Indonesian Journal *of* Digital Crime *and* Criminal Justice
Vol. 1 No. 1. January 2025. pp. 1-44
E-ISSN: 3164-0567 | P-ISSN: 3163-8368
Original Research Article

Artificial Intelligence and Criminal Liability: Reassessing Mens Rea in the Age of Autonomous Systems

Subkhi Amhadi Muhammad Nazhir [1]*, Anton Sandi Utama [2]
[1] Universiti Teknologi MARA, Shah Alam, Malaysia
[2] Universitas Pasundan, Bandung, Indonesia
* Corresponding author: subkhinazhir@uitm.edu.my

ABSTRACT

The rapid development of artificial intelligence (AI) and autonomous systems challenges traditional doctrines of criminal liability, particularly the concept of *mens rea*. Existing criminal law frameworks are primarily designed for human actors, assuming intentionality, knowledge, or negligence as mental elements of crime. However, AI systems increasingly operate with varying degrees of autonomy, raising complex questions about attribution of criminal responsibility when harm occurs. This article critically examines whether current criminal law doctrines are sufficient to address offences involving AI-driven decision-making or whether doctrinal reform is necessary. Using doctrinal legal research and comparative analysis of selected jurisdictions, this study explores three potential models of liability: operator liability, developer liability, and autonomous system accountability. The findings indicate significant conceptual gaps in attributing intent in AI-related harm, particularly where machine learning systems produce unpredictable outcomes. The study argues that strict reliance on traditional *mens rea* may lead to accountability deficits or overextension of human liability. It proposes a recalibrated framework combining strict liability for high-risk AI deployment and enhanced regulatory oversight. The article concludes that criminal law must evolve toward a hybrid accountability model to ensure justice in the era of autonomous technologies.

KEYWORDS: Artificial Intelligence, Mens Rea, Criminal Liability, Autonomous Systems, Digital Crime, AI Governance

INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) has transformed the way decisions are generated and implemented across numerous sectors of society.

Available online at https://journal.criminallawinstitute.org/JDCCJ/article/view/16, since January 15, 2026
Submitted: November 11, 2024 ■ Reviewed: November 27, 2024 ■ Revised: January 5, 2026 ■ Accepted: January 15, 2025

Unlike traditional software systems that merely execute pre-programmed instructions, contemporary AI systems increasingly possess the capacity to learn from data, adapt to changing environments, and make decisions with varying degrees of autonomy. Recent developments in machine learning, deep learning, and generative AI have accelerated the deployment of autonomous systems in transportation, healthcare, financial services, public administration, and security operations. These technological innovations promise substantial benefits in terms of efficiency, accuracy, and productivity; however, they also introduce unprecedented legal and ethical challenges, particularly when autonomous systems cause harm or contribute to criminal conduct (Russell and Norvig 2021).

The growing integration of AI into critical infrastructures has intensified concerns regarding accountability and responsibility. Autonomous vehicles, for example, are capable of making real-time navigational decisions without direct human intervention, while AI-assisted medical systems increasingly participate in diagnostic and treatment recommendations. In the financial sector, algorithmic trading systems execute transactions at speeds beyond human capability, and predictive analytics tools influence lending decisions and fraud detection mechanisms. Similarly, law enforcement agencies employ AI technologies for surveillance, facial recognition, and predictive policing. While these applications offer significant societal advantages, they also create risks of accidents, discrimination, privacy violations, financial losses, and other forms of harm that may trigger legal consequences (Calo 2015; Sartor and Lagioia 2020).

The potential harms generated by autonomous systems differ fundamentally from those associated with conventional technologies because AI systems may act in ways that were neither specifically programmed nor fully anticipated by their developers. Machine-learning algorithms frequently produce outputs through complex computational processes that remain difficult to explain, even to their creators. This phenomenon, commonly referred to as the "black box" problem, complicates efforts to determine causation, foreseeability, and accountability when harmful outcomes occur (Burrell 2016). Consequently, incidents involving autonomous systems raise

important questions regarding whether existing legal frameworks are capable of assigning criminal responsibility in situations where human control over decision-making is significantly diminished.

Criminal law has traditionally been constructed upon the assumption that criminal acts are committed by human actors possessing consciousness, intention, and moral agency. The fundamental principle of criminal liability generally requires the coexistence of two elements: *actus reus* (the prohibited act) and *mens rea* (the guilty mind). Mens rea functions as a cornerstone of criminal responsibility because it reflects the moral blameworthiness of an offender. Depending on the legal system, mens rea may encompass intention, knowledge, recklessness, or negligence, each requiring some degree of cognitive awareness or volitional control over the prohibited conduct (Simester, Spencer, and Sullivan 2019). The emergence of AI challenges these assumptions because autonomous systems can influence or execute actions without possessing consciousness, moral understanding, or legal personhood.

The difficulty becomes particularly evident when harmful conduct is performed by an AI system acting with substantial autonomy. If an autonomous vehicle causes a fatal collision after independently processing environmental data, it becomes difficult to determine whose mental state should be examined for purposes of criminal liability. The programmer may not have intended the harmful outcome, the manufacturer may have exercised reasonable care, and the user may have lacked effective control over the system at the relevant moment. Since the AI itself lacks legal and moral capacity, conventional doctrines of mens rea encounter significant limitations when applied to such circumstances (Pagallo 2013). As a result, criminal law faces the challenge of addressing situations in which harmful outcomes occur without a clearly identifiable human actor possessing the requisite guilty mind.

This challenge has generated what scholars frequently describe as the "accountability gap." The concept refers to situations in which responsibility for harmful outcomes cannot be satisfactorily attributed to any individual or entity under existing legal doctrines, despite the occurrence of significant

harm (Matthias 2004). Accountability gaps are particularly problematic in criminal law because criminal sanctions are designed not only to punish wrongful conduct but also to express societal condemnation and deter future wrongdoing. If autonomous systems can produce harmful consequences without generating identifiable criminal responsibility, the legitimacy and effectiveness of criminal justice systems may be undermined. Consequently, there is increasing scholarly debate concerning whether existing legal principles remain adequate for addressing criminal conduct involving AI.

Against this background, a central question concerns the continuing relevance of traditional mens rea doctrine in the age of autonomous systems. Some scholars argue that existing criminal law frameworks remain sufficiently flexible to attribute liability to programmers, manufacturers, operators, or corporate entities whose actions contribute to AI-related harms (Hallevy 2015). Others contend that the unique characteristics of autonomous systems require substantial doctrinal reform because traditional concepts of intention, knowledge, and recklessness were developed for human decision-makers rather than algorithmic agents (Abbott 2020). This debate highlights the need to critically assess whether mens rea, as currently understood, can adequately accommodate emerging forms of technologically mediated conduct.

A second issue concerns the identification of appropriate subjects of criminal responsibility. Various attribution models have been proposed in contemporary legal scholarship. These include direct liability for human operators, negligence-based liability for developers, corporate criminal liability for organizations deploying AI systems, and hybrid approaches that distribute responsibility among multiple actors. More controversial proposals have suggested limited forms of electronic personhood or AI-specific legal status, although such proposals remain highly contested due to concerns regarding moral agency, punishment, and legal legitimacy (European Parliament 2017; Turner 2019). Determining the most appropriate model requires careful consideration of both doctrinal coherence and practical enforceability.

A third issue relates to the development of future criminal liability

frameworks capable of responding effectively to increasingly autonomous technologies. Existing legal mechanisms may require adaptation to accommodate algorithmic decision-making processes, shared human-machine agency, and complex causal chains. Reform proposals have included expanded negligence standards, strict liability regimes for high-risk AI applications, organizational liability models, and new regulatory approaches emphasizing ex ante risk management and accountability by design (Sartor and Lagioia 2020). Evaluating these alternatives is essential for ensuring that criminal law remains capable of protecting societal interests while simultaneously encouraging technological innovation.

This study therefore seeks to address three interrelated research questions: first, whether the traditional concept of mens rea remains adequate for criminal cases involving AI systems; second, who should bear criminal responsibility when autonomous systems contribute to criminal conduct or harmful outcomes; and third, which model of criminal liability is most appropriate for governing autonomous systems. To answer these questions, the research aims to analyze the adequacy of contemporary mens rea doctrine, examine competing models of criminal responsibility attribution, and formulate recommendations for the reform of criminal law in response to technological developments. The significance of this study lies both in its theoretical contribution to ongoing debates concerning criminal responsibility and legal personhood, and in its practical relevance for legislators, regulators, and policymakers seeking to develop effective legal frameworks for the governance of artificial intelligence.

