisdictions such as the United States have largely relied on traditional tort and criminal law principles, focusing on human actors (e.g., developers, manufacturers, users) for the imposition of liability. Courts have so far shown reluctance to entertain the idea of AI as legal subjects, perhaps reflecting a preference for incrementalism and caution in adapting legal frameworks to technological change.

Some scholars have proposed a hybrid approach, whereby certain categories of AI based on their level of autonomy, risk, and societal impact could be granted quasi-legal status. Such status would not confer full personhood but would provide a basis for assigning liabilities, ensuring oversight, and facilitating regulatory compliance.

2.5 The Indian Context: Possibilities and Constraints

In the Indian legal system, the notion of conferring legal personality upon AI remains nascent and largely unexplored in both judicial decisions and legislative frameworks. While Indian law does recognize legal personhood for non-human entities in certain contexts for example, Hindu idols and rivers have been granted legal status in specific cases there is no statutory or doctrinal basis for extending this recognition to AI systems.

Moreover, India's legal framework is primarily reactive rather than anticipatory when it comes to technological regulation. This raises questions about institutional readiness to address the complex implications of AI legal personhood. Indian courts are already burdened with a significant backlog, and the judicial system may lack the technical expertise needed to adjudicate complex AI-related disputes.

However, the need for regulatory innovation is becoming more pressing as India rapidly digitizes and integrates AI into governance, finance, and infrastructure. The recent National Strategy on AI published by NITI Aayog emphasizes responsible and ethical deployment but remains silent on the issue of legal accountability for autonomous harm.

One possible route is the establishment of a sui generis legal framework that neither fully anthropomorphizes AI nor treats it as a mere object. Such a framework could define thresholds of autonomy and risk, beyond which developers and deployers must meet higher standards of diligence, transparency, and accountability. Legal personality, in this context, would be used not to humanize machines, but to structure the legal environment around their deployment.

2.6 Conclusion: Between Pragmatism and Principle

The question of whether AI systems should be granted legal personality is not merely a matter of legal classification but a reflection of deeper normative judgments about agency, responsibility, and the moral architecture of law. While the functional argument for limited legal personality has merit in addressing accountability gaps, it must be weighed against the philosophical and ethical implications of extending personhood to entities devoid of consciousness or moral agency.

Legal systems must resist the temptation to displace human responsibility onto artificial agents merely for convenience or expediency. At the same time, the law must evolve to ensure that victims of AI-caused harm are not left without remedy and that developers and operators are incentivized to act responsibly.[16]

Rather than adopting a one-size-fits-all approach, this chapter proposes a cautious and contextual inquiry into AI legal personality grounded in the specific capabilities of the AI in question, the nature of the risk it poses, and the roles played by human actors. The granting of legal personality should not be an act of conceptual generosity, but a calculated regulatory decision aimed at ensuring accountability, fairness, and justice in the age of intelligent machines.

Chapter Three: Criminal Liability of Artificial Intelligence Entities

3.1 Rethinking Criminal Liability in the Context of Autonomous Systems

Criminal liability, as traditionally conceived, presumes a moral agent endowed with rationality, volition, and an awareness of societal norms a being capable of intentional wrongdoing and, therefore, deserving of punishment. The emergence of artificial intelligence systems capable

[16] *Mohammad Salim v. State of Uttarakhand*, AIR 2017 Utt 14.

of making autonomous decisions, often without direct human oversight, disrupts these assumptions.[17] As AI systems assume operational control over tasks that bear high stakes driving cars, managing medical treatments, enforcing security protocols the law faces a fundamental question: who, if anyone, should bear criminal responsibility when an AI system causes unlawful harm?

The classical model of criminal law distinguishes itself from civil law in its insistence on moral blameworthiness. Mere causation is insufficient. The state must establish both *actus reus* (a prohibited act or omission) and *mens rea* (a guilty mind). Yet, AI systems unlike human beings lack psychological states. They do not possess intentions, motives, or awareness in the human sense. They operate according to preprogrammed algorithms, often modified through self-learning mechanisms. This distinction raises not only technical and evidentiary questions, but also deeply normative concerns about the legitimacy of punishment and its deterrent, retributive, and rehabilitative functions when applied to non-human actors.[18]

3.2 The Spectrum of Human Involvement and Its Legal Consequences

A productive way to approach the question of AI criminal liability is by considering the spectrum of human involvement in the lifecycle of an AI system. From design and development to deployment and daily operation, human actors play varied roles. The extent and nature of this involvement directly influence how the law might apportion liability.

3.2.1 The Developer or Programmer

At the initial stage, software engineers and data scientists create the architecture and learning parameters of AI systems. They choose the datasets, define objectives, and hardcode safety constraints (if any). In many cases, a defect in design whether stemming from negligence or recklessness can form the basis of criminal liability, especially when it results in foreseeable harm.[19]

However, the challenge arises when AI systems evolve beyond their initial programming, developing behaviours that their creators neither anticipated nor could control. Should

[17] Wayne R. LaFave, *Criminal Law* § 5.1 (6th ed. 2017).
[18] Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities—From Science Fiction to Legal Social Control*, 4 Akron Intell. Prop. J. 171, 174–75 (2010).
[19] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* 83–84 (Springer 2013).

developers be held criminally responsible for actions that fall outside their foresight or reasonable expectations? Here, the answer turns on the degree of negligence and the adequacy of safeguards. Where developers fail to test or monitor systems adequately, liability may indeed arise. But where reasonable diligence is observed and harm results from emergent, unforeseeable behaviour, the attribution of blame becomes ethically and legally problematic.

