African Journal of Law and Justice System (AJLJS)

ISSN 2753-3115 (Print) ISSN 2753-3123 (Online)

Indexed by SCOPUS, IBSS, EBSCO and SABINET

Volume 4, Number 3, December 2025

Pp 205-225

Criminal Responsibility in the Age of AI: Rethinking Legal Accountability

DOI: https://doi.org/10.31920/2753-3123/2025/v4n3a10

Dr. Yasein Hassan M. Osman[1]

Associate Professor. School of Law, Department of Public Law (Criminal), Kampala International University, Kampala, Uganda - yasein@kiu.ac.ug
Associate Professor, Department of Public Law (Criminal Law), Faculty of Law and Sharia, University of Nyala, Sudan –
Email: yasseinhassan@yahoo.com,
https://orcid.org/0009-0006-8830-8019

***Damina Joshua John*[2]**

School of Law Kampala International University Kampala Uganda
Department of public and Comparative Law.
john.damina@kiu.ac.ng
Corresponding author***

Dalia Abass Ahmed[3]

Stardom University & Majmmah University
Faculty of Law-Civil & Comparative Department
Email: dalyaabass4@gmail.com

Dr. Rowaida Musa Abdelazez Mohamed[4]

Assistant Professor. Buraydah Private,
Colleges - College of Law, Private Law Department,
Al-Qassim, Buraydah, Saudi Arabia,
E-mail: rowida439@gmail.com

&

Dr. Rasheeda Mahmoud Said Ahmed Ali[5]
Assistant Professor. Buraydah Private, Colleges - College of Law, Public Law Department, Al-Qassim, Buraydah, Saudi Arabia
E-mail rasheedaali555@gmail.com

Abstract

Artificial intelligence (AI) is increasingly capable of making autonomous decisions that can cause harm. This prompts important questions of criminal liability. The existing legal frameworks built on agency and intent are inadequate to accommodate situations where there is no direct human actor that caused the harm. This paper assesses the ongoing debate about whether AI should just be treated as a tool for which humans are liable for or whether AI ought to be treated as its own legal entity that can be held accountable. The research applies the retributive theory of punishment and classical legal rationales relating to mens rea, the causal relationship and legal personhood to identify whether there are conceptual issues with attaching liability to AI. The research used a qualitative descriptive legal-philosophical approach to review secondary literature, international political documents and comparative legal models across multiple jurisdictions. The research identified key areas of law which are not currently adequately reflected in liability regimes, naming some areas of connection with corporate liability and product liability, as well as considering emerging ideas such as electronic personhood, strict liability frameworks and preventative regulatory frameworks. The preliminary conclusion is that although there is no consensus on whether there will be direct criminal liability for AI, much more reform is needed to avoid accountability gaps and to ultimately preserve public confidence in the law. The paper concluded with policy recommendations for better clarity between human accountability and possibilities for regulated AI personhood, requirements for safety and transparency, training for legal agents, the scope of international cooperation, etc., which would work towards a legal framework which serves the interests of justice in relation to autonomous AI systems while incorporating key conventions.

Commented

Keywords: *Artificial intelligence, Criminal responsibility, Legal personhood, Mens rea, Autonomous systems, Liability frameworks*

Introduction

Artificial intelligence (AI) is spreading fast in our lives. This includes self-driving cars, medical diagnostics, and automated finance. According to

Bajwa[615], these systems can act without direct human control. This raises urgent legal questions. Traditional criminal law assumes that only humans (or legal persons like companies) can form a guilty intention (mens rea) and do a forbidden act (actus reus).[616] AI, however, has no mind or intent in the human sense. As one scholar notes, AI systems are limited and lack "the psychological capabilities" that qualify them to bear criminal responsibility.[617] If a smart car runs a red light and kills someone, who is at fault – the owner, the programmer, the car itself, or no one? Morales[618] warns that if AI is exempted from liability, we will end up with a void where power exists without answerability. In practice, most countries still treat AI-caused harm as human fault: Iran and Saudi Arabia, for example, agree that "AI itself cannot bear criminal responsibility, and instead liability is attributed to human actors."[619] But such rules may leave dangerous gaps as AI grows smarter.

Historically, criminal law has revolved around the notion of a morally and legally responsible human agent – someone who can act with *mens rea* (a guilty mind) and commit *actus reus* (a guilty act).[620] These concepts assume the capacity for intention, understanding, and moral agency – traits that machines do not naturally possess. However, the modern legal landscape must now grapple with scenarios where no human actor's conduct neatly fits the chain of causation. For example, an AI trading system could manipulate financial markets in ways unforeseen by its

---

615 Junaid Bajwa and others, 'Artificial Intelligence in Healthcare: Transforming the Practice of Medicine' (2021) 8 Future Healthc J e188 <https://www.sciencedirect.com/science/article/pii/S2514664524005277> accessed 22 September 2025.

616 Alessandro Stasi, 'Actus Reus and Mens Rea' [2021] General Principles of Thai Criminal Law 25 <https://link.springer.com/chapter/10.1007/978-981-15-8708-5_3> accessed 22 September 2025.

617 Elina Nerantzi and Giovanni Sartor, '"Hard AI Crime": The Deterrence Turn' (2024) 44 Oxf J Leg Stud 673; Elina Nerantzi, '"All AIs Are Psychopaths"? The Scope and Impact of a Popular Analogy' (2025) 38 Philos Technol 1 <https://link.springer.com/article/10.1007/s13347-025-00856-x> accessed 22 September 2025.