CONCEPTUAL AND THEORETICAL FRAMEWORK

Criminal liability constitutes one of the foundational concepts of modern criminal law. It refers to the legal mechanism through which an individual or legal entity is held accountable for conduct prohibited by law and subjected to penal sanctions. The doctrine of criminal liability serves not merely as a procedural device for punishment but as a normative framework that distinguishes blameworthy conduct from morally neutral behavior. Across

both common law and civil law traditions, criminal liability is generally predicated upon the existence of specific legal elements that collectively justify the imposition of punishment (Duff 2007).

The first essential element of criminal liability is *actus reus*, commonly understood as the external or physical component of a criminal offense. Actus reus encompasses conduct, circumstances, and consequences that constitute a legally prohibited act. Criminal law traditionally requires that a defendant perform a voluntary act or omission that causes a prohibited result. The emphasis on voluntary conduct reflects the broader principle that punishment should be imposed only when individuals exercise some degree of control over their actions. Consequently, involuntary movements, reflexive behavior, or purely accidental occurrences generally fall outside the scope of criminal responsibility (Ashworth 2021).

A second and equally important element is *mens rea*, which refers to the mental state accompanying the prohibited conduct. Mens rea functions as the subjective component of criminal liability because it evaluates the defendant's state of mind at the time of the offense. The requirement of mens rea reflects the moral intuition that criminal punishment should ordinarily be reserved for individuals who act with a culpable mental attitude. Without proof of the required mental state, criminal liability may be difficult to justify, particularly for serious offenses carrying significant penalties (Herring 2022).

In addition to actus reus and mens rea, criminal liability often requires proof of causation. Causation establishes the necessary connection between the defendant's conduct and the resulting harm. Legal systems typically distinguish between factual causation and legal causation. Factual causation asks whether the harm would have occurred "but for" the defendant's actions, whereas legal causation examines whether the resulting harm was sufficiently foreseeable and closely connected to the conduct to justify liability. The concept becomes particularly important in technologically complex environments where multiple actors and systems contribute to a harmful outcome (Ormerod and Laird 2023).

The culmination of these elements produces the broader concept of criminal responsibility. Criminal responsibility concerns the conditions under

which an individual may properly be regarded as answerable for criminal wrongdoing. It incorporates considerations of culpability, autonomy, capacity, and fairness. A person may satisfy the objective elements of an offense yet remain exempt from responsibility due to defenses such as insanity, duress, or lack of capacity. Thus, criminal responsibility represents a normative judgment about whether punishment is justified under the circumstances (Duff 2018).

Within traditional criminal law doctrine, mens rea is generally categorized into several distinct forms of culpability. The highest degree of culpability is intention. A person acts intentionally when he or she seeks to bring about a particular consequence or acts with the purpose of achieving a prohibited result. Intention occupies a central place in criminal law because it reflects the strongest degree of alignment between the actor's will and the harmful outcome. Consequently, intentional crimes often attract the most severe sanctions (Smith, Hogan, and Ormerod 2018).

A second category is knowledge. Knowledge exists when an individual is aware that particular circumstances exist or that certain consequences are practically certain to result from the conduct undertaken. Unlike intention, knowledge does not require a desire to bring about the prohibited outcome. Rather, liability arises because the actor consciously proceeds despite awareness of relevant facts. Criminal law frequently treats knowledge as a sufficiently culpable mental state because it demonstrates conscious engagement with risk and wrongdoing (Alexander and Ferzan 2009).

The third category is recklessness. Recklessness generally involves conscious risk-taking. A reckless actor recognizes a substantial and unjustifiable risk that harm may occur but nevertheless chooses to proceed. Modern criminal law often regards recklessness as a significant form of culpability because it reflects a deliberate disregard for legally protected interests. Although less blameworthy than intention, recklessness still demonstrates conscious indifference toward the possibility of harmful consequences (Robinson 1997).

Negligence occupies the lowest tier of traditional culpability. Unlike intention, knowledge, or recklessness, negligence does not require subjective

awareness of risk. Instead, liability is based on a failure to perceive a risk that a reasonable person would have recognized under similar circumstances. Criminal negligence therefore incorporates an objective standard of conduct. The doctrine seeks to ensure that individuals exercise appropriate care when engaging in activities capable of causing harm to others (Moore and Hurd 2011).

The intellectual foundations of criminal responsibility are deeply rooted in philosophical theories of punishment. One of the most influential perspectives is retributive theory. Retributivism maintains that punishment is justified because offenders deserve it in proportion to their wrongdoing. According to this view, punishment is fundamentally backward-looking, focusing on the moral culpability of the offender rather than on future social benefits. Criminal sanctions are legitimate because they express society's condemnation of wrongful conduct and restore moral balance disrupted by the offense (von Hirsch 1993).

Retributive theory places particular emphasis on individual autonomy and moral agency. Since punishment is justified by desert, liability can only be imposed upon actors capable of making meaningful choices. This requirement has significant implications for discussions concerning artificial intelligence. If criminal responsibility presupposes autonomous moral agency, questions inevitably arise regarding whether AI systems can ever become appropriate subjects of criminal punishment. Most contemporary scholars remain skeptical because current AI systems lack consciousness, moral understanding, and the capacity for genuine choice (Gardner 2007).

In contrast, utilitarian theories justify punishment on the basis of its beneficial social consequences. Classical utilitarian thinkers argue that criminal sanctions are legitimate when they prevent future harms through deterrence, incapacitation, rehabilitation, or social protection. Under this framework, the primary objective of criminal law is not moral retribution but the promotion of collective welfare. The effectiveness of punishment therefore depends upon its capacity to influence future behavior rather than its correspondence to moral desert (Bentham 1789/1996).

Despite their differences, both retributive and utilitarian approaches

converge on the importance of moral blameworthiness. Moral blameworthiness functions as a central criterion for distinguishing conduct deserving criminal sanction from conduct that merely produces unfortunate outcomes. Criminal law traditionally assumes that blameworthiness arises from conscious choices made by rational actors. This assumption becomes increasingly problematic in contexts involving autonomous technologies, where harmful outcomes may emerge from algorithmic processes that lack human-like cognition or intention (Cane 2002).

The emergence of artificial intelligence has therefore generated significant challenges for traditional theories of criminal responsibility. AI is generally defined as the capacity of computational systems to perform tasks that ordinarily require human intelligence, including learning, reasoning, perception, prediction, and decision-making. Contemporary AI systems differ from conventional software because they can improve performance through experience rather than relying solely on predetermined instructions (Kaplan and Haenlein 2019).

A major driver of modern AI development is machine learning. Machine learning refers to computational methods that enable systems to identify patterns within large datasets and generate predictive models. Rather than being explicitly programmed for every possible scenario, machine-learning systems derive behavioral rules from statistical analysis of training data. This characteristic enhances adaptability but also reduces predictability, creating challenges for legal systems that depend upon identifying foreseeable consequences and responsible actors (Alpaydin 2021).

Deep learning represents a more advanced subset of machine learning that employs artificial neural networks with multiple computational layers. These systems are capable of processing highly complex information, including images, language, and behavioral data. Deep learning has achieved remarkable success in areas such as autonomous driving, facial recognition, and medical diagnostics. However, the complexity of neural-network architectures often makes it difficult to understand how specific decisions are generated, thereby complicating legal assessments of causation and responsibility (Goodfellow, Bengio, and Courville 2016).

The concept of autonomous decision-making further intensifies these concerns. Autonomous systems can collect environmental information, evaluate alternatives, and execute actions without immediate human intervention. Although humans remain involved in the design, deployment, and supervision of such systems, real-time decisions may occur independently of direct human control. As autonomy increases, the causal relationship between human intentions and machine actions becomes progressively more attenuated, raising difficult questions regarding criminal attribution (Bostrom 2014).

These developments have given rise to what legal scholars describe as the accountability gap. Accountability gap theory refers to situations in which harmful outcomes occur, yet existing legal doctrines fail to identify a sufficiently culpable actor. The problem emerges because autonomous systems can contribute to decisions and actions that were neither intended nor fully foreseeable by any individual participant in the technological ecosystem. Consequently, responsibility becomes dispersed among programmers, developers, operators, manufacturers, and users, none of whom may satisfy traditional requirements of mens rea (Coeckelbergh 2020).

The accountability gap is closely associated with two interrelated challenges: the human control problem and the explainability problem. The human control problem arises when human oversight becomes too remote to justify direct attribution of responsibility, while the explainability problem arises when the internal reasoning processes of AI systems cannot be adequately understood or reconstructed. Together, these challenges expose significant limitations within conventional criminal law doctrines and underscore the need for new theoretical approaches capable of reconciling technological autonomy with principles of legal accountability. As autonomous systems continue to evolve, resolving these tensions will become increasingly important for preserving the legitimacy and effectiveness of criminal law.

