HUMAN-IN-THE-LOOP JUSTICE: CRIMINAL RESPONSIBILITY AND MENS REA IN AI-MEDIATED OFFENSES

Gauri Lamba, Student, BBA LL.B. (Hons.), Symbiosis Law School, Pune

ABSTRACT

In everyday language, the current law does not really treat AI as a criminal in itself. Instead, it sees AI as a powerful tool used by people and organizations, sometimes legally and sometimes illegally. When issues arise, courts still ask the traditional questions: Who created this system? Who decided to use it this way? Who wrote the prompt that led to such an outcome? And who should have stepped in to prevent harm but did not? Prosecutors and defense lawyers then debate these answers, citing examples such as algorithmic sentencing tools to discuss fairness, responsibility, and what constitutes "intent" when a machine is involved.

This paper explores the challenges of incorporating artificial intelligence (AI) into criminal law, especially concerning liability, intent, and accountability. It argues that AI systems, although capable of causing harm, cannot be considered independent criminal actors because they lack consciousness and moral agency. Instead, AI should be viewed as a tool of crime, with responsibility ultimately traceable to the human actors involved in its design, deployment, and use. The paper focuses on three main issues: whether AI systems can commit crimes, how to establish criminal intent (mens rea) in cases involving algorithms, and how the roles of prosecutors and defense attorneys are changing in these cases. Using principles of criminal law, comparative case studies, and practical legal considerations, it shows that traditional doctrines of mens rea and causation can be adapted to AI-related contexts by examining human knowledge, foreseeability, and control. The conclusion is that the future of criminal justice in the age of AI depends on maintaining human accountability while updating legal frameworks to address the complexities of automated systems' decision-making.

Key Words: *- Criminal Liability, AI Accountability, Criminal Instrumentality, Algorithmic Bias, Legal Personhood, Automated Offences, AI Evidence, Prosecutorial Strategy, Defence Strategy, Cybercrime, Data Governance, Human-in-the-Loop*

I. CAN AI SYSTEMS COMMIT CRIMES?

With the advent of artificial intelligence across every sector today, concerns about accountability are of paramount importance. This paper examines projected trends in the accountability of artificial intelligence (AI) over the next five to ten years. The paper argues that accountability will shift from abstract principles to measurable, auditable practices, driven by a combination of hard regulation, market incentives, and technological advances that enable traceability and oversight. This paper maps projected trends that will shape how governments, companies, and civil society hold AI systems and actors accountable[1]. The paper delves into the concepts of reliability and Accountability[2] that an entity must demonstrate for the parts it plays in the project it has worked on[3]. Accountability also requires that the results of this work are verifiable and not hallucinated.

AI accountability means that there is always someone answerable for what an AI does, especially when it causes harm. It rejects the idea that "the algorithm did it" and insists that creators, programmers, and actors remain responsible for outcomes they may cause.
AI accountability focuses on ensuring AI systems are transparent[4], auditable, and responsible for their decisions. Projections showing rapid market growth and regulatory evolution through 2030. Responsible AI markets are expected to expand from $1.09 billion in 2024 to $10.26 billion by 2030 at a 45.3% CAGR, driven by governance tools and ethical standards[5]. This paper examines key projected trends, regulatory shifts, technological advancements, and challenges.

The urgency of holding actors' accountable stems from the profound impact of AI on everyday life. Algorithms based on recommendation models[6] help determine the most basic aspects of our daily lives, such as what news people see and which advertisements they encounter. As a result, accountability is shifting from being "nice to have" to a legal necessity[7].

[1]Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).
[2]H.L.A. Hart & Tony Honoré, *Causation in the Law* (2nd edn., Oxford University Press 1985) 1–25.
[3]David Leslie et al., *AI Accountability in Practice* (Alan Turing Institute 2024).
[4]David Leslie, *Understanding Artificial Intelligence Ethics and Safety* (Alan Turing Institute 2019).
[5]MarketsandMarkets, *Responsible AI Market – Global Forecast to 2030* (2024).
[6]Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press 2018).
[7]Danielle Keats Citron, "Technological Due Process" 85 *Washington University Law Review* 1249 (2008).