618 Janielle Morales and A Vhon, 'WHEN MACHINES BREAK THE LAW: THE CASE FOR CRIMINAL LIABILITY & LEGAL PERSONHOOD FOR ARTIFICIAL INTELLIGENCE'.

619 Zahra Meghdadi and Mahdi Pourcheriki, 'Comparative Study of Criminal Responsibility of AI in the Legal Framework of Iran and Saudi Arabia' (2025) 2 Journal of Innovations in Computer Science and Engineering (JICSE) 87 <https://jicse.sbu.ac.ir/article_105767.html> accessed 22 September 2025.

620 Ariane Bigenwald and Valerian Chambon, 'Criminal Responsibility and Neuroscience: No Revolution Yet' (2019) 10 Front Psychol 454562 <www.frontiersin.org> accessed 22 September 2025.

developers; a delivery drone could autonomously violate airspace laws; or an AI-driven medical diagnostic tool could misclassify a critical illness, leading to preventable death. In each instance, the harm is real, but the locus of blame is legally ambiguous.

This tension has prompted a growing scholarly and policy debate across jurisdictions. Some argue that AI should be treated purely as a tool,[621] placing full liability on the human designers, operators, or deployers. Others contend that AI's increasing autonomy calls for the recognition of a distinct legal status—maybe akin to corporate personhood—enabling the direct attribution of responsibility to the system itself.[622] International bodies such as the European Union, the OECD, and the Council of Europe have begun to explore regulatory models, though consensus remains elusive.[623] The challenge is not simply one of drafting new laws; it is a deeper philosophical problem concerning the boundaries of legal personhood, culpability, and the function of criminal punishment in a world where decision-making agents may not even be human at all.[624]

As societies continue to weave AI into critical infrastructure, the stakes grow higher. Without a coherent framework for assigning criminal responsibility in cases involving autonomous systems, there is a risk of legal gaps and inconsistent rulings, as well as diminished public trust in both technology and the justice system. This article seeks to examine the evolving discourse on AI and criminal liability. It analyses existing legal

621 'Intellectual Property Rights and AI-Generated Content — Issues in Human Authorship, Fair Use Doctrine, and Output Liability | by Adnan Masood, PhD. | Medium' <https://medium.com/@adnanmasood/intellectual-property-rights-and-ai-generated-content-issues-in-human-authorship-fair-use-8c7ec9d6fdc3> accessed 22 September 2025; Wei Li and others, 'AI Creativity and Legal Protection for AI-Generated Works in Posthuman Societal Scenarios' (2025) 9 Sustainable Futures 100749 <https://www.sciencedirect.com/science/article/pii/S2666188825003156> accessed 22 September 2025.

622 Arto Laitinen and Otto Sahlgren, 'AI Systems and Respect for Human Autonomy' (2021) 4 Front Artif Intell; Rafael Dean Brown, 'Property Ownership and the Legal Personhood of Artificial Intelligence' (2021) 30 Information & Communications Technology Law 208.

623 Miguel Ángel Presno Linera and Anne Meuwese, 'Regulating AI from Europe: A Joint Analysis of the AI Act and the Framework Convention on AI' [2025] The Theory and Practice of Legislation 1.

624 Ryan Benjamin Abbott and Alex F Sarch, 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction' [2019] SSRN Electronic Journal; Filippo Santoni de Sio and Giulio Mecacci, 'Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them' (2021) 34 Philos Technol 1057.

doctrines, emerging theoretical approaches, and the potential pathways for reform in an increasingly automated world. It begins by explaining core legal concepts of responsibility. It then reviews existing laws and proposals around the world. The following part discusses the key challenges in attributing crime and intent when AI is involved and suggested solutions. The conclusion reflects on these findings, and the recommendations suggest ways forward. Through this analysis, we aim to show that law must evolve to address AI's autonomous behavior while still upholding justice and public safety.

Theoretical Framework

One relevant theory is the retributive theory of punishment as applied to AI. Abbott and Sarch[625] argue that when an autonomous system commits a crime in a way that cannot be reduced to any human wrongdoer – what they call a "Hard AI Crime" – there may be value in holding the AI itself criminally accountable. From this view, punishing the AI could have a deterrent effect and express societal condemnation of the harm. However, Abbott and Sarch caution that this approach risks "conceptual confusion" and other costs.[626] In other words, a machine cannot literally suffer or understand punishment, so treating it like a guilty person can blur important legal concepts. This theory's flaw is that it stretches the meaning of responsibility: is it fair to punish an AI that cannot be moral? Here, a theoretical tension exists between the utility of deterrence and the demand for moral blame.

Another perspective is classical criminal doctrine. This perspective posits that an individual with a "guilty mind" (mens rea) must decide to commit a "guilty act" (actus reus) in order to commit a crime.[627] In the 18th and 19th centuries, thinkers such as Kant and legal codifiers argued that only rational agents could face punishment for crimes. Today's laws continue this view. For example, many sources note that current criminal rules are not compatible with the reality of crimes involving smart machines.[628] The doctrine implies that only beings capable of intent can be criminals. Its weakness in the AI context is that it leaves a moral void: if no human planned an AI's action, traditional theory has no ready answer.

---

625 Abbott and Sarch (n 10).

626 ibid.

627 'Classical School of Criminology | Overview, Theory & Principles - Lesson | Study.Com' <https://study.com/academy/lesson/the-classical-school-of-criminology-its-influence-today.html> accessed 22 September 2025.

628 Nerantzi and Sartor (n 3); Nerantzi (n 3).

In practice, courts attempt to stretch these ideas (for example, by imputing intent through negligence or foreseeability). But as we will see, this approach alone struggles with autonomous AI.