METHODS

This study employs doctrinal legal research, a method that focuses on the systematic analysis of legal norms, principles, doctrines, and regulatory frameworks governing criminal liability in the context of artificial intelligence (AI). Doctrinal legal research is particularly appropriate because the study seeks to evaluate the adequacy of existing criminal law doctrines rather than to investigate empirical patterns of behavior. The analysis concentrates on legal rules concerning *mens rea*, causation, criminal responsibility, and corporate criminal liability, as well as emerging regulatory principles applicable to autonomous systems. Through this approach, the research examines whether traditional criminal law doctrines remain capable of addressing harms caused by AI-enabled decision-making systems (Hutchinson and Duncan 2012; McConville and Chui 2017).

The research adopts three complementary approaches: a statutory approach, a conceptual approach, and a comparative approach. The statutory approach examines relevant criminal law provisions governing intention, knowledge, recklessness, negligence, causation, and liability attribution, together with contemporary regulatory frameworks addressing AI governance. Particular attention is given to legal norms concerning accountability, risk management, and human oversight in emerging AI regulations. The conceptual approach analyzes the legal doctrines and theoretical foundations of criminal responsibility, focusing on the compatibility of traditional *mens rea* requirements with autonomous decision-making systems. In addition, the comparative approach evaluates how different jurisdictions and legal systems respond to accountability challenges arising from AI-related harms, enabling the identification of common principles and potential legal reforms (Michaels 2009; Smits 2017).

The legal materials used in this study consist of both primary and secondary sources. Primary legal materials include criminal statutes, AI-related regulations, judicial decisions, and other authoritative legal instruments relevant to criminal accountability. Secondary legal materials comprise scholarly books, peer-reviewed journal articles, legal commentaries,

and reports issued by international organizations and regulatory institutions. These materials are analyzed using qualitative legal analysis and comparative legal reasoning. The analytical process involves interpreting legal norms, assessing doctrinal coherence, identifying normative gaps, and evaluating competing models of criminal liability attribution. Through this method, the study seeks to formulate a coherent legal framework capable of addressing questions of criminal responsibility in the age of autonomous AI systems (Zweigert and Kötz 1998; Taekema 2018).

THE CHALLENGE OF MENS REA IN AI-DRIVEN HARM

The doctrine of *mens rea* occupies a central position in modern criminal law because it serves as the primary basis for attributing moral and legal blameworthiness to an offender. Criminal liability is generally not imposed solely because harm has occurred; rather, the law requires proof that the accused possessed a culpable mental state at the time of the prohibited conduct. The requirement of mens rea reflects the principle that punishment is justified only when an individual consciously chooses to engage in wrongful behavior or fails to exercise the degree of care expected by law (Ashworth 2021). Consequently, criminal law distinguishes between intentional wrongdoing, reckless conduct, negligent behavior, and purely accidental events.

The importance of mens rea extends beyond technical legal doctrine. It is closely linked to the broader concept of moral culpability. Criminal punishment carries social condemnation and therefore presupposes that the offender is morally responsible for the prohibited act. A person who intentionally commits fraud or knowingly causes harm is generally considered more blameworthy than someone who produces the same result through inadvertence or mistake. This relationship between culpability and punishment has long been regarded as a fundamental principle of criminal justice systems across jurisdictions (Duff 2007). Without a culpable mental state, the moral legitimacy of criminal punishment becomes difficult to justify.

The emergence of artificial intelligence challenges these traditional assumptions because AI systems are capable of producing harmful outcomes without possessing the psychological characteristics that criminal law associates with culpability. Unlike human beings, AI systems do not possess consciousness, emotions, desires, or subjective awareness. Their actions are generated through computational processes rather than intentional choices. As a result, the legal concepts traditionally used to evaluate human conduct do not easily apply to autonomous technologies (Turner 2019).

One of the most significant obstacles concerns the absence of intention. In criminal law, intention generally requires a conscious objective to bring about a particular result. AI systems may generate outputs that resemble intentional behavior, yet such outputs are not accompanied by genuine mental states. An autonomous vehicle may change direction to avoid an obstacle, and an algorithmic trading system may execute thousands of transactions to maximize profit, but neither system forms intentions in the legal or philosophical sense. Their behavior results from programmed objectives, data processing, and optimization functions rather than conscious decision-making (Abbott 2020).

Similarly, AI lacks moral agency. Moral agency refers to the capacity to understand the moral significance of one's actions and to make choices based upon ethical considerations. Criminal responsibility traditionally assumes that offenders are capable of distinguishing right from wrong and acting accordingly. Contemporary AI systems do not possess such capacities. Although advanced algorithms may simulate ethical reasoning or follow predefined rules, they do not genuinely comprehend moral obligations. Consequently, attributing criminal blame directly to AI would conflict with the foundational assumptions underlying criminal liability (Coeckelbergh 2020).

A further complication arises from the fact that AI systems lack legal personality. Criminal law generally recognizes natural persons and, in some jurisdictions, legal entities such as corporations as subjects of criminal responsibility. These entities can be punished because they possess recognized legal status. AI systems, however, remain legal objects rather than

legal subjects. They may be owned, controlled, deployed, or regulated, but they do not possess rights, duties, or liabilities independent of human actors. As a result, criminal liability must ultimately be attributed to individuals or organizations connected to the system rather than to the AI itself (Pagallo 2013).

The challenge becomes even more complex when machine-learning systems are involved. Traditional software operates according to explicit instructions written by programmers, making its behavior relatively predictable. Machine-learning systems, by contrast, derive patterns and decision rules from data. As these systems become increasingly sophisticated, developers may not fully understand how particular outputs are generated. The gap between design and outcome creates difficulties when attempting to identify the mental state of any specific human actor associated with the system (Alpaydin 2021).

One manifestation of this difficulty is the so-called black-box problem. Many contemporary AI systems, particularly those based on deep learning architectures, produce results through highly complex computational processes that cannot easily be interpreted by humans. Even when a harmful outcome occurs, it may be impossible to reconstruct the precise reasoning process that led to the decision. This lack of transparency complicates efforts to establish causation, foreseeability, and culpability, all of which are essential components of criminal liability (Burrell 2016).

Another challenge stems from emergent behavior. Emergent behavior occurs when a system exhibits actions or outcomes that were not explicitly anticipated by its developers. Such behavior may arise from interactions between algorithms, datasets, environmental conditions, and user inputs. Although the underlying system operates according to computational rules, its resulting conduct may differ substantially from the expectations of designers and operators. This unpredictability makes it difficult to determine whether any individual possessed the requisite mens rea regarding the eventual harm (Bostrom 2014).

The issue is particularly evident in self-learning systems. Unlike traditional technologies that remain relatively static after deployment, self-

learning AI systems continue to adapt based on new information and experiences. Over time, the behavior of these systems may diverge from their original programming. As a result, harmful outcomes may emerge long after development, making it increasingly difficult to attribute criminal intent, knowledge, or recklessness to programmers, manufacturers, or users. The capacity for continuous adaptation therefore challenges conventional assumptions regarding foreseeability and control (Goodfellow, Bengio, and Courville 2016).

Autonomous vehicle accidents provide a useful illustration of these difficulties. Self-driving vehicles are designed to process environmental information and make rapid decisions in real time. During unavoidable collision scenarios, an autonomous vehicle may be required to choose among competing risks, such as protecting passengers, pedestrians, or other road users. If a fatal accident occurs, determining criminal liability becomes highly complex. The vehicle itself cannot possess mens rea, while programmers and manufacturers may not have intended or foreseen the specific outcome. Furthermore, the human occupant may have exercised little or no control over the relevant decision-making process (Marchant and Lindor 2012).

AI-based financial fraud presents another example. Algorithmic trading systems can execute transactions at extraordinary speed and scale. In certain circumstances, these systems may engage in market manipulation, exploit regulatory loopholes, or contribute to financial instability. While the resulting conduct may resemble intentional fraud, proving criminal intent on the part of developers, operators, or corporate executives can be challenging. The complexity of financial algorithms often obscures the relationship between human decision-making and harmful outcomes, thereby complicating the attribution of mens rea (Bathaee 2018).

The challenges become even more pronounced in the context of cybercrime. Autonomous malware can independently identify vulnerabilities, propagate through networks, and adapt to defensive measures without continuous human intervention. Similarly, AI-assisted hacking tools can automate reconnaissance, password cracking, phishing campaigns, and exploitation activities. If such systems produce unforeseen

consequences or evolve beyond their original parameters, determining the criminal responsibility of developers or users becomes increasingly difficult. The autonomy of the technology may obscure the connection between human intent and criminal conduct (Brenner 2010).

These examples demonstrate that the primary legal challenge is not merely the occurrence of harm but the identification of a guilty mind. Criminal law traditionally seeks to establish who intended, knew, foresaw, or disregarded the risk of the prohibited outcome. In AI-related cases, however, harmful consequences may result from interactions between multiple actors, datasets, algorithms, and environmental factors. Responsibility becomes fragmented across developers, manufacturers, operators, users, and organizations, none of whom may individually satisfy the conventional requirements of mens rea (Hallevy 2015).