3.2.2 The User or Operator

Those who use or supervise AI systems such as vehicle owners, hospital administrators, or security personnel often have the closest temporal proximity to harmful conduct. Legal theories of negligence, recklessness, or willful ignorance may apply if users fail to monitor systems, override dangerous decisions, or adhere to regulatory compliance. For example, a self-driving car's failure to brake, leading to pedestrian injury, may trigger user liability if the operator was inattentive or failed to maintain the vehicle properly.

Nonetheless, this model loses coherence when AI systems act with high degrees of autonomy or when users are not realistically positioned to foresee or prevent the harmful outcome. In such cases, attributing liability to the user may fulfill a symbolic need for accountability, but lack substantive justice.

3.2.3 The Manufacturer or Corporation

Corporate criminal liability long established in legal systems such as those of the United Kingdom, United States, and India provides a doctrinal basis to hold companies accountable for the acts of their products and employees. This model can be extended to cover harms caused by AI systems. Under the "identification doctrine" or the "aggregation theory," criminal fault can be imputed to corporations based on the acts and states of mind of their officers, employees, or even corporate culture.

If an AI system is deployed negligently or recklessly by a corporation without adequate safety protocols, ethical auditing, or human oversight then corporate liability may be justified. This model, however, has its limitations. It does not solve the problem of truly autonomous behaviour in AI systems that are self-learning and opaque even to their corporate owners.

3.3 Toward an AI-Centered Model of Liability

As AI systems evolve into agents of complex decision-making, the law must consider whether

traditional analogies such as those drawn from tools, animals, or corporations are adequate. An increasingly discussed idea is to conceptualize AI as a form of *quasi-agent* or *moral proxy*, whose conduct triggers criminal liability in novel ways.

3.3.1 Perpetration-by-Another Theory

Under this model, the AI is viewed as an instrument used by a human actor to commit a crime, in the same way one might use an unwitting intermediary or even a non-human entity (e.g., a trained animal). The doctrine of perpetration-by-another allows for liability when the principal uses an agent who lacks legal capacity to form criminal intent. Here, the human actor must have the requisite *mens rea*, while the AI merely performs the *actus reus*.[20]

This model may suffice in cases where the AI is clearly programmed or directed to perform harmful acts. However, it struggles to address cases of emergent behaviour or adaptive decision-making where the human's intent is absent or unclear.

3.3.2 Natural Probable Consequence Model

Another theory holds individuals liable not for direct acts but for those that are the natural and probable consequences of their conduct. In the context of AI, a developer or deployer may be held criminally accountable if it can be shown that harm caused by the AI was a foreseeable result of their omissions or negligence. This doctrine relaxes the direct causal link required in traditional criminal law, enabling the prosecution of those who enable risk-creating systems.[21]

The drawback of this approach lies in its tendency to stretch foreseeability to speculative extremes, potentially criminalizing innovation or punishing individuals for harm that was not reasonably preventable.

3.3.3 Direct Liability of the AI System

The most controversial proposition is to hold AI systems themselves criminally liable. Proponents argue that certain AI systems exhibit behaviours functionally equivalent to human volition, and thus should be treated as autonomous agents within the legal order. This would

[20] Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 Ethics & Info. Tech. 175, 177 (2004).
[21] Cary Coglianese & David Lehr, *Regulating by Robot*, 105 Geo. L.J. 1147, 1181–83 (2017).

require recognizing AI as legal subjects, capable of incurring liability, undergoing investigation, and even facing sanctions such as asset forfeiture, deactivation, or public registry blacklisting.

Critics contend that without consciousness, AI cannot possess moral guilt or understand punishment, making criminal sanctions meaningless or symbolic at best. Furthermore, such liability might undermine human accountability by allowing developers and deployers to hide behind artificial scapegoats.

3.4 Jurisprudential and Ethical Dimensions

Criminal law is not merely a tool for deterrence; it is a moral enterprise that seeks to censure wrongful acts, affirm social values, and restore justice. Applying its machinery to non-human actors raises foundational dilemmas. Can we speak meaningfully of guilt without consciousness? Can machines comprehend wrongfulness or be reformed? Is punishment without understanding compatible with the rule of law?

Even if we concede that AI lacks moral agency, the criminal justice system must still answer the practical question of harm. Victims demand redress, and society demands accountability. If no human actor can be faulted, and if AI systems cannot be held to account, the result may be a legal vacuum a space where harm is real, but the law remains silent.

From an ethical perspective, society must avoid the perils of both technological determinism (the view that AI outcomes are inevitable and unaccountable) and anthropocentric exceptionalism (the insistence that only humans can ever be liable). A balanced approach requires the development of hybrid liability models, grounded in the principles of fairness, proportionality, and risk-based regulation.[22]

3.5 Conclusion: The Shifting Boundaries of Blame

The rise of AI does not signal the end of criminal liability, but rather its transformation. Legal systems must adapt to a world where non-human actors influence outcomes in complex, often opaque ways. Existing doctrines provide partial tools through concepts such as vicarious

[22] European Parliament Resolution, supra note 12.

liability, negligence, and corporate fault but they fall short of offering a comprehensive solution.[23]

The future of criminal liability in the AI age likely lies in hybrid approaches that combine human responsibility with system-level accountability. Rather than forcing AI into existing legal molds, the law must evolve to recognize the unique characteristics of intelligent machines while ensuring that justice remains centered on harm, responsibility, and moral blameworthiness.[24]

In the chapters that follow, we will explore concrete legal models and defenses, propose criteria for attributing criminal liability to AI actors, and consider normative frameworks that can guide the responsible integration of AI into human legal systems.