Artificial Intelligence is gaining indulgence[8] in diverse professional fields. This rapidly advancing technology offers users a range of benefits, enabling them to work more efficiently and produce high-quality output. This has been made possible through advancements in machine learning [ML], which is the core method for building AI systems that can make independent decisions. However, it has often been observed that AI suffers from a lack of accountability[9] due to the large language models and large sample data feeding, which in turn affects the output data it generates. This can lead to inaccuracies and inefficiencies, which subsequently degrade productivity and yield inaccurate outputs. One of the ten OECD AI Principles[10] refers to the accountability that AI actors bear for the proper functioning of the AI systems they develop and deploy[11]. This means that AI actors must take measures to ensure their AI systems are trustworthy, transparent, and safe to use[12]. To achieve this, actors need to manage risks throughout the lifecycle of their AI systems, from planning to data collection and processing, model building and validation, to deployment, operation, and monitoring[13].

According to the OECD AI Principles, trustworthy[14] AI refers to systems that promote human well-being and sustainable development, respect human rights, fairness, democracy, and the rule of law, and remain transparent, explainable, robust, secure, and safe throughout their lifecycle. AI actors are expected to ensure accountability for the proper functioning of AI systems and enable human oversight and redress where necessary.[15]

AI systems create accountability gaps, where decisions are made by automated systems that cannot be legally held accountable. AI systems operate through complex technical processes, making it challenging to determine who should be held responsible when something goes wrong. Responsibility is spread across several human actors, including system designers, developers, and policymakers. AI systems are often described as "black boxes" because their decision-making processes are difficult for humans to fully understand or explain[16]. This lack of transparency has direct implications for AI accountability, as decisions are generated through

[8]S. Chakravarthy Naik, "Exploring the Role of Technology in Ensuring Accountability of Artificial Intelligence" 2 *IPR Journal of Maharashtra National Law University Nagpur* 76 (2024).
[9]OECD, *OECD Principles on Artificial Intelligence* (2019).
[10]OECD, *Advancing Accountability in Artificial Intelligence*, supra note 2.
[11]OECD, *OECD Principles on Artificial Intelligence* (2019).
[12]OECD, *OECD Principles on Artificial Intelligence*, supra note 11.
[13]OECD, *Advancing Accountability in Artificial Intelligence*, supra note 2.
[14]OECD, *OECD Principles on Artificial Intelligence*, supra note 11.
[15]OECD, *Advancing Accountability in Artificial Intelligence*, supra note 2.
[16]David Leslie et al., *AI Accountability in Practice*, supra note 3, at 11–12.

complex machine-learning models, such as deep neural networks and large language models, whose reasoning is not traceable. Since the basis of an AI-driven outcome cannot be clearly explained, it becomes challenging to identify responsibility, provide justification, and address the concerns of those affected by such decisions[17].

GLOBAL FRAMEWORKS-

1. Artificial Intelligence Act of the European Union (2024–2026)

AI is categorized as unacceptable, high, limited, or minimal under risk-based regulation. AI is defined in Article 3. High-risk systems are subject to responsibilities in Chapter IV (transparency, human oversight, record-keeping, accuracy), Article 65–70 enforcement, and Article 71 sanctions. Prohibits biased social scoring and biometric mass monitoring. An international standard for AI governance that is safety-driven[18].

2. AI Principles of the OECD (2019)[19]

First basis for international soft law. Focuses on human-centric AI, sustainability, accountability, transparency, robustness, fairness, and privacy. Emphasizes the need for explainability and risk minimization. Adopted by more than 40 countries, it is non-binding but serves as the foundation for national legislation, such as the U.S. algorithmic risk standards and the EU AI Act.[20].