Methodology

Research design

This study adopted a qualitative, descriptive legal-philosophical approach. It focuses on a conceptual and normative analysis of criminal responsibility in the context of Artificial Intelligence (AI). The study was structured as a doctrinal legal analysis combined with comparative elements. The aim was to evaluate existing criminal law principles, such as *mens rea*, causation and legal personhood, against emerging AI capabilities. The scope was deliberately international, encompassing regional instruments (e.g., EU guidelines, African Union frameworks), global policy documents (e.g., UNICRI reports), and domestic laws. This ensured a broad yet focused coverage of both common law and civil law traditions.

Data source identification

The primary sources of data included secondary legal literature (peer-reviewed journal articles, law review commentaries, and scholarly monographs addressing AI and criminal liability), official documents and policy papers (European Commission white papers, UN guidelines on AI governance, OECD AI principles, and national AI strategies), and case law and statutory provisions from relevant jurisdictions where legal debates or precedents on AI accountability have emerged. Current sources were prioritized to reflect the most current legal developments and debates, while also referencing seminal earlier works where necessary for conceptual grounding.

Data collection process

Literature and documents were collected through systematic searches in major academic databases such as HeinOnline, JSTOR, Westlaw, LexisNexis, and Google Scholar, as well as official institutional repositories. Search terms included "AI criminal liability", "robot personhood law", "mens rea and AI", "autonomous systems and

culpability", and "algorithmic accountability in criminal law". Additional materials were obtained through citation tracking from key articles and policy documents to ensure completeness. Only credible, peer-reviewed, or officially endorsed sources were included.

Data analysis

The gathered materials were subjected to thematic content analysis to identify recurrent concepts, controversies, and legal gaps. The thematic analysis involves identifying, analyzing, and reporting patterns (themes) in the data. The process includes familiarizing oneself with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report. Each theme was analyzed in light of both doctrinal principles and policy considerations. Cross-jurisdictional findings were compared to reveal patterns and divergences and to formulate proposed legal models for reform. Where conflicting interpretations arose, these were critically examined and triangulated against multiple authoritative sources to ensure validity.

Ethical considerations

As the study relies solely on secondary data, no human participants were involved, and thus no institutional ethical approval was required. However, interpretive bias was minimized through transparent documentation of search strategies, criteria for source inclusion, and a consistent application of thematic coding across all materials reviewed.

Analysis

Concepts of Criminal Responsibility and AI

Criminal responsibility traditionally requires a subject (a person or legal entity) and a mental state. In law, a "person" is defined as a subject of rights and duties.[629] Alehime notes that by contrast, a "thing" is property that cannot hold rights or duties. Under this framework, AI systems are treated as things (objects), not individuals. For example, French legal theory notes that "legal personhood" moves an entity from a passive thing

[629] Philippe Fauquet Alekhine, 'Can the Robot Be Considered a Person? The European Perspective' (2022) 23 Advances in Research on Teaching 100 <https://hal.science/hal-05197912> accessed 22 September 2025.

to an active subject who can assert rights and bear duties.[630] To date, AI systems have no legal personhood: they cannot hold rights or be punished. Existing criminal law reflects this: AI itself cannot be charged. Instead, the law typically treats AI as a "human instrument." Like a gun or a corporation, the AI can only commit a crime through a human actor. Indeed, one study finds that the dominant view is to assign any smart machine's wrongdoing to a natural person.[631]

Nerantzi[632] explains that machines "lack the psychological capabilities" to qualify as criminals, so responsibility must be attached to humans (designers, owners, operators). Similarly, in Iran and Saudi Arabia, scholars conclude that "AI itself cannot bear criminal responsibility", and thus blame falls on developers or users.[633] Some legal systems recognize nonhuman persons. Corporations can be convicted under controlled circumstances. The Model Penal Code allows corporate criminal liability if an employee acts within their scope.[634] But this is a fiction because a corporation's "mind" is essentially the consensus of its people. AI differs because it can learn and act without direct human orders.

This distinction is not merely technical but doctrinal. Corporate liability preserves the foundational link between human intent and legal blame – directors or employees must have acted with requisite mens rea for the corporation to be liable. AI, by contrast, may generate harmful outcomes through emergent behaviors that no human foresees or authorizes. To stretch corporate personhood to cover such cases would not extend criminal law but distort it, substituting legal fiction for actual culpability. This reveals a core tension: criminal responsibility requires a locus of blameworthiness, yet autonomous AI dissolves that locus. Unless the law changes, AI remains in the realm of things.

One of the most pressing conceptual issues is whether existing legal doctrines can be adapted to encompass AI-driven conduct without creating entirely new legal categories. In some jurisdictions, corporate

630 ibid; Udo Thiel, 'A Theory of Legal Personhood' [2019] The Early Modern Subject: Self-Consciousness and Personal Identity from Descartes to Hume 1 <https://papers.ssrn.com/abstract=3446600> accessed 22 September 2025.

631 Haroon Sheikh, Corien Prins and Erik Schrijvers, 'Artificial Intelligence: Definition and Background' (2023).

632 Nerantzi (n 3).

633 Meghdadi and Pourcheriki (n 5).

634 Abbott and Sarch (n 10).

criminal liability provides a partial analogy.[635] This is because corporations, though artificial entities, can be held accountable for crimes committed through their agents. This suggests that legal responsibility need not be limited to natural persons. But, unlike corporations, AI does not have shareholders, directors, or employees in the traditional sense[636], making it harder to identify whose intentions and knowledge should be imputed to the system. Scholars such as Walter[637] argue that attempting to fit AI into corporate liability frameworks risks oversimplifying its unique operational dynamics.