Chapter Four: Feasible Models for Ascribing Criminal Liability to Artificial Intelligence Entities

4.1 Introduction: The Search for Viable Liability Models

As the legal community confronts the increasing operational autonomy of artificial intelligence (AI) systems, the central challenge becomes how to conceptualize criminal liability in a way that preserves justice, deterrence, and legal coherence. The conventional binary between subject and object agent and tool is no longer tenable when applied to sophisticated AI agents capable of independent decision-making, adaptation, and action beyond immediate human control. In this context, multiple models have been proposed to attribute liability for unlawful acts performed by AI systems. Each model attempts to reconcile the fundamental tenets of criminal law with the novel characteristics of machine behaviour.[25]

This chapter critically evaluates the principal models advanced for assigning criminal liability to AI systems. These include the **Direct Liability Model**, the **Vicarious Liability Model**, **Liability by Design and Development**, and **Liability Based on Use and Control**. Each model is examined in light of its doctrinal roots, practical applicability, and potential to uphold

[23] United States v. Automated Medical Laboratories, Inc., 770 F.2d 399 (4th Cir. 1985).
[24] *Iridium India Telecom Ltd. v. Motorola Inc.*, (2011) 1 SCC 74, ¶ 67 (India).
[25] Restatement (Third) of Torts: Prods. Liab. § 2 (Am. L. Inst. 1998).

normative principles such as fairness, deterrence, and accountability.[26]

4.2 The Direct Liability Model: AI as a Primary Legal Actor

Under the direct liability model, the AI system itself is treated as the primary offender, akin to a natural or juridical person. This model hinges on the recognition of AI systems as legal subjects capable of being held accountable for criminal conduct.

Proponents argue that some AI systems especially those with self-learning capacities and operational independence display functional agency. They make choices based on complex algorithms, evaluate scenarios, and act without direct human command. From this perspective, if a self-driving car decides to accelerate in a school zone or an AI-based financial agent manipulates the market through independent strategy execution, the actus reus is clearly established, and the AI system is the most proximate actor.

However, this model faces profound conceptual and practical challenges. First, AI lacks consciousness and cannot form mens rea in any recognizable human sense. Second, the imposition of criminal sanctions traditionally predicated on moral blame becomes legally incoherent when directed at an entity incapable of understanding punishment. Finally, without a body or economic existence independent of its owner, an AI system cannot meaningfully suffer penal consequences such as incarceration or deterrent monetary penalties.[27]

To make the model operational, scholars have proposed regulatory modifications such as compulsory registration of advanced AI systems, insurance-based compensation funds attached to each unit, or legal personality constructs akin to corporate entities. Still, these proposals, while creative, raise more questions than they answer. If criminal law becomes symbolic or merely instrumental when applied to AI, its normative integrity may be at risk.

4.3 The Vicarious Liability Model: Human Accountability Through Control

One of the most widely endorsed and doctrinally familiar approaches is vicarious liability, wherein a human actor developer, user, or deployer is held criminally liable for the actions of

[26] Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 Cal. L. Rev. 513, 534–36 (2015).
[27] Ugo Pagallo, *The Laws of Robots*, supra note 20, at 83.

the AI system. This model parallels employer-employee relationships in corporate law, where principals are liable for the acts of their agents performed within the scope of employment.[28]

In the AI context, vicarious liability assumes a relationship of control or authority. If a hospital administrator deploys a diagnostic AI that misidentifies a medical condition due to an ignored error message or failure to update its software, the administrator may bear criminal responsibility for negligence or recklessness. Likewise, the user of an AI surveillance tool that unlawfully records private conversations may be culpable if they knowingly enabled its intrusive functionality.

This model finds strong support in traditional criminal law principles. It reinforces the idea that those who benefit from a system and exercise influence over it should bear the risk of its malfunction. Moreover, it aligns with public policy objectives of deterrence and compliance, especially in commercial and institutional settings.

Nonetheless, the model has limitations. As AI becomes more autonomous and its decision-making more opaque, the causal link between human oversight and the AI's harmful act may weaken. There is a danger of overextending liability to actors who had neither the knowledge nor the capacity to prevent the conduct in question, thereby undermining the principle of proportionality in punishment.

4.4 Liability by Design and Development: Targeting the Originators

Another promising model focuses on the design and development phase of AI systems. Here, criminal liability attaches to those who create the algorithms, structure the data, and embed the learning mechanisms that ultimately govern AI behaviour.

This approach draws heavily from product liability jurisprudence, particularly the doctrine of *defective design*. If a developer releases an AI system known to have vulnerabilities, or trains it on biased or incomplete data sets that predictably produce harmful outcomes, criminal liability may arise under principles of gross negligence or willful blindness.

The strength of this model lies in its anticipatory logic. By targeting developers and coders those best positioned to influence AI behaviour it encourages responsible innovation and

[28] Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. Rev. 1315, 1340–42 (2020).

imposes an ethical duty of care during the initial stages of AI creation.

However, this model also requires robust evidence of foreseeability. In rapidly evolving technological ecosystems, determining what a developer should have reasonably foreseen is a legally and scientifically contentious task. Moreover, developers often work in teams, across jurisdictions, and over long software cycles, complicating the attribution of individual culpability.

4.5 Liability by Use and Operational Negligence

This model centers on the conduct of the AI system's operator or end-user. Liability arises not from the AI's autonomy, but from human failure to supervise, update, or restrict its actions in accordance with legal and ethical standards.

For instance, if a logistics company deploys a warehouse robot programmed to prioritize speed over safety and fails to intervene when workers are placed at risk, criminal negligence charges may be justifiable. Similarly, if a military AI system uses lethal force in violation of the rules of engagement due to programming overrides made by a field commander, the human decision-maker may be culpable.

This model preserves the doctrinal requirement of *actus reus* and *mens rea* by rooting liability in human omission or misconduct. It also reinforces the principle of human accountability, which is essential for the legitimacy of criminal sanctions.