This phenomenon has been described as the "mens rea gap." The mens rea gap arises when criminal harm occurs, yet no clearly identifiable human actor possesses the mental state required for criminal liability. Closely related to this problem is the difficulty of proving foreseeability. Traditional doctrines of recklessness and negligence depend upon demonstrating that a risk was known or reasonably foreseeable. In autonomous and self-learning systems, however, harmful outcomes may emerge from complex and evolving computational processes that exceed human expectations. Consequently, the gap between technological autonomy and human culpability challenges one of the most fundamental assumptions of criminal law: that every punishable harm can ultimately be traced to a blameworthy human mind. Addressing this challenge will require a reconsideration of traditional liability doctrines and the development of legal frameworks capable of accommodating increasingly autonomous forms of decision-making (Matthias 2004; Abbott 2020).

COMPARATIVE ANALYSIS OF CRIMINAL LIABILITY APPROACHES

The emergence of artificial intelligence has prompted jurisdictions worldwide to reconsider traditional approaches to criminal liability and accountability.

Although no major legal system currently recognizes AI as an independent subject of criminal responsibility, significant differences exist regarding how liability should be attributed when autonomous systems contribute to harmful outcomes. These differences are shaped by each jurisdiction's criminal law traditions, regulatory philosophies, and approaches to technological governance. A comparative analysis demonstrates that while legal systems continue to rely primarily on human responsibility, increasing attention is being devoted to organizational accountability, risk management, and human oversight.

United States

The United States continues to address AI-related harms primarily through existing criminal law doctrines rather than through AI-specific criminal legislation. American criminal law remains grounded in the traditional requirements of *actus reus* and *mens rea*, meaning that criminal liability generally attaches only to natural persons or legal entities capable of satisfying recognized standards of culpability. Consequently, when an AI system causes harm, legal analysis focuses on the conduct and mental states of developers, operators, users, or corporations involved in the deployment of the technology (Dressler 2022; Kerr 2023).

A distinctive feature of the U.S. legal system is its expansive doctrine of corporate criminal liability. Under the principle of *respondeat superior*, corporations may be held criminally liable for criminal acts committed by employees acting within the scope of employment and for the benefit of the corporation. This doctrine provides a potential mechanism for addressing AI-related harms because liability may be attributed to organizations responsible for the design, implementation, or commercialization of AI systems, even where identifying an individual offender proves difficult (Buell 2016; Coffee 2020).

At the same time, scholars have noted that traditional criminal law doctrines encounter significant difficulties when applied to machine-learning systems. The autonomous and adaptive characteristics of AI may complicate the attribution of intention, knowledge, or recklessness to any particular

actor. Consequently, legal commentators increasingly debate whether existing negligence standards and corporate liability doctrines are sufficient to address accountability gaps created by AI-driven decision-making (Balkin 2015; Abbott 2020).

European Union

The European Union has adopted a markedly different approach by emphasizing preventive regulation and risk management. Rather than relying primarily on ex post liability mechanisms, the EU seeks to reduce potential harms through a comprehensive governance framework that imposes obligations upon AI developers, providers, and deployers before harmful outcomes occur (Veale and Borgesius 2021).

This approach is reflected in the European Union Artificial Intelligence Act (AI Act), which establishes a risk-based regulatory framework for AI systems. The AI Act classifies AI applications according to levels of risk and imposes stricter requirements on systems categorized as high risk. These requirements include transparency obligations, technical documentation, record-keeping mechanisms, conformity assessments, and continuous monitoring throughout the lifecycle of the system (European Parliament and Council 2024).

One of the most important aspects of the AI Act is its emphasis on meaningful human oversight. High-risk AI systems must be designed in a manner that allows human operators to supervise, understand, and intervene in automated decision-making processes when necessary. This requirement reflects the EU's broader objective of preserving human accountability and preventing situations in which harmful outcomes cannot be attributed to identifiable actors (Smuha 2021; Veale and Borgesius 2021).

Comparative Findings

The comparative analysis demonstrates several common trends across jurisdictions. First, none of the examined legal systems currently recognizes AI as an independent bearer of criminal responsibility. Criminal liability continues to be attributed to natural persons and legal entities rather than to

autonomous systems themselves. Second, all jurisdictions emphasize the importance of accountability, although they adopt different mechanisms to achieve that objective. Third, policymakers increasingly recognize that traditional doctrines of individual fault may be insufficient when dealing with highly autonomous technologies (Abbott 2020; Turner 2019).

Nevertheless, important differences remain. The United States relies heavily on traditional criminal law doctrines and corporate liability principles, whereas the European Union has adopted a comprehensive regulatory framework centered on risk management and human oversight. The United Kingdom continues to refine its corporate liability framework, while Singapore, Japan, and China have emphasized governance-oriented approaches that combine regulatory oversight with technological innovation. Despite these differences, a common trend can be observed toward strengthening organizational accountability and ensuring meaningful human control over AI systems (Wells 2020; Yeung, Howes, and Pogrebna 2020).

MODELS OF CRIMINAL LIABILITY FOR AI-RELATED HARM

The increasing autonomy of artificial intelligence systems has generated significant debate regarding the most appropriate model of criminal liability for harms caused or facilitated by AI. Traditional criminal law assumes that harmful conduct can ultimately be attributed to identifiable human actors possessing the requisite *mens rea*. However, as AI systems become more autonomous, adaptive, and capable of independent decision-making, identifying a single responsible actor becomes increasingly difficult. This challenge has prompted legal scholars and policymakers to explore various models of liability attribution designed to address accountability gaps while preserving the fundamental principles of criminal law (Abbott 2020; Hallevy 2015).

Any proposed liability model must satisfy several normative objectives. First, it must ensure that harmful conduct does not escape legal accountability merely because it involves advanced technology. Second, it must maintain fairness by imposing liability only upon actors who possess an appropriate

degree of responsibility for the harmful outcome. Third, it should promote socially beneficial behavior, including safe AI design, responsible deployment, and effective organizational governance. Against this background, four principal models have emerged in contemporary legal scholarship: operator liability, developer liability, corporate liability, and autonomous system accountability (Turner 2019; Pagallo 2013).

Operator Liability Model

The operator liability model attributes criminal responsibility to the individual or entity exercising direct control over the AI system at the time the harmful conduct occurs. Under this approach, the operator is regarded as the primary decision-maker because the deployment and use of AI ultimately remain subject to human choice. The model is consistent with the traditional legal principle that individuals who introduce potentially dangerous technologies into society should bear responsibility for their consequences (Marchant and Lindor 2012).

The primary rationale underlying operator liability is the concept of human control. Even highly autonomous AI systems are typically activated, configured, monitored, or deployed by human users. Consequently, operators are often viewed as the actors best positioned to assess risks, supervise system performance, and intervene when necessary. This reasoning reflects broader regulatory trends emphasizing meaningful human oversight as a prerequisite for responsible AI deployment (Yeung, Howes, and Pogrebna 2020).

One significant advantage of the operator liability model is its compatibility with existing criminal law doctrines. Courts and prosecutors are already familiar with assigning liability to individuals whose actions directly contribute to harmful outcomes. Extending this principle to AI-assisted conduct requires relatively limited doctrinal modification. The model also preserves the traditional relationship between criminal liability and human agency, thereby avoiding controversial questions regarding the legal status of AI systems (Ashworth 2021).

A second advantage is the clarity of attribution. By identifying the

operator as the primary bearer of responsibility, the model reduces uncertainty concerning who should be investigated and prosecuted when harm occurs. This clarity may enhance deterrence because operators are incentivized to exercise caution when deploying potentially risky technologies (Duff 2018).

Despite these strengths, the operator liability model faces important limitations. In highly autonomous environments, operators may possess limited knowledge of how an AI system reaches particular decisions. Self-learning systems can evolve beyond their initial configurations, making harmful outcomes difficult to predict or prevent. Imposing criminal liability under such circumstances risks holding individuals responsible for outcomes they neither intended nor reasonably foresaw (Bathaee 2018).

The model may therefore result in over-criminalization. Operators who exercise reasonable care may nevertheless face liability for harms generated by complex algorithmic processes beyond their practical control. Such outcomes conflict with fundamental principles of culpability that require a meaningful connection between fault and punishment. Consequently, operator liability alone may be insufficient to address accountability challenges associated with advanced AI systems (Matthias 2004).

Developer Liability Model

An alternative approach focuses on developers and designers responsible for creating AI systems. Under the developer liability model, criminal responsibility is attributed to individuals or organizations involved in designing, programming, training, or testing AI technologies that subsequently cause harm. This approach reflects the view that many risks associated with AI originate during the development stage rather than during deployment (Bostrom 2014).