Another conceptual challenge lies in the attribution of mens rea to AI. Even if an AI system's actions satisfy the actus reus element, criminal law requires proof of a mental state such as intent, knowledge, recklessness, or negligence.[638] AI cannot "intend" in the human sense, but it can be programmed to pursue certain objectives, adapt to changing circumstances, and make decisions based on probabilistic reasoning. Some legal theorists propose treating certain high-level algorithmic functions as proxies for mental states[639], while others caution that such analogies blur the line between human culpability and machine behavior, potentially undermining fundamental principles of justice. This debate exposes a deeper methodological problem. Criminal law's mens rea doctrine is not merely about observable behavior but about moral agency. Even if an AI "pursues a goal" in a functional sense, it does so without understanding, regret, or the capacity to choose otherwise. These are qualities that

---

635 Kingsley O Mrabure and Alfred Abhulimhen-Iyoha, 'A Comparative Analysis of Corporate Criminal Liability in Nigeria and Other Jurisdictions' (2020) 11 Beijing Law Review 429.

636 Muzaffer Eroğlu and Meltem Karatepe Kaya, 'Impact of Artificial Intelligence on Corporate Board Diversity Policies and Regulations' (2022) 23 European Business Organization Law Review 541.

637 Yoshija Walter, 'Managing the Race to the Moon: Global Policy and Governance in Artificial Intelligence Regulation—A Contemporary Overview and an Analysis of Socioeconomic Consequences' (2024) 4 Discover Artificial Intelligence 14.

638 Kevin Cole, 'Knowledge and Belief as Criminal Law Mental States' (2018) 16 Ohio State Journal of Criminal Law <https://heinonline.org/HOL/Page?handle=hein.journals/osjcl16&id=453&div=&collection=> accessed 22 September 2025.

639 Angelo J Golia, 'The Death of the Legal Subject: How Predictive Algorithms Are (Re)Constructing Legal Subjectivity' [2022] Global Constitutionalism <https://papers.ssrn.com/abstract=4047345> accessed 22 September 2025; Katrina Geddes, 'The Death of the Legal Subject' (2023) 25 Vanderbilt J Entertain Technol Law <https://heinonline.org/HOL/Page?handle=hein.journals/vanep25&id=11&div=&collection=> accessed 22 September 2025.

underpin desert-based punishment. Treating algorithmic optimization as intent risks reducing criminal law to a system of outcome-based liability, thereby abandoning its retributive core in favor of pure social control.
It is worth noting that the European Parliament has floated the idea of "electronic personhood" for robots. A 2017 EU report suggested granting sophisticated autonomous robots an "electronic person" status.[640] This would make them legal entities rather than property. However, this idea is still only a recommendation and has not been adopted. Currently, criminal law does not recognize AI as a person, which creates a conceptual gap.

Legal Frameworks and Gaps

To date, no country has a clear law holding AI itself criminally liable. Instead, legal regimes attribute crimes involving AI to humans. For example, if a self-driving car causes a death, police will investigate the driver or manufacturer, not prosecute the car. Civil law is ahead in this regard. Many jurisdictions are revising liability rules to help victims recover damages. The EU, for instance, is updating its Product Liability Directive and proposing an AI Liability Directive.[641] The Parliament has even recommended a common strict liability regime for operators of what it classed as high-risk AI systems and requires them to have insurance or pay for harm.[642] But these are civil remedies (compensation), not criminal punishment. This civil-criminal divide is legally significant. Civil liability prioritizes loss distribution and victim compensation, tolerating strict or no-fault models. Criminal law, however, is anchored in moral blameworthiness. Imposing criminal penalties without fault, such as prosecuting a programmer for an AI's unforeseeable action, would violate the principle that punishment must be deserved. Thus, while civil strict liability may be appropriate for AI harms, its extension into criminal law would represent not innovation but erosion of a foundational safeguard. Criminal law generally adapts old concepts to new tech. In practice, some theories are applied:

---

640 Alekhine (n 15).

641 Tambiama Madiega, 'BRIEFING EU Legislation in Progress Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)'.

642 'Carriages Preview | Legislative Train Schedule' <https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-liability-directive> accessed 22 September 2025.

Vicarious or strict liability: One approach is to hold companies or humans strictly liable for AI harms. The EU Parliament's stance is an example of this mindset. However, legal scholars, such as Levenson,[643] view strict criminal liability skeptically (punishing the innocent) unless guided by safety regulations.

***Negligence and foreseeability*:** Courts may treat a programmer's or owner's failure to foresee AI's risk as negligence. If a reasonable person prevented the harm, criminal negligence could apply. This is like manslaughter or involuntary conduct charges. The gap is that AI learning can produce truly unpredictable outcomes, [644] making it hard to prove a human should have foreseen a specific harm. This unpredictability is not just a practical evidentiary hurdle; it challenges the very premise of negligence-based criminal liability. Negligence assumes a baseline of human reasonableness, but when even experts cannot predict an AI's behavior, holding someone criminally liable for "failing to foresee" becomes a form of constructive fault. Such an approach may deter innovation without enhancing justice, revealing a mismatch between traditional culpability models and the epistemic limits of AI development.