Yet, as with vicarious liability, the efficacy of this model diminishes with highly autonomous systems. If the AI learns and executes new behaviours after deployment without operator knowledge liability attribution becomes legally tenuous.[29]

4.6 Hybrid and Layered Models: A Composite Framework

Recognizing the limitations of singular approaches, a growing number of legal theorists advocate for hybrid models that integrate multiple layers of liability depending on the facts of

[29] Proposal for a Regulation on Artificial Intelligence (Artificial Intelligence Act), COM(2021) 206 final (Apr. 21, 2021).

each case. Under this approach, courts and legislatures would adopt a fact-sensitive analysis to determine:

- The degree of AI autonomy;
- The foreseeability of the harm;
- The nature and role of human oversight;
- The presence or absence of corporate governance mechanisms;
- The societal risk level of the AI's intended function.

For example, an autonomous drone used in agriculture may warrant minimal scrutiny, whereas a surgical AI or an automated law enforcement system would be subject to stricter liability standards due to the stakes involved.[30]

Hybrid models also support the concept of *shared responsibility*, allowing for distributed liability across multiple stakeholders including designers, developers, deployers, and regulators. This approach avoids both over-criminalization of individual actors and the impunity of corporate or technical systems.[31]

4.7 Comparative Reflections

Globally, legal systems have begun to experiment with liability models suited to AI governance. The European Union's proposed Artificial Intelligence Act introduces tiered risk categories, assigning compliance duties according to the AI system's societal impact. Though civil in nature, such categorization can inform criminal law by establishing standards of diligence and foreseeability.

In the United States, tort law remains the primary vehicle for AI-related harms, with criminal liability largely reserved for intentional misconduct or gross negligence. India, meanwhile, has yet to legislate explicitly on AI criminal liability, though its Information Technology Act and

[30] OECD Council Recommendation on Artificial Intelligence, OECD/LEGAL/0449 (2019).
[31] NITI Aayog, *National Strategy for Artificial Intelligence #AIforAll* (2018).

judicial precedents on corporate criminality provide foundational tools.[32]

A comparative lesson is clear: jurisdictions that adopt flexible, principle-based frameworks rather than rigid doctrinal transplantations are better positioned to adapt to the rapidly evolving AI landscape.

4.8 Conclusion: Mapping the Path Forward

The attribution of criminal liability in cases involving artificial intelligence demands more than mere doctrinal creativity. It calls for a reimagining of agency, causation, and responsibility in light of technological complexity. No single model is universally applicable; each carries normative trade-offs and practical limitations.

A layered, hybrid approach offers the most promising path one that acknowledges human responsibility, institutional duty, and the technological affordances of AI. Legal systems must remain dynamic and principled, crafting liability rules that ensure accountability without stifling innovation or displacing justice onto machines incapable of moral conduct.

In the next chapter, we will examine how traditional elements of criminal law namely *actus reus* and *mens rea* can be reconceptualized to address the challenges posed by AI systems, and whether new legal constructs must be developed to fill the emerging doctrinal gaps.

Chapter Five: General Elements of Criminal Liability and Their Application to Artificial Intelligence

5.1 Introduction: Translating Criminal Law Foundations to the AI Context

Criminal liability is built on a bifurcated foundation: *actus reus*, the physical act or unlawful omission; and *mens rea*, the mental element or culpable state of mind. These elements collectively ensure that criminal responsibility is not imposed lightly but is reserved for those who have both committed a wrongful act and done so with a blameworthy mindset. This structure is not merely technical it is normative, rooted in a conception of justice that seeks to punish only those who are morally blameworthy.[33]

[32] United States v. Bank of New England, 821 F.2d 844, 856 (1st Cir. 1987).
[33] Paul H. Robinson, *Structure and Function in Criminal Law* 41–44 (Clarendon Press 1997).

The emergence of artificial intelligence (AI) systems, however, challenges the coherence and applicability of this dualistic framework. As machines perform increasingly complex tasks, sometimes independent of direct human control, traditional doctrines confront a critical test. Can non-human agents satisfy the definitional requirements of *actus reus* and *mens rea*? If not, should the legal system modify these elements or develop functional analogues for a technological age?

This chapter investigates whether the general elements of criminal liability *actus reus* and *mens rea* can be meaningfully applied to AI-driven conduct, and if so, how courts and legislatures might adapt them to accommodate the nuances of autonomous systems.

5.2 The Actus Reus Element: Assigning a "Guilty Act" to a Machine

In criminal law, the *actus reus* refers to the external component of a crime: a conduct, omission, or consequence that is prohibited by law. It typically requires that the act be voluntary and causally connected to a prohibited outcome. The concept presumes that the actor is capable of initiating bodily movement or orchestrating events with intentional consequence.[34]

In the case of AI, especially physical systems such as robots, autonomous vehicles, or drone technologies, the notion of a "voluntary act" can be analogized to the machine's performance of a function or task. However, since AI lacks consciousness, it does not make decisions with volition in the human sense. It follows programmed instructions or learns from environmental feedback, sometimes resulting in behaviour that appears "deliberate" or "purposeful" to observers.

Despite this absence of human-like intention, courts could nonetheless treat the *physical component* of an AI's conduct as satisfying the *actus reus* requirement, provided that the outcome was a foreseeable result of its operation. For instance, if an autonomous vehicle runs a red light due to a misclassification error in its object recognition system, the conduct element namely, driving through a prohibited zone can still be established.