The principal rationale for developer liability is design responsibility. Developers make critical decisions concerning system architecture, training data, safety mechanisms, risk controls, and operational parameters. These design choices significantly influence how an AI system behaves after deployment. If foreseeable risks arise from negligent or reckless design

practices, assigning liability to developers may be justified because the harmful outcome can be traced to deficiencies embedded within the system itself (Sartor and Lagioia 2020).

One of the strongest arguments supporting this model is its capacity to encourage safer innovation. The prospect of legal liability creates incentives for developers to invest in safety testing, transparency mechanisms, cybersecurity protections, and robust risk assessment procedures. By internalizing the costs of harmful outcomes, developer liability may promote higher standards of care throughout the AI development lifecycle (Calo 2015).

Furthermore, the model aligns with established legal principles governing defective products and professional negligence. Just as engineers, manufacturers, and medical professionals may be held accountable for foreseeable harms arising from substandard conduct, AI developers may similarly bear responsibility when negligent design contributes to criminally relevant outcomes (Abbott 2020).

Nevertheless, significant challenges arise in proving causation. AI systems often involve multiple developers, training datasets, software updates, and external inputs. Harmful outcomes may emerge years after deployment and result from interactions that no individual developer could reasonably predict. Establishing a direct causal link between a specific design decision and a particular criminal harm may therefore be extremely difficult (Bathaee 2018).

The complexity of machine-learning systems further complicates attribution. Modern AI frequently generates outputs through adaptive processes that exceed the intentions of its creators. As autonomy increases, the causal relationship between development decisions and operational outcomes becomes increasingly indirect. Consequently, excessive reliance on developer liability may unfairly impose responsibility for harms that arise from factors beyond the developer's effective control (Goodfellow, Bengio, and Courville 2016).

Corporate Liability Model

Many scholars argue that corporate liability provides the most practical framework for addressing AI-related harms. Under this model, responsibility is attributed to organizations that develop, own, deploy, or profit from AI systems rather than to individual actors alone. The model reflects the reality that modern AI technologies are typically created and managed within complex organizational structures involving numerous employees, departments, and decision-making processes (Wells 2020).

The concept of organizational responsibility is particularly relevant because AI-related harms often result from systemic failures rather than isolated individual actions. Decisions regarding resource allocation, safety protocols, risk management, employee training, and compliance oversight are frequently made at the organizational level. Consequently, attributing responsibility solely to individual operators or developers may overlook broader institutional factors contributing to harm (Gobert and Punch 2003).

Corporate liability also addresses compliance failures. Many AI-related risks can be mitigated through appropriate governance structures, internal controls, auditing mechanisms, and safety procedures. Where an organization fails to implement adequate safeguards despite foreseeable risks, criminal liability may serve as a powerful incentive for responsible governance. This approach aligns with contemporary regulatory trends emphasizing organizational accountability and risk management (Coffee 2020).

Another advantage of the corporate model is its compatibility with emerging AI governance frameworks. Regulatory instruments such as the EU AI Act impose obligations on providers and deployers rather than on AI systems themselves. Requirements relating to transparency, human oversight, documentation, and risk assessment are directed primarily toward organizations responsible for AI deployment. Corporate liability therefore complements existing governance mechanisms by reinforcing compliance obligations through potential criminal sanctions (Veale and Borgesius 2021).

However, corporate liability is not without limitations. Critics argue that organizations may absorb penalties as operational costs without significantly

altering behavior. Furthermore, attributing criminal fault to corporations raises complex questions concerning collective intention, organizational culture, and institutional blameworthiness. Despite these concerns, many scholars regard corporate liability as one of the most realistic approaches for addressing harms generated by increasingly complex technological systems (Wells 2020).

Autonomous System Accountability Model

The most controversial proposal involves granting some form of legal accountability directly to autonomous systems themselves. This approach is often associated with debates concerning electronic personhood, which suggest that highly autonomous AI systems may eventually require a distinct legal status comparable to corporate personality (Pagallo 2013).

Proponents argue that electronic personhood could help address accountability gaps created by autonomous decision-making. If AI systems become sufficiently independent to make consequential decisions without meaningful human intervention, attributing responsibility solely to human actors may appear increasingly artificial. Granting limited legal status to AI systems could provide a mechanism for assigning legal obligations, maintaining accountability records, and facilitating compensation for victims (Koops, Hildebrandt, and Jaquet-Chiffelle 2010).

Supporters further contend that legal personhood has historically been extended to entities lacking human characteristics, including corporations, foundations, and public institutions. From this perspective, legal personality is a practical legal construct rather than a reflection of moral agency. Consequently, extending a limited form of legal status to highly autonomous AI systems may not be conceptually impossible (Bayern 2021).

Despite these arguments, opposition to electronic personhood remains substantial. Critics emphasize that AI systems lack consciousness, intentionality, moral understanding, and the capacity to experience punishment. Since criminal liability is fundamentally connected to moral blameworthiness, attributing criminal responsibility to AI may undermine the normative foundations of criminal law. Punishing an algorithm does not

express moral condemnation in the same manner as punishing a human offender (Duff 2007).

Moreover, recognizing AI as a legal person may create opportunities for responsibility avoidance. Corporations, developers, and operators could attempt to shift blame onto autonomous systems, thereby weakening incentives for responsible design and governance. Rather than eliminating accountability gaps, electronic personhood might simply relocate responsibility away from actors capable of preventing harm (Turner 2019).

For these reasons, most contemporary legal scholars reject the idea of granting AI systems independent criminal liability. Current AI technologies remain tools rather than autonomous moral agents, and existing legal frameworks continue to regard humans and organizations as the appropriate subjects of criminal responsibility (Abbott 2020).

Comparative Evaluation

A comparison of the four models reveals that no single approach provides a complete solution to the accountability challenges posed by artificial intelligence. Operator liability offers doctrinal familiarity and clear attribution but risks over-criminalization where human control is limited. Developer liability promotes safer system design yet encounters significant evidentiary difficulties concerning causation and foreseeability. Autonomous system accountability attempts to address the autonomy problem directly but remains conceptually controversial and normatively problematic (Hallevy 2015; Turner 2019).

Among the available options, corporate liability appears best positioned to address contemporary accountability gaps. Modern AI systems are generally developed, deployed, and governed by organizations rather than isolated individuals. Corporate liability recognizes the collective nature of technological decision-making and creates incentives for comprehensive governance, compliance, and risk management mechanisms. Furthermore, this model aligns closely with emerging international regulatory frameworks that emphasize organizational responsibility and human oversight throughout the AI lifecycle (Coffee 2020; Veale and Borgesius 2021).

Nevertheless, the most effective solution is likely a hybrid framework combining elements of operator, developer, and corporate liability. Under such a model, responsibility would be allocated according to the specific role each actor played in creating, deploying, or supervising the AI system. Operators would remain accountable for misuse, developers for negligent design, and organizations for governance failures. This multi-layered approach better reflects the distributed nature of AI ecosystems and provides a more comprehensive response to accountability gaps than any single model operating in isolation. As AI systems continue to evolve, criminal liability frameworks must similarly adapt to ensure that technological innovation does not undermine the fundamental principle that serious harms should remain subject to meaningful legal accountability.

TOWARD A RECALIBRATED MENS REA FRAMEWORK

The preceding discussion demonstrates that traditional criminal law faces significant difficulties when applied to harms caused by increasingly autonomous artificial intelligence systems. Contemporary doctrines of criminal liability were developed on the assumption that harmful conduct originates from human actors capable of conscious decision-making, moral reasoning, and intentional action. However, AI technologies challenge these assumptions by introducing forms of decision-making that are neither entirely human nor entirely independent. As a result, the conventional understanding of *mens rea* may no longer be sufficient to address the complexities of AI-related harms. Rather than abandoning the principle of culpability altogether, legal systems must reconsider how criminal responsibility can be attributed in environments where human and machine agency interact in increasingly sophisticated ways (Abbott 2020; Turner 2019).

The need for recalibration arises from the growing gap between technological capabilities and legal concepts. Autonomous systems can influence, recommend, or execute decisions without direct human intervention, yet they remain incapable of possessing intention, knowledge, or moral awareness in the traditional legal sense. Consequently, applying

conventional *mens rea* doctrines often results in situations where significant harm occurs but no individual actor clearly satisfies the required mental element of a criminal offense. This challenge necessitates the development of a liability framework capable of preserving accountability while accommodating technological realities (Matthias 2004; Hallevy 2015).