***Existing precedents*:** There are few direct precedents. Some road traffic cases have held drivers responsible when using autopilot.[645] According to Morales and Vhon,[646] vehicles driven using Autopilot recorded one crash per 7.08 million miles. In contrast, drivers who did not use Autopilot experienced an accident every 1.29 million miles. But no court has blamed an autonomous system independently. Internationally, organizations like UNESCO emphasize human accountability in AI ethics, but no global treaty specifies AI criminality. Therefore, current frameworks leave many

[643] Laurie L Levenson, 'Good Faith Defenses: Reshaping Strict Liability Crimes' (1992) 78 Cornell Law Rev <https://heinonline.org/HOL/Page?handle=hein.journals/clqv78&id=439&div=&collection=> accessed 22 September 2025.

[644] Roman V. Yampolskiy, 'Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent' (2020) 07 Journal of Artificial Intelligence and Consciousness 109.

[645] 'Tesla Autopilot Accidents: Legal Rights & Liability Explained' <https://teamjustice.com/tesla-autopilot-accidents-legal-rights-liability/> accessed 22 September 2025.

[646] Morales and Vhon (n 4).

uncertainties. Many jurisdictions effectively say if AI does something wrong, a human is to blame.[647]

Legal precedents, while not directly addressing AI, offer useful analogies for framing future accountability. For instance, product liability cases that involve defective software or autonomous systems, such as rulings on autopilot-related accidents in the United States and Europe, have sometimes extended manufacturer responsibility beyond traditional negligence standards.[648]. Similarly, there is the angle of court decisions on corporate criminal liability, where a legal entity is held accountable for the acts of its agents.[649] This illustrates how non-human actors can be attributed intent and fault through human intermediaries. Precedents in cybercrime prosecutions, where intent is inferred from indirect digital actions,[650] also provide models for assessing culpability when direct human control is absent. Although none of these precedents perfectly map onto AI's autonomous decision-making, they collectively demonstrate the judiciary's capacity to adapt legal doctrines to novel technological contexts.

Challenges of Intention and Causation

Mens rea (guilty mind) is central to criminal law. The big question is, can an AI ever form intent? In human crimes, intent often requires goals, beliefs, or emotions. AI can have "goals" programmed or learnt but lacks consciousness. Still, some argue AI can fulfill the functional criteria for knowledge and intent. For example, the Saudi/Iran study notes that knowledge (awareness of facts) is possible for AI through data input.[651] They argue that specific intent might correspond to an AI's goal-setting process. In other words, if an AI system "knowingly and intentionally" pursues an outcome (like delivering a lethal strike), that could mirror

[647] Daniel B Shank, Alyssa DeSanti and Timothy Maninger, 'When Are Artificial Intelligence versus Human Agents Faulted for Wrongdoing? Moral Attributions after Individual and Joint Decisions' (2019) 22 Inf Commun Soc 648 <https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1568515> accessed 22 September 2025; Jordan Richard Schoenherr and Robert Thomson, 'When AI Fails, Who Do We Blame? Attributing Responsibility in Human–AI Interactions' (2024) 5 IEEE Transactions on Technology and Society 61.

[648] Miriam C Buiten, 'Product Liability for Defective AI' (2024) 57 European Journal of Law and Economics 239.

[649] Mrabure and Abhulimhen-Iyoha (n 21).

[650] Paul Arnell and Bukola Faturoti, 'The Prosecution of Cybercrime – Why Transnational and Extraterritorial Jurisdiction Should Be Resisted' (2023) 37 International Review of Law, Computers & Technology 29.

[651] Meghdadi and Pourcheriki (n 5).

human intent. However, the main obstacle is human-like psychology. AI cannot feel hate or revenge – emotions often underlie crimes. Abbott and Sarch[652] point out that AI lacks the strong emotions (love, hatred, and jealousy) that drive many offences. Nevertheless, they also note emotions are not required elements for most crimes. This suggests that absence of emotion alone should not bar finding intent. The bigger issue is proof. Currently, a court cannot easily ask an AI what it intended.

Chain of causation is another challenge. Suppose an AI acts autonomously and causes harm. How do we trace responsibility? One model, according to Hallevy,[653] is the *perpetration-by-another* approach. Here, the AI is an innocent agent like a child or disabled person. The law then treats the human who set the scheme in motion as the true perpetrator. For instance, the law holds a programmer and user liable if they create a violent AI or use it to harm someone. The Iranian/Saudi study explains that under this model, the programmer and the user are fully criminally liable, while the AI entity itself bears no criminal liability.[654] In this context, AI serves as a tool, and the human operator bears the responsibility.

This model works if the crime was intended or foreseeable by the human. But a problem arises when the AI acts with its own learnt logic. Hallevy[655] proposes scenarios where AI decides to commit a crime based on its accumulated experience and is not programmed to perform criminal acts. If no human programmed or intended that outcome, it becomes hard to pin blame on anyone. This "accountability gap" is not merely a policy concern but a doctrinal crisis. Criminal law assumes a continuous chain of human agency, from decision to action. Autonomous AI fractures this chain, creating scenarios where harm occurs without any culpable human actor. The perpetration-by-another model, while useful in cases of intentional misuse, collapses when the AI's action is neither intended nor foreseeable. In such cases, forcing attribution to a human risks punishing the innocent, which is a direct violation of the presumption of innocence and the requirement of personal culpability. In this case, the chain of cause is severed, and the AI's decision emerges internally. Attribution models try to address this. According to Judicial Council of California Criminal Jury

652 Abbott and Sarch (n 10).

653 Gabriel Hallevy, 'AI vs. IP : Criminal Liability for Intellectual Property Offences of Artificial Intelligence Entities' [2020] Artificial Intelligence and the Law 222 <https://www.taylorfrancis.com/chapters/edit/10.4324/9780429344015-10/ai-vs-ip-gabriel-hallevy> accessed 22 September 2025.