3. UNESCO's 2021 AI Ethics Recommendation

One hundred ninety-three states have accepted a universal human rights-based AI guideline. Articles 10 (non-discrimination & fairness), 17 (data governance), 21 (transparency), and 23 (accountability) are the focus areas. Requires public decision systems to conduct bias audits, explainability, environmental sustainability, and AI effect evaluations[21].

4. The proposed U.S. Algorithmic Accountability Act

mandates bias testing, risk reporting, algorithmic impact assessments (AIA), and transparency requirements for AI applied in employment, housing, credit, and law enforcement. Extends anti-discrimination legislation to automated systems in accordance with the Civil Rights Act, Fair Credit Reporting Act, and Americans with Disabilities Act. FTC enforcement.

5. Framework Convention on AI, Council of Europe (2024 Draft)

The first international treaty-style mechanism to align AI with democracy, human rights,

[17]David Leslie et al., *AI Accountability in Practice*, supra note 3.
[18]Regulation (EU) 2024/___ (Artificial Intelligence Act), arts. 3, 5, 9–15, 65–71
[19]OECD, *OECD Principles on Artificial Intelligence*, supra note 11.
[20]OECD, *OECD Principles on Artificial Intelligence*, supra note 11.
[21]UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (2021).

and the rule of law. Oversight organizations, corrective measures, and harm-prevention responsibilities are among the provisions. Focuses on due process rights, risk assessment, bias governance, data protection, and auditability[22].

6. The Digital Personal Data Protection Act of 2023 in India[23]

Addresses user consent, data processing, and privacy obligations. AI is not directly regulated; however, it serves as the foundation for standards related to data fairness and openness. Accountability, risk assessment, explainability, and bias auditing methods are anticipated to be covered by India's future AI regulatory mission.

AI Accountability using Theoretical Foundation-

I. Hart and Honore's Theory of Causation-

- They say that legal causation isn't a technical concept but relies on common sense, which is actually rooted in plain man's distinction from everyday life, which contains responsibility. It focuses on Normal and Abnormal facts, Intervening facts, and Foreseeability and human agency v. natural events[24].
- To link theory of causation with AI accountability- AI involves long, opaque, causal chains, which involve training data, model design, system changes, and end users' decisions that are common in AI systems. But according to their theory:
- Abnormal events may include hallucinations, biases, or malfunctions in the AI system.
- The actor with control and predictability, typically the developer or deployer, remains the primary cause. Since AI is not an autonomous moral agent, its autonomy doesn't break the causal chain.

II. Joel Feinberg's Harm Principle[25]

- According to him, actions that create injury or unjustified danger of harm to others are grounds for state involvement, even if there is no malice; still,[26] this principle is applicable.

[22]Council of Europe, *Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law* (Draft 2024).
[23]Digital Personal Data Protection Act, 2023 (Act 22 of 2023).
[24]H.L.A. Hart & Tony Honoré, *Causation in the Law* (2nd edn., Oxford University Press 1985) 1–25.
[25]Joel Feinberg, *Harm to Others* (Oxford University Press 1984) 9–26.
[26]John Rawls, *A Theory of Justice* (Revised edn., Harvard University Press 1999) 60–75.

- When utilizing AI, its program produces Risks that are predictable, like discrimination, bias, etc. Risks of high magnitude, such as autonomous vehicles, medical issues, and diffuse dangers, like surveillance or privacy infringement.
- But as per Harm theory (State Intervention)- Regulation is justified solely by the size and predictability of the risk. When private actors utilize technologies that could cause widespread harm, the state has a responsibility to intervene.

III. Rawlsian's Fairness Principle-

- The Justice Principle of John Rawls says Equal Fundamental freedom for all, Equitable Access to opportunities, and the difference principle states that disparities are only permissible if they help the least fortunate.
- When utilizing an AI-algorithmic system, it discriminates against groups that are protected, strengthens historical prejudices, decreases equitable opportunities in financing, employment, and law enforcement, and more severely harms vulnerable groups.
- When applying this theory, AI must not make the situation of underprivileged populations worse. Transparency and bias audits are required by law and morality. Equal opportunity must be improved, not diminished, by automated decision-making.