Limitations of Traditional Mens Rea

One of the most fundamental limitations of traditional *mens rea* doctrine is its human-centered nature. Criminal law historically assumes that liability should be imposed upon actors capable of forming intentions, understanding consequences, and exercising meaningful control over their conduct. Concepts such as intention, knowledge, recklessness, and negligence are therefore constructed around human cognitive capacities. These concepts function effectively when criminal conduct is committed directly by human actors but become increasingly problematic when harmful outcomes emerge from autonomous computational processes (Duff 2007).

The doctrine of intention illustrates this challenge. In conventional criminal law, intention refers to a conscious objective or purpose to bring about a particular result. AI systems may produce outcomes that resemble intentional conduct, yet such systems do not possess desires, motivations, or subjective awareness. Consequently, attributing intention to AI is conceptually difficult, while attributing intention to human actors connected to the system may be equally problematic when they neither intended nor anticipated the harmful outcome (Pagallo 2013).

Similar difficulties arise with respect to knowledge and recklessness. Criminal liability often depends upon demonstrating that an individual knew of a risk or consciously disregarded a foreseeable danger. However, machine-learning systems frequently generate outputs through adaptive and opaque processes that may not be fully understood even by their creators. When harmful outcomes emerge from complex algorithmic interactions, establishing what developers, operators, or organizations actually knew becomes significantly more difficult than in traditional criminal cases (Bathaee 2018).

Furthermore, traditional *mens rea* doctrine assumes a relatively direct relationship between human decision-making and harmful consequences. Autonomous systems disrupt this relationship by introducing multiple layers of technological mediation. Harm may result from interactions among developers, users, training datasets, software updates, and environmental conditions rather than from a single identifiable decision-maker. Consequently, conventional doctrines often struggle to identify a sufficiently culpable actor despite the occurrence of substantial harm (Abbott 2020).

These limitations suggest that future liability frameworks should move beyond an exclusive focus on subjective mental states and incorporate broader assessments of responsibility, risk creation, and governance failures. Such an approach would not eliminate the importance of culpability but would recognize that accountability in AI ecosystems may depend upon factors extending beyond traditional concepts of intention and knowledge (Cane 2002).

Risk-Based Criminal Liability

One promising alternative is the adoption of a risk-based approach to criminal liability. Rather than focusing exclusively on whether a specific actor possessed a particular mental state, risk-based liability evaluates whether individuals or organizations created, ignored, or inadequately managed foreseeable risks associated with AI systems. This approach reflects developments in regulatory law, where responsibility increasingly depends upon risk management and preventive governance rather than solely upon proof of subjective fault (Yeung and Lodge 2019).

A central feature of this model is the concept of foreseeable risk. Under a risk-based framework, liability may arise when an actor deploys or develops an AI system despite reasonably foreseeable dangers associated with its operation. The focus shifts from proving actual intention to examining whether the harmful outcome was sufficiently predictable that preventive measures should have been implemented. Such an approach is particularly relevant for AI technologies because many risks arise not from deliberate misconduct but from inadequate anticipation of system behavior (Marchant

and Wallach 2015).

Closely related to foreseeability is the concept of duty of care. Developers, operators, and organizations involved in AI deployment may be required to meet specific standards of care proportionate to the risks generated by their systems. Failure to implement reasonable safeguards, conduct adequate testing, monitor system performance, or respond to known vulnerabilities could justify criminal liability even in the absence of traditional intent. This approach preserves the principle of culpability while adapting it to technologically complex environments (Calo 2015).

Risk-based liability also encourages proactive behavior. Rather than merely punishing harmful outcomes after they occur, it creates incentives for responsible innovation, continuous monitoring, and effective governance. In this respect, the framework aligns with contemporary regulatory approaches emphasizing prevention and accountability throughout the lifecycle of AI systems (Yeung and Lodge 2019).

Strict Liability for High-Risk AI

While risk-based liability may be appropriate for many AI applications, certain categories of AI systems present risks so substantial that stricter forms of responsibility may be justified. High-risk AI systems possess the potential to cause significant harm to human life, public safety, national security, or critical societal functions. In such circumstances, legal systems may consider limited forms of strict liability that do not require proof of traditional *mens rea* (Schwartz 1997).

Autonomous vehicles provide a notable example. Self-driving vehicles operate in dynamic environments where errors can result in serious injuries or fatalities. Although developers and operators may exercise reasonable care, the potential consequences of system failure are sufficiently severe to justify heightened accountability standards. Strict liability mechanisms may ensure compensation, encourage rigorous safety practices, and reduce evidentiary difficulties associated with proving subjective fault (Marchant and Lindor 2012).

Military AI systems present even greater concerns. Autonomous

weapons capable of selecting and engaging targets raise profound legal and ethical questions regarding accountability for unlawful killings, disproportionate attacks, or violations of international humanitarian law. Given the potentially catastrophic consequences associated with military AI, reliance solely upon traditional mens rea doctrines may be inadequate. Enhanced liability standards may therefore be necessary to ensure meaningful accountability for decisions involving lethal force (Asaro 2012).

Similarly, AI systems deployed within critical infrastructure sectors—including energy networks, healthcare systems, transportation networks, and financial markets—can generate widespread societal harm if they malfunction or are improperly managed. Because failures within these sectors may affect large populations, strict liability regimes may provide a stronger incentive for organizations to prioritize safety, resilience, and risk management (Bostrom and Yudkowsky 2014).

Importantly, strict liability should not be applied universally to all AI systems. Excessive reliance on strict liability could discourage innovation and impose unfair burdens upon developers and operators. Instead, such liability should be reserved for narrowly defined categories of high-risk applications where the magnitude of potential harm justifies heightened legal obligations (Schwartz 1997).

Enhanced Regulatory Obligations

Reforming criminal liability alone is unlikely to resolve all accountability challenges associated with AI. Effective governance also requires robust regulatory obligations designed to prevent harm before it occurs. Increasingly, policymakers and scholars recognize that transparency, explainability, human oversight, and auditability constitute essential components of responsible AI governance (Floridi et al. 2018).

Transparency refers to the availability of information concerning the design, purpose, capabilities, and limitations of AI systems. Transparency enables regulators, courts, and affected individuals to understand how systems operate and to identify potential sources of harm. Without transparency, assigning responsibility becomes significantly more difficult

because relevant decision-making processes remain obscured (Wachter, Mittelstadt, and Russell 2018).

Closely related is the principle of explainability. Explainability requires that AI-generated decisions be understandable to human users and regulators. Although complete transparency may not always be technically feasible, organizations should provide sufficient explanations to permit meaningful evaluation of system behavior. Explainability is particularly important in criminal liability contexts because establishing causation and foreseeability often depends upon understanding how a particular outcome was produced (Doshi-Velez and Kim 2017).

Human oversight constitutes another critical safeguard. Autonomous systems should not operate entirely beyond human supervision, particularly in contexts involving significant risks to life, liberty, or public safety. Meaningful human oversight ensures that responsible actors retain the capacity to monitor, intervene, and override AI-generated decisions when necessary. Such oversight strengthens accountability by preserving a clear connection between technological decisions and human responsibility (Smuha 2021). Finally, auditability provides mechanisms for documenting and reviewing AI system behavior. Audit trails, record-keeping requirements, and independent assessments facilitate post-incident investigations and improve the ability of courts and regulators to determine responsibility. By creating verifiable records of system operation, auditability reduces the evidentiary difficulties commonly associated with AI-related harms (Mittelstadt 2019).

Hybrid Accountability Model

Given the complexity of AI ecosystems, no single liability model is capable of addressing all accountability challenges. A more effective solution is a hybrid accountability model that combines multiple layers of responsibility. Such a framework recognizes that AI-related harms often result from interactions among developers, operators, organizations, and regulatory systems rather than from the actions of a single actor (Abbott 2020).

Under this model, operators would remain responsible for negligent or improper use of AI systems. Where individuals fail to exercise reasonable care in deploying or supervising AI technologies, criminal liability could be imposed based upon established principles of negligence or recklessness. This preserves traditional concepts of personal responsibility while recognizing the continuing importance of human oversight (Ashworth 2021).

Developers would bear responsibility for foreseeable harms arising from defective design, inadequate testing, insufficient safety measures, or reckless disregard of known risks. Such liability would encourage responsible innovation and incentivize the incorporation of safety mechanisms throughout the development process (Calo 2015).

Organizations and corporate entities would be accountable for governance failures, compliance deficiencies, and inadequate risk management structures. Because most advanced AI systems are developed and deployed within organizational environments, corporate accountability provides an essential mechanism for addressing systemic risks that cannot be attributed to individual actors alone (Wells 2020).

Finally, regulatory compliance obligations would function as a complementary layer of accountability. Organizations deploying high-risk AI systems would be required to satisfy transparency, explainability, oversight, and auditing requirements established by law. Failure to comply with these obligations could trigger administrative, civil, or criminal consequences depending upon the severity of the violation (Veale and Borgesius 2021).