654 Meghdadi and Pourcheriki (n 5).

655 Hallevy (n 39).

Instructions[656], natural-probable-consequence (akin to accomplice liability) holds a human liable if a crime is the foreseeable result of their actions, even without intent. For example, if an engineer knows his AI can accidentally kill but does nothing to prevent it, it argues he could be negligent. But proving foreseeability in novel AI cases is tricky.

In addition to intent, punishment and capacity are also important issues. Criminal sanctions usually aim to deter or reform a human being or to incapacitate them.[657] As seen in the discussion above, AI cannot feel pain or regret. If one were to punish an AI, the usual penalties (prison, fines) would be meaningless for a machine. The only available sanctions are material – deactivate it, wipe its memory, or seize resources. Though it might be questioned whether these count as true punishment or just disabling a hazard. This distinction is crucial. Disabling a hazard is a regulatory or public safety measure; punishment is a moral judgment. If criminal law sanctions against AI serve only the former function, they lose their expressive and retributive meaning. This suggests that any "punishment" of AI must be carefully distinguished from traditional criminal penalties (maybe reclassified as administrative sanctions) to preserve the integrity of criminal justice.

Proposed models and responses

Given these challenges, scholars and policymakers have proposed various models:

a) ***Human agent focus:*** Most suggestions reinforce human accountability. Some have called for clearer laws stating that AI's creators or users are responsible for its criminal use.[658] For instance, one recommendation is to ensure designers bear criminal responsibility whether acting alone or jointly with AI. Strengthening negligence and

656 'CALCRIM No. 402. Natural and Probable Consequences Doctrine (Target and Non-Target Offenses Charged) :: California Criminal Jury Instructions (CALCRIM) (2025) :: Justia' <https://www.justia.com/criminal/docs/calcrim/400/402/> accessed 22 September 2025.

657 Hsin Wen Lee, 'Taking Deterrence Seriously: The Wide-Scope Deterrence Theory of Punishment' (2017) 36 Crim Justice Ethics 2 <https://www.tandfonline.com/doi/abs/10.1080/0731129X.2017.1298879> accessed 22 September 2025.

658 Meghdadi and Pourcheriki (n 5).

corporate liability rules could close gaps because programmers might face stricter duties to foresee and prevent AI harms.

b) ***Electronic personhood*****:** Other proposals go further by treating AI as its own legal person. Ellamey and Elwakad[659] argue for granting AI an "electronic legal personality" so it can "bear civil and criminal responsibility" for its acts. This would be analogous to corporate personhood. Under this model, an autonomous AI could in principle be charged (though in practice it might face asset forfeiture, disabling orders, or other sanctions suited to machines). Al Mulla,[660] similarly recommends recognizing a machine's legal personality and creating a punishment system that match the nature of the smart machine. The advantage is that it fills the accountability gap directly. The clear weakness is philosophical. It treats an object as a moral agent, raising the conceptual issues noted earlier. It would also require a fundamental legal reform across many areas of law.

Strict liability and insurance**:** The EU Parliament has suggested strict liability for what it termed "high-risk AI" under civil law.[661] A comparable idea for criminal law would be to mandate insurance or equivalent financial penalties for AI producers. If an AI crime occurs, the insurance fund could compensate victims (just like in corporate fines), and the human beneficiaries could still face action. This system does not punish the AI directly but enforces deterrence. However, a drawback is that this approach still relies on human enforcement and may not adequately address questions of intent.

Preventive regulation: Some experts emphasize preventing crime through design. Measures like mandatory audits, transparency requirements, and even an AI registry have been proposed. For example, Morales[662] envisions a public "AI Registry" where high-risk systems are recorded and monitored, along with mandatory algorithmic audits to trace decisions. This would aid accountability after the fact. It reflects a model where regulation and oversight, rather than traditional crime-fault logic,

---

659 Yasser Ellamey and Amr Elwakad, 'THE CRIMINAL RESPONSIBILITY OF ARTIFICIAL INTELLIGENCE SYSTEMS: A PROSPECTIVE ANALYTICAL STUDY' (2023) 5 Corporate Law & Governance Review 2023 <https://doi.org/10.22495/clgrv5i1p8> accessed 22 September 2025.

660 Andrea Roth, 'Trial by Machine' (2015) 104 Georgetown Law Journal <https://heinonline.org/HOL/Page?handle=hein.journals/glj104&id=1259&div=&collection=> accessed 22 September 2025.

661 Madiega (n 27).

662 Morales and Vhon (n 4).

keep AI aligned with law. The downside is it requires adequate governance infrastructure and does not solve the core question of criminal blame when prevention fails.

***Philosophical criteria*:** Morales[663] also suggests a framework of four principles (volition, intent, culpability and autonomy [VICA]—to gauge AI's role. Under this view, an AI would only be held accountable to the extent it meets thresholds of autonomous decision-making and culpability, even if partial. This hybrid model would blend elements of personhood with human oversight. If the AI acted almost on its own (high autonomy and intent), we might lean towards something like direct liability; if not, responsibility shifts to people. This sophisticated approach is largely theoretical at present, but it shows how philosophical models could inform legal rules. Taken together, these models reveal a fundamental trade-off: human-centered approaches preserve doctrinal coherence but risk injustice in high-autonomy cases; AI-centered models fill accountability gaps but threaten to hollow out the moral foundations of criminal law. The path forward likely lies not in choosing one over the other but in a tiered framework: human liability for foreseeable or controllable AI behavior and system-level regulatory sanctions (not criminal punishment) for harms arising from genuine autonomy. Such a hybrid model respects both justice and practicality. It anchors blame where moral agency exists, while ensuring no harm goes entirely unaddressed.