RELEVANT CASE-

Mata v. Avianca (S.D.N.Y. 2023)[27]

This includes a personal injury claim in which the plaintiff's attorneys submitted a brief containing six fictitious case citations generated by ChatGPT. Instead of confirming the authorities, counsel doubled down when questioned. Judge Castel found subjective bad faith and imposed Rule 11 sanctions, including a $5,000 fine and correction letters to affected judges, after ruling that reliance on AI did not absolve a professional of their duties. The case became a global legal and ethical landmark, indicating the need for independent verification of AI-assisted research.

As of 2025, the case continues to set a precedent for attorney liability in cases.[28] Involving AI misuse.

[27]*Mata v. Avianca, Inc.*, 2023 WL 4114965 (S.D.N.Y. 2023).

A. CRIMINAL CAPACITY AND LEGAL PERSONHOOD

Artificial intelligence systems, no matter how intelligent they appear, cannot be treated as criminals undercriminal law. The foundation of criminal liability rests on two key elements: actus reus (the physical act of committing a crime) and mens rea (the guilty mind or intent behind it). For someone to be legally accountable for their actions, they must have both the capacity to act and the conscious awareness of their wrongdoing[29].

AI, by contrast, lacks thesehuman qualities. It lacks consciousness, emotions, and moral agency. It does not make choices in the way people do; instead, it follows patterns, data, commands, and programming. Due to this, the law does not recognize AI systems as independent legal actors capable of forming criminal intent or bearing punishment. In simple terms, a machine cannot be held accountable for committing a crime in the way a person can[30].

But that does not mean AI is irrelevant in criminal law. History shows that harm can be caused through tools or agents that themselves lack moral awareness, whether it's a dangerous animal, a faulty machine, or an automated process. In those cases, the law turns its focus to the human beings who designed, deployed, or controlled these tools. Similarly, when AI systems cause harm or contribute to an offence, responsibility ultimately falls on the humans behind them, such as the programmers, operators, or organizations that set them in motion.

Thus, AI does not fit into the category of "criminal subject," but rather "criminal instrumentality." The real challenge for modern criminal law lies in tracing responsibility through this chain of human and machine actions, ensuring that accountability remains both just and practicable in an age of intelligent technology.

B. AI AS A CRIMINAL INSTRUMENTALITY

Artificial intelligence has become a powerful tool that is reshaping how crimes are planned, executed, and concealed. While popular discussions sometimes imagine AI as a self-directing, rogue offender, classical criminal law makes it clear that only a person with intent and legal capacity can be a criminal. AI lacks both of these qualities. It cannot form intent or understand moral responsibility. Yet, this does not render AI irrelevant to the field of criminal law. Instead,

[28]Restatement (Second) of Torts § 283 (American Law Institute 1965).
[29]David Leslie et al., *AI Accountability in Practice*, supra note 3.
[30]UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, supra note 21.

it now acts as an instrument through which human intentions, whether lawful or unlawful, are carried out[31].

AI's significance lies in its ability to amplify human capabilities. Just as an automobile can be used for safe transportation or for an act of violence, AI can either serve legitimate purposes or become a conduit for harm. The difference lies entirely in human decision-making. As AI systems become more advanced and autonomous, they blur traditional notions of responsibility, prompting legal systems to reassess how intent, causation, and control are understood in this new technological context[32].

1. AI-Augmented Traditional Crimes

One of the clearest ways AI affects criminal law is by enhancing traditional crimes such as fraud, identity theft, harassment, and extortion. These offenses are not new, but their execution has undergone significant changes with the aid of intelligent software. Where earlier crimes required time, effort, and personal involvement, AI now allows them to be scaled up, refined, and automated.

For instance, AI-driven fraud has become far more sophisticated. In the past, a scammer might have had to craft deceptive emails or manually impersonate officials. Today, advanced AI models can produce personalized messages that mimic the tone and writing style of trusted individuals or organizations. Deepfake technology can replicate voices or faces, making it nearly impossible for victims to distinguish between genuine and fabricated content. Such scams are more efficient and credible, putting a greater number of people at risk while requiring minimal direct effort from the perpetrator[33].