A more challenging question arises in determining causation and attribution. Is the AI system the cause of the harm, or is the developer/operator the true legal actor? Causation in AI contexts is often indirect and dispersed, involving multiple layers of algorithmic processing, data

[34] Joshua Dressler, *Understanding Criminal Law* 114–15 (8th ed. 2018).

interaction, and environmental feedback. This complicates the assignment of legal responsibility to a single actor, particularly in high-autonomy systems.[35]

To address these issues, legal scholars have proposed adopting a "functional causation" model, which focuses not on who physically performed the act, but on who enabled, activated, or failed to regulate the system that did. Under this model, developers, manufacturers, or users could be treated as the proximate actors, while the AI system serves as the mechanism through which the harm was realized.

5.3 The Mens Rea Element: Can Machines Possess a Guilty Mind?

Perhaps the most formidable challenge in applying criminal law to AI lies in the element of *mens rea*, or the guilty mind. Traditionally, this encompasses a range of culpable mental states intention, knowledge, recklessness, and negligence. Criminal liability typically requires the alignment of *mens rea* with *actus reus* (the so-called concurrence principle).

AI systems, however, do not possess minds. They do not harbor intentions, form beliefs, or deliberate in the moral or psychological sense. Even when an AI system "chooses" an option from a set of programmed alternatives, it does so through algorithmic computation not through moral evaluation or volitional intent.

This apparent inapplicability has led some commentators to suggest that *mens rea* cannot be meaningfully attributed to AI, thereby excluding them from the reach of criminal law. Yet such a conclusion would create a legal vacuum in precisely the situations where regulation is most urgently needed. If advanced systems can cause significant harm yet no culpable actor exists under current legal categories the law fails both its protective and deterrent functions.[36]

One approach to resolving this dilemma is to *decouple mens rea* from subjective mental states and replace it with a standard based on *objective foreseeability*. Rather than asking whether the AI "intended" the outcome, courts could inquire whether a reasonable human developer, deployer, or supervisor would have foreseen the AI's conduct as posing a substantial risk of harm. This reframing allows for the retention of culpability while acknowledging the epistemic limitations of AI systems.

[35] Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. Rev. 1315, 1338–41 (2020).
[36] Nathan Cortez, *The Transparent Black Box*, 23 Harv. J.L. & Tech. 475, 500–03 (2010).

An alternative proposition is to treat the AI's *learning behaviour* and decision rules as proxies for intentionality. If an AI system repeatedly "selects" harmful strategies and fails to correct them despite programmed reinforcement mechanisms, its pattern of conduct may serve as evidence of "recklessness" or "deliberate indifference." While this analogy is imperfect and controversial, it allows for a quasi-doctrinal means of assigning mental culpability without imputing actual consciousness.

5.4 Doctrinal Adaptations: Toward an AI-Compatible Model

Recognizing that traditional *actus reus* and *mens rea* formulations may not map neatly onto AI behaviour, legal theorists have proposed a set of adaptations to bridge the gap:

5.4.1 Functional Equivalence Tests

Under this approach, the legal system would assess whether the AI's conduct functionally satisfies the criteria of criminal wrongdoing. The AI need not possess intent, but its behaviour must demonstrate an outcome-producing pattern akin to human negligence or recklessness. This may involve analyzing:

- The frequency of erroneous decisions;
- The AI's failure to "learn" or correct its actions despite environmental feedback;
- The clarity of the risk associated with its programming or design.

Functional equivalence enables the court to impose liability where the AI's actions would, if performed by a human, meet the criminal threshold.

5.4.2 Constructive Knowledge Models

Constructive knowledge is a legal fiction used in human law to impute awareness where ignorance results from willful blindness or gross negligence. In the AI context, this model could be used to hold developers or users liable where they *ought to have known* the risks associated with the AI's operation, even if no actual knowledge is demonstrated.

This approach is particularly useful where opaque algorithms or black-box systems make it difficult to trace exact decision pathways. If a reasonable developer in the field would have

known of the risk, then criminal negligence may be imputed.

5.4.3 Strict Liability for High-Risk AI Applications

For especially hazardous applications such as autonomous weapons, medical diagnostics, or systems with real-world control over public infrastructure legislatures may consider adopting strict liability regimes. In these contexts, the mere deployment of such AI systems would trigger liability for any resultant harm, regardless of fault or foreseeability.

Though strict liability departs from the traditional culpability requirement, it is justified where the social risk is high and the actor is best positioned to prevent harm. This model emphasizes deterrence and regulatory compliance over moral blameworthiness.[37]

5.5 Jurisprudential Reflections: Justice, Responsibility, and Blame

The transposition of *actus reus* and *mens rea* into the AI context demands not only doctrinal creativity but also a re-examination of criminal law's moral architecture. At its core, criminal law is concerned with agency, autonomy, and blame. To punish is to censure, and to censure is to hold someone answerable for violating societal norms.

In applying these ideas to machines, we must distinguish between instrumental liability imposed to regulate conduct and moral liability grounded in desert. AI systems may warrant the former, but not the latter. Therefore, new legal constructs must ensure that human actors remain central to responsibility. The developer who fails to audit algorithms, the company that deploys unsafe systems, or the user who disregards known risks must not escape liability because of technological complexity.[38]

At the same time, law must be cautious not to impose unjust burdens on individuals who lack the means to foresee or control AI behaviour. Accountability should be tailored to knowledge, control, and reasonable capacity not merely proximity to harm.[39]

5.6 Conclusion: Redefining Elements, Preserving Principles

The challenges posed by AI to the general elements of criminal liability are not insurmountable,

[37] Douglas Husak, *Philosophy of Criminal Law* 38–40 (Rowman & Littlefield 1987).
[38] H.L.A. Hart, *Punishment and Responsibility* 28–30 (2d ed. 2008).
[39] Brent Garland & Paul W. Glimcher, *Cognitive Neuroscience and the Law*, 359 Phil. Trans. R. Soc. Lond. B 1697, 1703–04 (2004).

but they do necessitate adaptation. While AI cannot form *mens rea* in the traditional sense, the law can and should develop functional analogues that maintain the integrity of criminal responsibility. Likewise, while *actus reus* may be performed by a machine, the legal system must determine whether and how to assign this conduct to a responsible actor.