In summary, we note here that no single solution is settled. The literature ranges from keeping all liability on humans to creating a new category of AI offender. Each model has trade-offs. What is clear is that the law must reckon with AI's unique nature. As Morale warns, if AI is permitted to act with impunity, law itself becomes obsolete. Adapting criminal law will likely require a mix of clarifying human duties, updating legal personhood concepts and embedding technical safeguards in the AI design.

Conclusion

Artificial intelligence challenges the foundations of criminal law. Our analysis shows that traditional legal concepts struggle with autonomous AI. AI systems cannot fit neatly into the old model of a person with a guilty mind. Most current laws see the machine as an innocent agent and hold some human responsible. But this can leave serious harms

[663] ibid.

unaccounted for or unfairly blamed on people who did not intend the outcome. Scholars have therefore proposed radical changes. Granting AI an "electronic personality" so it can be charged directly, or imposing strict liability on operators with insurance, are examples. These ideas reflect a tension between retributive notions of blame (is it fair to call a machine a criminal?) and utilitarian needs (how else to deter and compensate for high-tech harms?). If machines can develop knowledge and goal-seeking behavior, maybe a new legal theory of "technological individuals" is needed. But many experts still favor refining human culpability over punishing code.

Finally, the law cannot remain static. It must evolve with contemporary events and technology. This means rethinking who or what counts as a legal actor and how intention is assigned. The substantive issues above – intent, causation, personhood – will be at the heart of future reforms. As AI grows more capable, the need for clear rules becomes urgent. Without adaptation, either wrongful acts will go unchecked or the wrong parties will be punished, both of which undermine justice.

Recommendations

To prepare the legal system for AI, we suggest:

1. ***Clarifying human accountability*:** It is necessary to update criminal statutes or guidelines to state clearly that AI actions are attributed to identifiable humans (developers, operators, owners) when those humans could foresee or control the AI's use. This prevents gaps where no one is blamed.

2. ***Considering regulated personhood*:** There is a need to conduct careful study on limited electronic legal personality for advanced AI. For high-autonomy systems, create a legal status (with tied assets or insurance) that can be penalized (e.g., through fines or mandates).

3. ***Mandatory safeguards and insurance:*** Mandatory insurance or financial reserves for high-risk AI, similar to liability insurance, are required. It is also advantageous to impose strict safety standards (certification, testing) so that AI developers must minimize dangerous autonomy.

4. ***Increasing transparency and oversight:*** An AI registry and audit system should be established. Developers should log key design details and

decision data, and independent auditors should have authority to examine AI systems, especially after an incident.

5. ***Training legal authorities:*** Judges, prosecutors and police should be educated about AI capabilities and risks. Again, it is important to create expert advisory bodies that can analyze AI behaviors in criminal cases.

6. ***International cooperation:*** It is crucial to work with other countries and international bodies to harmonize AI liability rules. Shared principles (through the UN, OECD, or other forums) can prevent "forum shopping" and ensure that AI-related crimes are addressed globally.

References

Abbott, R. B. & Sarch, A. F. (2019, February 1). Punishing Artificial Intelligence: Legal Fiction or Science Fiction. *53 UC Davis Law Review 1*(323) http://dx.doi.org/10.2139/ssrn.3327485

Al Mulla, M. S. (2024). The Historical Dimensions of the Evolution of the Theory of Criminal Liability and the Dialectic of its Application in the Era of Artificial Intelligence: An Analytical and Prospective Study. *Kuwait International Law School Journal, 1*(49).

Arnell, P., & Faturoti, B. (2022). The prosecution of cybercrime – why transnational and extraterritorial jurisdiction should be resisted. *International Review of Law, Computers & Technology*, *37*(1), 29–51. https://doi.org/10.1080/13600869.2022.2061888

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, *8*(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

Bigenwald, A., & Chambon, V. (2019). Criminal Responsibility and Neuroscience: No Revolution Yet. *Frontiers in psychology*, *10*, 1406. https://doi.org/10.3389/fpsyg.2019.01406

Brown, R. D. (2020). Property ownership and the legal personhood of artificial intelligence. *Information & Communications Technology Law*, *30*(2), 208–234. https://doi.org/10.1080/13600834.2020.1861714

Buiten, M. C. (2024). Product liability for defective AI. *European Journal of Law and Economics*, *57*(1), 239-273. https://doi.org/10.1007/s10657-024-09794-z

Christensen, D. (2025, February 24). *Tesla Autopilot Accidents: Legal Rights and Liability. DJC Law*. https://teamjustice.com/tesla-autopilot-accidents-legal-rights-liability/

Cole, K. (2018). Knowledge and Belief as Criminal Law Mental States. *Ohio St. J. Crim. L.*, *16*, 441.

Cornford, A. (2024). The aims and functions of criminal law. *The Modern Law Review*, *87*(2), 398_429. https://doi.org/10.1111/1468_2230.12846

Ellamey, Y., & Elwakad, A. (2023). The criminal responsibility of artificial intelligence systems: A prospective analytical study. *Corporate Law and Governance Review*, *5*(1), 92-100.

Eroğlu, M., & Karatepe Kaya, M. (2022). Impact of artificial intelligence on corporate board diversity policies and regulations. *European Business Organization Law Review*, *23*(3), 541 572. https://doi.org/10.1007/s40804-022-00251-5

European Parliament (2022, September 29). *Artificial intelligence liability directive*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf#:~:text=Parliament%20recommended%20setting%20up%20a,INI.