The hybrid accountability model therefore represents the most balanced response to the challenges posed by autonomous technologies. Rather than attempting to fit AI-related harms within rigid traditional doctrines or granting legal personality to AI systems themselves, this approach distributes responsibility across all relevant actors according to their respective roles and capacities. By integrating criminal liability, regulatory governance, and risk management principles, a recalibrated mens rea framework can preserve accountability while remaining adaptable to the evolving realities of artificial intelligence.

POLICY IMPLICATIONS AND FUTURE LEGAL REFORM

The rapid development of artificial intelligence requires a reconsideration of existing criminal law frameworks and regulatory approaches. Although current legal systems generally maintain the principle that criminal responsibility must ultimately be attributed to human actors or legal entities, the increasing autonomy of AI systems creates circumstances in which traditional concepts of responsibility may become insufficient. Future legal reform should therefore not seek to replace fundamental criminal law principles but rather recalibrate them to address new forms of technological risk, distributed decision-making, and complex causation. A balanced approach should preserve individual accountability while introducing mechanisms capable of addressing organizational and systemic failures associated with AI deployment (Abbott 2020; Turner 2019).

Reforming Criminal Codes

One of the most important areas for reform concerns the modernization of criminal codes. Existing criminal statutes were largely developed in an era where harmful conduct was assumed to originate from identifiable human decisions. Consequently, concepts such as intention, knowledge, negligence, and causation may require reinterpretation when applied to autonomous systems. Criminal law reform should consider expanding the understanding of responsibility beyond direct human action and include failures related to the design, deployment, supervision, and governance of AI technologies (Hallevy 2015).

A possible reform approach involves introducing explicit provisions addressing responsibility for unsafe AI deployment. Such provisions could establish criminal offences for organizations or individuals that release, operate, or maintain high-risk AI systems without adequate safety mechanisms, testing procedures, or monitoring arrangements. Similar to regulatory offences in environmental and corporate criminal law, liability could be based on failures to comply with legally established duties of care rather than requiring proof of traditional intentional wrongdoing (Coffee

2020).

However, reform should avoid creating excessive criminalization that discourages technological innovation. Criminal sanctions should be reserved for serious failures involving reckless disregard of significant risks, deliberate avoidance of safety obligations, or systemic compliance failures. Minor technical errors or unforeseeable AI behavior should generally remain within civil liability or regulatory enforcement frameworks. This distinction is essential to maintain proportionality, which remains a fundamental principle of criminal justice (Ashworth 2021).

Furthermore, criminal codes may need to recognize new forms of organizational responsibility. Because AI systems are typically developed and managed through complex institutional structures, responsibility should not depend exclusively on identifying a single individual possessing the required *mens rea*. Instead, legal systems may need to incorporate concepts of corporate accountability, governance failure, and organizational negligence to ensure that responsibility remains attached to actors capable of preventing harm (Wells 2020).

International Harmonization

AI technologies operate across national borders, creating significant challenges for fragmented regulatory approaches. An AI system may be developed in one jurisdiction, trained using global datasets, deployed through international platforms, and cause harm in another country. This transnational nature requires greater cooperation among states and the development of international standards concerning AI safety, accountability, and liability attribution (Calo 2015).

International harmonization does not necessarily require identical criminal laws across jurisdictions. Instead, it requires agreement on fundamental principles, including transparency obligations, human oversight requirements, risk assessment procedures, and minimum accountability standards. Similar to international cooperation in areas such as cybersecurity and environmental protection, global AI governance requires shared principles that enable states to address cross-border technological

risks effectively (Floridi et al. 2018).

The development of global standards is particularly important for high-risk AI applications, including autonomous vehicles, healthcare systems, financial algorithms, and military technologies. Without international coordination, organizations may engage in regulatory arbitrage by deploying risky systems in jurisdictions with weaker oversight. Harmonized standards can reduce this possibility by establishing common expectations for responsible AI development and deployment (Veale and Borgesius 2021).

International organizations and standard-setting institutions may therefore play an increasingly important role in shaping AI governance. Frameworks developed by bodies such as the Organisation for Economic Co-operation and Development (OECD) and the United Nations can provide common principles while allowing individual jurisdictions to adapt rules according to their legal traditions and social contexts (OECD 2019).

Ethical and Human Rights Considerations

Legal reform concerning AI liability must also incorporate ethical and human rights considerations. AI systems increasingly influence decisions affecting employment, healthcare access, financial opportunities, law enforcement, and individual freedoms. Therefore, accountability frameworks must ensure that technological efficiency does not undermine fundamental rights and democratic values (Mittelstadt 2019).

Fairness represents one of the central concerns. AI systems trained on biased or incomplete datasets may reproduce or amplify existing social inequalities. From a legal perspective, responsibility mechanisms should ensure that developers and deployers undertake adequate measures to identify and mitigate discriminatory outcomes. Failure to address foreseeable bias may constitute a governance failure requiring legal consequences (Barocas and Selbst 2016).

Transparency is another essential requirement. Individuals affected by AI-generated decisions should have access to meaningful information regarding how those decisions are produced. Transparency promotes accountability by enabling individuals, regulators, and courts to evaluate

whether AI systems operate lawfully and fairly. Without adequate transparency, affected persons may be unable to challenge automated decisions effectively (Wachter, Mittelstadt, and Russell 2018).

Due process considerations are equally important. Where AI systems influence decisions involving legal rights or significant personal consequences, individuals should have opportunities to contest automated outcomes and seek human review. The principle of due process requires that technological systems remain subject to legal oversight rather than becoming autonomous mechanisms of authority beyond human accountability (Citron and Pasquale 2011).

Future Challenges

Future developments in artificial intelligence will create increasingly complex challenges for criminal law. The emergence of Artificial General Intelligence (AGI)—systems capable of performing a broad range of intellectual tasks comparable to human cognitive abilities—raises questions that extend beyond current AI governance debates. If AGI systems eventually demonstrate advanced autonomy, adaptability, and strategic decision-making, existing concepts of responsibility may require even deeper reconsideration (Bostrom 2014).

Fully autonomous systems represent another significant challenge. Current AI systems generally operate within boundaries established by human designers and organizations. However, future systems may possess greater independence in pursuing objectives, modifying strategies, and interacting with complex environments. In such circumstances, determining whether responsibility should remain exclusively human-centered or whether new legal categories are required will become an increasingly important theoretical and practical issue (Bryson 2018).

The possibility of highly autonomous AI also raises broader philosophical questions regarding criminal law in post-human technological environments. Criminal law has traditionally been based on assumptions of human agency, consciousness, and moral responsibility. If technological systems eventually develop capabilities approaching human-level autonomy,

legal scholars may need to reconsider whether existing concepts of personhood, responsibility, and accountability remain adequate (Gunkel 2018).

Nevertheless, current developments do not justify abandoning human-centered criminal responsibility. AI systems remain technological artifacts rather than moral agents capable of understanding obligations or experiencing punishment. Therefore, the most appropriate future direction is likely the development of sophisticated accountability structures that maintain human responsibility while recognizing the distinctive risks created by autonomous technologies (Abbott 2020).

Therefore, the future of criminal law in the age of artificial intelligence will depend on achieving a balance between technological innovation and legal accountability. Reform should not focus solely on assigning blame after harm occurs but should also promote preventive governance, responsible design, and meaningful human control. Through adaptive legal frameworks, international cooperation, and rights-based regulation, criminal law can remain effective in governing technological environments characterized by increasing autonomy and complexity.

CONCLUSION

The development of artificial intelligence and autonomous systems has created fundamental challenges for traditional criminal liability frameworks, particularly regarding the application of *mens rea* as the basis of criminal responsibility. Conventional criminal law is primarily designed around human actors who possess consciousness, intention, and moral agency, while AI systems operate through algorithmic processes that may generate unpredictable outcomes without possessing a guilty mind. This situation creates significant accountability gaps because existing doctrines often struggle to identify whether responsibility should be attributed to operators, developers, corporations, or other actors involved in the AI lifecycle. Nevertheless, the solution should not be the abandonment of *mens rea* as a foundational principle of criminal law, as the requirement of culpability

remains essential to ensure fairness and proportionality in punishment. Instead, criminal law should be recalibrated through a hybrid accountability framework that distributes responsibility according to the role and level of control exercised by different actors.

Future legal reform should focus on developing a more adaptive liability framework capable of addressing the unique risks associated with autonomous technologies. The adoption of risk-based liability, the introduction of stricter obligations for high-risk AI deployment, and the strengthening of regulatory mechanisms concerning transparency, explainability, human oversight, and auditability represent important steps toward closing existing accountability gaps. In addition, international cooperation is necessary to establish common standards for AI governance, considering the cross-border nature of emerging technologies. By combining traditional principles of criminal responsibility with innovative regulatory approaches, legal systems can maintain human accountability while ensuring that technological advancement develops within a framework of safety, responsibility, and respect for fundamental legal values.