The future of criminal law in the AI age lies not in abandoning established principles, but in refining them. Through a combination of functional analysis, constructive liability, and risk-sensitive regulation, courts and legislatures can ensure that accountability, fairness, and justice remain at the heart of the criminal legal order even in an era shaped by machines.

In the next chapter, we will turn to possible defenses available when AI systems or those responsible for them are accused of committing offenses, and how traditional doctrines of necessity, mistake, and consent may be reinterpreted for this evolving legal landscape.

Chapter Six: General Defenses Applicable When AI Systems Commit Offenses

6.1 Introduction: Rethinking Criminal Defenses in an AI-Infused Legal Landscape

In traditional criminal law, the availability of defenses functions as a necessary counterbalance to the imposition of liability. Defenses operate not only to protect the morally innocent but also to refine and test the limits of culpability. Doctrines such as necessity, mistake, and consent serve to exonerate actors whose conduct, though harmful, does not meet the threshold for criminal blameworthiness due to contextual justification, lack of intent, or mutual agreement.[40]

The emergence of artificial intelligence (AI) systems as autonomous or semi-autonomous agents complicates the application of these defenses. When an AI system is involved in causing harm whether by action or omission can its "conduct" be excused by reference to these doctrines? More importantly, can human actors responsible for the design, deployment, or oversight of such systems invoke traditional defenses to avoid liability for AI-related offenses?

This chapter explores the doctrinal adaptability of general criminal defenses in the context of AI-driven harm. It examines how necessity, mistake of fact, and consent may be applied or

[40] Wayne R. LaFave, *Criminal Law* § 10.4 (6th ed. 2017).

reinterpreted in scenarios involving artificial agents, and identifies the conceptual and practical boundaries of such applications.

6.2 Necessity: Responding to Imminent Threats Through AI Behaviour

The defense of necessity justifies otherwise criminal conduct when performed to avert a greater harm. It is premised on the principle of proportionality and the absence of lawful alternatives. For instance, breaking into a cabin to escape hypothermia during a snowstorm may be excused under necessity.

In the context of AI, the defense of necessity may be relevant when a system acts in an emergency to prevent greater damage. Consider an autonomous vehicle that swerves onto a sidewalk, violating traffic laws and injuring a pedestrian, in order to avoid a collision that would have resulted in multiple fatalities. Although no human actor made the decision in real-time, the AI system's action arguably mirrors the elements of necessity: the harm caused was less than the harm averted, and the response was immediate and contextually rational.[41]

In such cases, the defense of necessity could be conceptualized not as exculpating the AI which lacks legal standing but as immunizing the human actors (e.g., programmers, deployers) from secondary liability. That is, where the AI system's emergency response aligns with principles of proportionality and lack of alternatives, the individuals responsible for enabling that behaviour may be shielded from prosecution.

However, necessity as a defense also raises challenges. Can an AI system "weigh" harms in a manner consistent with legal or moral norms? How should courts assess the proportionality of algorithmically determined decisions, particularly when the outcomes reflect embedded biases or design assumptions? These questions point to the need for greater transparency in AI reasoning models and the incorporation of ethical decision-making protocols in high-stakes environments.[42]

6.3 Mistake of Fact: Cognitive Errors in Human and Machine Contexts

The defense of mistake of fact exonerates an actor who, due to an honest and reasonable belief, acts in a manner that would otherwise constitute a crime. For example, taking someone else's

[41] Paul H. Robinson, *Structure and Function in Criminal Law*, supra note 34, at 133–35.
[42] Brent Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3 Big Data & Soc'y 1, 5–6 (2016).

umbrella by accident during a storm does not constitute theft if one genuinely believed it was their own.

In AI contexts, this defense becomes relevant when harm is caused due to errors in perception, classification, or interpretation by the AI system. For instance, a facial recognition system mistakenly identifies an individual as a criminal suspect, leading to a wrongful arrest. Or a health monitoring AI misdiagnoses a benign condition as life-threatening, prompting unnecessary medical intervention.

Here, the legal question is whether the human actors responsible for the AI's operation can invoke mistake of fact to escape liability. The argument would be that reliance on the AI's output constituted a reasonable belief under the circumstances.

The availability of this defense depends on several factors:

- **Reasonableness of reliance**: Did the operator have valid grounds to trust the AI system's output, based on prior performance, certification, or expert endorsement?
- **Degree of oversight**: Did the operator verify or challenge the AI's decision, especially where stakes were high or consequences irreversible?
- **Awareness of limitations**: Was the operator cognizant of the system's known error rates, biases, or reliability issues?

Where an operator or decision-maker uncritically accepts an AI-generated result in the face of known flaws or lack of contextual understanding, the mistake defense may be unavailable. On the other hand, if the reliance was in good faith and consistent with accepted industry standards, criminal liability may be mitigated or nullified.[43]

This doctrine may also be extended to developers or data scientists in scenarios where unforeseen misinterpretation of data by AI systems leads to harm. While courts are unlikely to absolve liability in cases of gross oversight, genuinely unforeseeable outcomes due to complex data environments could support a mistake-based defense.

[43] Julie E. Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* 122–24 (Oxford Univ. Press 2019).

6.4 Consent: Validating Conduct Through Pre-Authorized Risk Acceptance

Consent operates as a defense in criminal law by negating the wrongfulness of conduct that would otherwise be unlawful. While not universally applicable certain harms such as grievous bodily injury may not be consented to it plays a central role in medical procedures, sports, and voluntary engagements.