European Parliament (2025, June 20). *AI liability directive: A new plan for Europe's sustainable prosperity and competitiveness*. https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-liability-directive

Fauquet-Alekhine, F. (2022). Can the Robot be Considered a Person? The European Perspective. *Advances in Research on Teaching*, *23*(6), 100-105.

Ferguson, F., & Boyd, B. (2023, November 21). *Classical school of criminology: Overview, theory & principles*. Study.com. https://study.com/academy/lesson/the-classical-school-of-criminology-its-influence-today.html

Geddes, K. (2022). The Death of the Legal Subject: How Predictive Algorithms Are (Re) constructing Legal Subjectivity. *Vanderbilt Journal of Entertainment & Technology Law*, *25*.

Geddes, K. (2023). The death of the legal subject. *V and. J. Ent. & Tech. L.*, *25*, 1.

Hallevy, G. (2019, June 11). *The basic models of criminal liability of AI systems and outer circles*. https://ssrn.com/abstract=3402527.

Hallevy, G. (2021). AI vs. IP, criminal liability for intellectual property offences of artificial intelligence entities. In D. J. Baker & P. H. Robinson (Eds.), *Artificial intelligence and the law: Cybercrime and criminal liability* (pp. 231–239). Routledge.

InsideEVs (2025, April 29). *Only 1 crash for every 7.44 million miles driven using Tesla Autopilot in Q1 2025*. https://www.cybertruckownersclub.com/forum/threads/only-1-crash-for-every-7-44-million-miles-driven-using-tesla-autopilot-in-q1-2025.40936/

Judicial Council of California Criminal Jury Instructions (2025). *CALCRIM No. 402. Natural and Probable Consequences Doctrine (Target and Non Target Offenses Charged)*. JUSTIA. https://www.justia.com/criminal/docs/calcrim/400/402/

Kurki, V. A (2019, August 13). Theory of Legal Personhood. A Theory of Legal Personhood. *Helsinki Legal Studies Research Paper, 58*. https://ssrn.com/abstract=3446600

Laitinen, A., & Sahlgren, O. (2021). AI Systems and Respect for Human Autonomy. *Frontiers in artificial intelligence*, *4*, 705164. https://doi.org/10.3389/frai.2021.705164

Lee, H. W. (2017). Taking deterrence seriously: The wide-scope deterrence theory of punishment. *Criminal justice ethics*, *36*(1), 2-24.

Levenson, L. L. (1993). *Good Faith Defenses: Reshaping Strict Liability Crimes. Cornell Law Review, 78*(3). http://scholarship.law.cornell.edu/clr/vol78/iss3/2

Li, W., Song, R., Zhang, B., & Yu, K. (2025). AI Creativity and Legal Protection for AI-Generated Works in Posthuman Societal Scenarios. *Sustainable Futures, 9*. 100749. https://doi.org/10.1016/j.sftr.2025.100749

Masood, A. (2025, April 4). *Intellectual property rights and AI-generated content — issues in human authorship, fair use doctrine, and output liability*. Medium. https://medium.com/@adnanmasood/intellectual-property-rights-and-ai-generated-content-issues-in-human-authorship-fair-use-8c7ec9d6fdc3

Meghdadi, Z., & Pourcheriki, M. (2025). Comparative Study of Criminal Responsibility of AI in the Legal Framework of Iran and Saudi Arabia. *Journal of Innovations in Computer Science and Engineering, 2*(1), 87-93.

Morales, J. V. A. (2025). *When machines break the law: the case for criminal liability & legal personhood for artificial intelligence*. Emmanuel Q. Fernando Undergraduate Conference. https://philosophy.upd.edu.ph/wp-content/uploads/2025/08/When-Machines-Break-The-Law-Morales.pdf

Mrabure, K., & Abhulimhen-Iyoha, A. (2020) A Comparative Analysis of Corporate Criminal Liability in Nigeria and Other Jurisdictions. *Beijing Law Review*, 11, 429-443. https://doi.org/10.4236/blr.2020.112027.

Nay, J. J. (2022). Law informs code: A legal informatics approach to aligning artificial intelligence with humans. *Nw. J. Tech. & Intell. Prop.*, *20*, 309.

Presno Linera, M. Á., & Meuwese, A. (2025). Regulating AI from Europe: a joint analysis of the AI Act and the Framework Convention on

AI. *The Theory and Practice of Legislation*, 1_20. https://doi.org/10.1080/20508840.2025.2492524

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy Technology. 34*, 1057_1084. https://doi.org/10.1007/s13347-021-00450-x

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial Intelligence: Definition and Background. In: *Mission AI. Research for Policy*. Springer. https://doi.org/10.1007/978-3-031-21448-6_2

Simmler, M., Canova, G., & Schedler, K. (2021). Smart criminal justice: Phenomena and normative requirements1. *International Review of Administrative Sciences*, *89*(2), 415_432. https://doi.org/10.1177/00208523211039740

Stasi, A. (2021). Actus Reus and Mens Rea. In *General Principles of Thai Criminal Law* (pp. 25-30). Singapore: Springer Singapore.

Walter, Y. (2024). Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation – A contemporary overview and an analysis of socioeconomic consequences. *Discover Artificial Intelligent,* 4(14). https://doi.org/10.1007/s44163-024-00109-4

Yampolskiy, R. V. (2020). Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent. *Journal of Artificial Intelligence and Consciousness*, *7*(01), 109_118. https://doi.org/10.1142/S2705078520500034.