ACKNOWLEDGMENTS

None.

FUNDING INFORMATION

None.

CONFLICTING INTEREST AND ORGINALITY STATEMENT

The authors state that there is no conflict of interest in the publication of this article.

GENERATIVE AI STATEMENT

None.

REFERENCES

Abbott, Ryan. 2020. *The Reasonable Robot: Artificial Intelligence and the Law*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108670874.

Alexander, Larry, and Kimberly Kessler Ferzan. 2009. *Crime and Culpability: A Theory of Criminal Law*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511817449.

Alpaydin, Ethem. 2021. *Machine Learning*. Revised and Updated Edition. Cambridge, MA: MIT Press.

Asaro, Peter M. 2012. "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making." *International Review of the Red Cross* 94 (886): 687–709. https://doi.org/10.1017/S1816383113000115.

Ashworth, Andrew. 2021. *Principles of Criminal Law*. 10th ed. Oxford: Oxford University Press.

Balkin, Jack M. 2015. "The Path of Robotics Law." *California Law Review Circuit* 6: 45–60.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671–732. https://doi.org/10.15779/Z38BG31.

Bathaee, Yavar. 2018. "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harvard Journal of Law & Technology* 31 (2): 889–938.

Bayern, Shawn. 2021. *The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems*. Stanford, CA: Stanford University Press.

Bentham, Jeremy. (1789) 1996. *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 316–34. Cambridge: Cambridge University Press.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford:

Oxford University Press.
Brenner, Susan W. 2010. *Cybercrime: Criminal Threats from Cyberspace*. Santa Barbara, CA: Praeger.
Bryson, Joanna J. 2018. "Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20: 15–26. https://doi.org/10.1007/s10676-018-9448-6.
Buell, Samuel W. 2016. *Capital Offenses: Business Crime and Punishment in America's Corporate Age*. New York: W.W. Norton.
Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 1–12. https://doi.org/10.1177/2053951715622512.
Calo, Ryan. 2015. "Robotics and the Lessons of Cyberlaw." *California Law Review* 103 (3): 513–63. https://doi.org/10.15779/Z38D93B.
Cane, Peter. 2002. *Responsibility in Law and Morality*. Oxford: Hart Publishing.
Citron, Danielle Keats, and Frank Pasquale. 2011. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89 (1): 1–33.
Coeckelbergh, Mark. 2020. *AI Ethics*. Cambridge, MA: MIT Press.
Coffee, John C. 2020. *Corporate Crime and Punishment: The Crisis of Underenforcement*. Oakland: Berrett-Koehler Publishers.
Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv preprint. https://doi.org/10.48550/arXiv.1702.08608.
Dressler, Joshua. 2022. *Understanding Criminal Law*. 9th ed. Durham, NC: Carolina Academic Press.
Duff, R. A. 2007. *Answering for Crime: Responsibility and Liability in the Criminal Law*. Oxford: Hart Publishing.
Duff, R. A. 2018. *The Realm of Criminal Law*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198808938.001.0001.
European Parliament and Council. 2024. *Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Brussels: European Union.
European Parliament. 2017. *Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*. Strasbourg: European

Parliament.
Floridi, Luciano, et al. 2018. "AI4People—An Ethical Framework for a Good AI Society." *Minds and Machines* 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5.
Gardner, John. 2007. *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*. Oxford: Oxford University Press.
Gobert, James, and Maurice Punch. 2003. *Rethinking Corporate Crime*. London: Butterworths.
Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.
Gunkel, David J. 2018. *Robot Rights*. Cambridge, MA: MIT Press.
Hallevy, Gabriel. 2015. *Liability for Crimes Involving Artificial Intelligence Systems*. Cham: Springer. https://doi.org/10.1007/978-3-319-10124-3.
Herring, Jonathan. 2022. *Criminal Law: Text, Cases, and Materials*. 9th ed. Oxford: Oxford University Press.
Hutchinson, Terry, and Nigel Duncan. 2012. "Defining and Describing What We Do: Doctrinal Legal Research." *Deakin Law Review* 17 (1): 83–119. https://doi.org/10.21153/dlr2012vol17no1art70.
Kaplan, Andreas, and Michael Haenlein. 2019. "Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence." *Business Horizons* 62 (1): 15–25. https://doi.org/10.1016/j.bushor.2018.08.004.
Kerr, Orin S. 2023. *Criminal Law*. 6th ed. St. Paul, MN: West Academic Publishing.
Koops, Bert-Jaap, Mireille Hildebrandt, and David-Olivier Jaquet-Chiffelle, eds. 2010. *Bridging the Accountability Gap: Rights for New Entities in the Information Society?* Nijmegen: Wolf Legal Publishers.
Marchant, Gary E., and Rachel A. Lindor. 2012. "The Coming Collision between Autonomous Vehicles and the Liability System." *Santa Clara Law Review* 52 (4): 1321–40.
Matsuo, Yutaka. 2019. *Artificial Intelligence and Deep Learning in Japan*. Tokyo: Springer Japan.
Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility

for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

McConville, Mike, and Wing Hong Chui, eds. 2017. *Research Methods for Law*. 2nd ed. Edinburgh: Edinburgh University Press.

Michaels, Ralf. 2009. "The Functional Method of Comparative Law." In *The Oxford Handbook of Comparative Law*, edited by Mathias Reimann and Reinhard Zimmermann, 339–82. Oxford: Oxford University Press.

Mittelstadt, Brent. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1 (11): 501–507. https://doi.org/10.1038/s42256-019-0114-4.

Moore, Michael S., and Heidi M. Hurd. 2011. *The Moral Foundations of Law*. Oxford: Oxford University Press.

OECD. 2019. *Recommendation of the Council on Artificial Intelligence*. Paris: OECD Publishing.

Ormerod, David, and Karl Laird. 2023. *Smith and Hogan's Criminal Law*. 17th ed. Oxford: Oxford University Press.

Pagallo, Ugo. 2013. *The Laws of Robots: Crimes, Contracts, and Torts*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-6564-1.

Robinson, Paul H. 1997. *Structure and Function in Criminal Law*. Oxford: Oxford University Press.

Russell, Stuart, and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Pearson.

Sartor, Giovanni, and Francesca Lagioia. 2020. "The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence." *European Parliamentary Research Service Study*. Brussels: European Parliament.

Schwartz, Gary T. 1997. "The Ethics and Economics of Tort Liability Insurance." *Cornell Law Review* 75 (2): 313–61.

Simester, Andrew P., John R. Spencer, Graham J. Virgo, and G. R. Sullivan. 2019. *Simester and Sullivan's Criminal Law: Theory and Doctrine*. 7th ed. Oxford: Hart Publishing.

Smith, J. C., Brian Hogan, and David Ormerod. 2018. *Smith, Hogan, and Ormerod's Criminal Law*. 15th ed. Oxford: Oxford University Press.

Smits, Jan M. 2017. *Comparative Law: A Very Short Introduction*. Oxford: Oxford

University Press. https://doi.org/10.1093/actrade/9780198791353.001.0001.

Smuha, Nathalie A. 2021. "From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence." *Law, Innovation and Technology* 13 (1): 57–84. https://doi.org/10.1080/17579961.2021.1898300.

Taekema, Sanne. 2018. "Theoretical and Normative Frameworks for Legal Research: Putting Theory into Practice." *Law and Method* 8: 1–18. https://doi.org/10.5553/REM/.000031.

Turner, Jacob. 2019. *Robot Rules: Regulating Artificial Intelligence*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-02487-9

Veale, Michael, and Frederik Zuiderveen Borgesius. 2021. "Demystifying the Draft EU Artificial Intelligence Act." *Computer Law Review International* 22 (4): 97–112. https://doi.org/10.9785/cri-2021-220402.

von Hirsch, Andrew. 1993. *Censure and Sanctions*. Oxford: Clarendon Press.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31 (2): 841–87.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology* 31 (2): 841–87.

Wells, Celia. 2020. *Corporations and Criminal Responsibility*. 2nd ed. Oxford: Oxford University Press.

Yeung, Karen, and Martin Lodge. 2019. *Algorithmic Regulation*. Oxford: Oxford University Press.

Yeung, Karen, Andrew Howes, and Ganna Pogrebna. 2020. "AI Governance by Human Rights-Centred Design, Deliberation and Oversight." *Philosophical Transactions of the Royal Society A* 378 (2183): 20190016. https://doi.org/10.1098/rsta.2019.0016.

Zweigert, Konrad, and Hein Kötz. 1998. *An Introduction to Comparative Law*. 3rd ed. Oxford: Clarendon Press.

This page is intetionally left blank