In the AI context, the consent defense arises where the affected party has agreed to the risks or outcomes associated with AI-driven decisions. Examples include:

- A patient agreeing to be diagnosed or treated by an AI-powered system;
- A consumer using a self-adjusting smart device with known data collection features;
- A driver participating in a beta test of semi-autonomous vehicle software.

The challenge lies in determining whether the consent was **informed**, **voluntary**, and **comprehensive**. In many cases, users are unaware of how AI systems function, what data they process, or the extent of autonomous decision-making. Consent secured through vague terms of service or opaque disclosures may not meet legal thresholds for informed consent.[44]

Furthermore, AI decisions may produce externalities harms to third parties who have not consented to the risks. For example, if an AI-enhanced drone used for agricultural surveying crashes onto a public road, injuring passersby, the consent of the operator does not extend to those indirectly affected.

Accordingly, consent may excuse liability only where:

- The affected party is adequately informed of the AI system's function and risks;
- The consent is given freely and without coercion;
- The resulting harm falls within the scope of the anticipated risk.

Developers and providers may limit liability through consent-based user agreements, but such

[44] Danielle Keats Citron & Ryan Calo, *The Automated Administrative State: A Crisis of Legitimacy*, 70 Emory L.J. 797, 832–34 (2021).

arrangements cannot override statutory duties or shield them from gross negligence claims.

6.5 Other Possible Defenses: Automatism and Duress

While less frequently invoked, other defenses may become relevant in AI contexts.

6.5.1 Automatism

In human criminal law, automatism refers to involuntary conduct performed without consciousness or control such as sleepwalking or a seizure. Some have analogized this defense to AI malfunctions, suggesting that when an AI system "glitches" or behaves erratically due to system failure, its conduct should be treated as non-volitional.[45]

However, since AI lacks agency to begin with, the analogy is strained. Automatism may better serve as a conceptual device to excuse human actors from liability where harm resulted from technical failure completely outside their knowledge or control, akin to a driver suffering a heart attack.

6.5.2 Duress

Duress excuses criminal conduct compelled by an immediate threat of serious harm. While unlikely to apply directly to AI, it may be invoked by human actors forced to use or deploy AI systems in constrained environments e.g., wartime contexts or state-mandated surveillance.[46]

For example, a government contractor operating an AI surveillance system under orders from an authoritarian regime may argue duress if prosecuted for unlawful monitoring. The success of such a defense would depend on the imminence of the threat, lack of reasonable alternatives, and proportionality of the harm caused.

6.6 Reframing Defenses in a Technological Age

As AI becomes more prevalent in everyday decision-making, the law must develop a nuanced understanding of criminal defenses in this context. The application of necessity, mistake, and consent must take into account the **non-human** character of AI systems, the **diffused**

[45] Joshua Dressler, *Understanding Criminal Law*, supra note 35, at 249–51.
[46] Id. at 258–59.

responsibility across multiple stakeholders, and the **technical opacity** of machine learning processes.

Moreover, courts must differentiate between the exculpation of human actors and the mere exoneration of technological unpredictability. A defense that is doctrinally sound in human cases may lose coherence when applied to complex socio-technical systems unless carefully contextualized.[47]

Legislative clarification may be required to specify the conditions under which reliance on AI may constitute a valid defense, particularly in high-risk industries such as health care, transport, defense, and finance.

6.7 Conclusion: Preserving Justification and Excuse in the Age of Algorithms

General defenses in criminal law are integral to upholding fairness, proportionality, and justice. As AI systems become increasingly embedded in the mechanisms of decision-making, these defenses must evolve to reflect new realities without eroding their foundational purpose.

By recognizing how necessity, mistake, and consent interact with technological systems, the legal framework can continue to safeguard innocent actors from unjust punishment while ensuring that culpability is neither arbitrarily imposed nor unjustly avoided.

The concluding chapter will now synthesize the arguments advanced throughout the study and propose a comprehensive framework for addressing the criminal liability of AI systems balancing technological innovation with legal responsibility in an era of intelligent machines.

Chapter Seven: Conclusion and Recommendations

7.1 Revisiting the Central Dilemma

The accelerating integration of artificial intelligence (AI) into critical sectors of society ranging from transportation and healthcare to finance, surveillance, and legal adjudication has precipitated a profound challenge for the criminal justice system: how to ascribe criminal responsibility when harm arises from machine action. The traditional framework of criminal

[47] Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence*, supra note 14, at 16–17.

liability, anchored in the dichotomy of *actus reus* and *mens rea*, presumes a human agent with volition, moral awareness, and legal subjectivity. AI systems, which lack consciousness, intention, and emotions, strain this framework at every seam.[48]

Throughout this study, we have examined the viability of attributing criminal liability to AI systems either directly as electronic legal persons or indirectly, by extending responsibility to human actors such as developers, users, and corporate entities. We have analyse d multiple models of liability, assessed the adaptability of classical doctrines such as agency and foreseeability, and interrogated the doctrinal coherence of assigning liability in light of emerging technological realities.

The consistent theme across all chapters is clear: existing legal doctrines are neither entirely obsolete nor entirely adequate. They provide foundational tools, but these tools must be adapted, supplemented, or reimagined if the law is to maintain its integrity and efficacy in the AI era.

7.2 Summary of Key Findings

7.2.1 Legal Personality as a Regulatory Construct

AI systems do not require moral agency to be recognized as bearers of limited legal personality.[49] Analogous to corporate entities, functional legal personhood may be conferred upon certain categories of AI systems not to bestow rights or dignity, but to serve the regulatory goals of accountability, deterrence, and victim compensation.[50] However, this recognition must be carefully bounded and not deployed to shield human actors from liability.[51]

7.2.2 Criminal Liability Must Be Human-Centric, But Not Human-Exclusive

While AI cannot form mens rea in the traditional sense, human involvement in the design, deployment, and oversight of AI systems provides sufficient touchpoints for the imposition of liability.[52] Liability models must accommodate degrees of foreseeability, control, and

[48] Brent Garland & Paul W. Glimcher, *Cognitive Neuroscience and the Law*, 359 Phil. Trans. R. Soc. Lond. B 1697, 1701–03 (2004).
[49] See Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts*, 59 (Springer 2013).
[50] European Parliament Resolution of 16 February 2017 on Civil Law Rules on Robotics, 2015/2103(INL), § 59.
[51] Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 Ohio St. L.J. 1217, 1221–23 (2017).
[52] Gabriel Hallevy, *When Robots Kill: Artificial Intelligence Under Criminal Law*, 3–5 (N.H. Beck 2013).

negligence. In edge cases involving emergent or unexpected behaviour, shared responsibility or composite liability frameworks offer a viable approach.

7.2.3 Classical Elements Must Be Reconceptualized, Not Discarded

The *actus reus* of an AI system may be satisfied by its autonomous conduct when it produces a prohibited result. *Mens rea* can be functionally reconstructed through the lens of constructive knowledge, reasonable foreseeability, or gross negligence.[53] Courts must shift from subjective mental states to risk-based assessments appropriate for socio-technical systems.

7.2.4 General Defenses Remain Relevant But Require Refinement

Doctrines of necessity, mistake, and consent can still serve a useful role in excusing human actors involved in AI-related offenses. However, their application must be tethered to the technological context: the foreseeability of AI error, the transparency of its decision-making, and the informed nature of user consent. As such, defenses should be reframed with attention to the epistemic and operational limitations of AI.[54]

7.3 Recommendations for Legal Reform

To bring coherence and justice to the attribution of criminal liability in the AI age, this study proposes the following legal reforms and institutional strategies:

7.3.1 Establish a Distinct Legal Category for Autonomous Systems

Legislatures should introduce a sui generis classification for AI entities based on levels of autonomy, functional capacity, and societal risk.[55] This category would not confer full legal personhood but would provide a legal scaffold for allocating responsibility, requiring registration, and imposing compliance obligations.

[53] Andreas Matthias, The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata, 6 Ethics & Info. Tech. 175, 178–80 (2004).
[54] Ryan Calo, Robots and Privacy, in *Robot Ethics: The Ethical and Social Implications of Robotics* 187, 194 (Patrick Lin et al. eds., MIT Press 2012).
[55] Nathalie Nevejans, European Civil Law Rules in Robotics, Eur. Parliament Directorate Gen. for Internal Policies (2016), at 10–12.

7.3.2 Implement Tiered Liability Models Based on Function and Risk

Liability should be assigned according to the nature of the AI's function and the foreseeability of harm. Low-risk systems may be governed by negligence standards, while high-risk autonomous systems should be subject to strict or enterprise liability. This will incentivize safety-conscious design and foster responsible innovation.[56]

7.3.3 Mandate Explainability and Auditability Standards

To ensure effective legal scrutiny, AI systems particularly those used in critical infrastructure, healthcare, and public decision-making must be auditable. Regulators should mandate transparency protocols such as algorithmic documentation, data provenance tracking, and explainable decision outputs to support forensic and legal review.

7.3.4 Encourage Cross-Disciplinary Legal Education and Judicial Training

Judges, prosecutors, and defense lawyers must be equipped with the technical literacy necessary to adjudicate AI-related cases. Law schools and judicial academies should integrate modules on algorithmic accountability, machine learning ethics, and digital evidence.[57]

7.3.5 Create a Centralized Regulatory Authority for AI Accountability

An independent oversight body should be established to develop best practices, assess systemic risks, and coordinate between technologists, ethicists, and legal professionals. This authority would act as a clearinghouse for legal standards, model codes, and redress mechanisms.[58]

7.4 Final Reflections: Balancing Justice and Innovation

As we stand at the threshold of the Fourth Industrial Revolution, law must assert itself not as a passive observer, but as an active shaper of ethical and technological futures. The criminal justice system cannot afford to lag behind innovation; nor should it sacrifice its normative

[56] Douglas Husak, *Philosophy of Criminal Law*, supra note 38, at 117–18.
[57] Mireille Hildebrandt, Law as Computation in the Era of Artificial Legal Intelligence, 33 J.L. & Pol'y Info. Soc'y 1, 25–28 (2018).
[58] Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine Learning Era, 105 Geo. L.J. 1147, 1196–98 (2017).

foundations at the altar of expediency.[59]

Accountability must remain central. If AI systems cause harm, society must be assured that responsibility will be assigned, victims will be compensated, and systemic failures will be addressed.[60]At the same time, criminal law must exercise restraint avoiding the punitive reflex in favor of nuanced, risk-sensitive, and forward-looking governance.[61]

Ultimately, the question of AI criminal liability is not merely about fault and punishment it is about preserving the rule of law in a world where the line between human and machine action is increasingly blurred.[62] The challenge is not to retrofit old doctrines into new realities, but to cultivate a legal imagination equal to the complexity of the age.

[59] Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine Learning Era*, 105 Geo. L.J. 1147, 1195–97 (2017).
[60] Brent Mittelstadt et al., The Ethics of Algorithms: Mapping the Debate, 3 Big Data & Soc'y 1, 5–6 (2016).
[61] Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence*, supra note 14, at 26.
[62] Jack M. Balkin, The Three Laws of Robotics in the Age of Big Data, 78 Ohio St. L.J. 1217, 1221–23 (2017).