AI has also transformed online harassment and defamation. Automated bots can flood social media with threatening or defamatory messages directed at specific individuals. The human offenders behind these campaigns may operate only a few accounts, but AI enables them to simulate thousands of distinct voices. This creates the illusion of collective opinion and intensifies psychological harm. The offense remains driven by human intention, but its effect is multiplied through technological means.

[31]Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).
[32]H.L.A. Hart & Tony Honoré, *Causation in the Law* (2nd edn., Oxford University Press 1985) 1–25.
[33]Danielle Keats Citron, "Sexual Privacy" 128 *Yale Law Journal* 1870 (2019).

Another striking example involves extortion and blackmail. Generative AI tools can now create fake but realistic images or recordings that depict victims in compromising or illegal situations. These are then used to coerce, threaten, or silence individuals. The deceptive nature of such material undermines the credibility of victims and complicates their ability to defend themselves publicly. Even though AI performs the act of generation, the underlying wrongdoing originates in the mind of its human user.

In all of these cases, criminal responsibility remains based on human intent and the use of force. The tools have changed, but the moral logic of the crime remains the same. Courts and investigators must, however, navigate new technical complexities, understanding how the AI functioned, how it was deployed, and whether its actions were foreseeable or controllable. As a result, establishing proof now often requires both legal and technological expertise.

2. Automation of Criminal Offenses

Beyond amplifying traditional crimes, AI has introduced a new challenge: the automation of criminal behavior. Some AI systems are designed to operate independently, performing actions continuously without human intervention. This raises difficult questions about liability and causation. When a machine acts on its own, to what extent can its human operator be held responsible?

The answer still lies within the principles of actus reus and mens rea. The act itself may be performed through the AI system. Still, the mental element[34], including the intention or foreseeability of harm, must be found in the human creator, programmer, or user. Criminal law attributes responsibility to people who deploy or fail to control systems that predictably cause harm.

Consider algorithmic trading software. Such programs can make thousands of trades in milliseconds, responding to market changes faster than any human could. In some cases, they unintentionally engage in practices such as "spoofing," which involves creating false signals that manipulate market prices. Even if no human being deliberately clicked a button to perform each unlawful act, the designers or operators may still bear responsibility if the outcome was foreseeable and adequate precautions were not taken.

A similar issue arises in cybercrime. AI-powered botnets can launch continuous, adaptive cyberattacks that adjust in real time to security defenses. Once unleashed, these systems may

[34]Andrew Ashworth, *Principles of Criminal Law* (8th edn., Oxford University Press 2016).

continue attacking targets long after their creators have stopped monitoring them. The legal question is not whether the machine made the decision to attack, but whether the human actor knowingly created or released a system capable of such conduct. Where foreseeability and lack of control can be demonstrated, human liability persists.

Automation also complicates enforcement. In traditional cases, investigators can track a human decision-maker, but when AI operates through self-learning algorithms, reconstructing the line of causation becomes far more complex. Legal authorities must often rely on expert analysis of algorithms, data inputs, and system design. The "black box[35]" kind of nature of machine learning, where even developers struggle to explain why the AI acted a certain way, adds an extra layer of difficulty.

In response, prosecutors may shift toward negligence-based or regulatory offenses. Rather than proving deliberate intent to commit a crime, investigators might show that the operator failed to use reasonable safeguards, allowed uncontrolled systems to function dangerously, or neglected foreseeable risks. Criminal law thus adapts by focusing on duties of care and supervision rather than direct command of the offense[36].

3. Implications for Criminal Doctrine

Understanding AI as a criminal instrumentality rather than a criminal subject helps preserve the integrity of criminal law. It reinforces the principle that only beings with moral and legal awareness can be held accountable for committing crimes. At the same time, it highlights the pressing need to refine existing doctrines so they remain effective in a world of intelligent systems.

First, the idea of causation must evolve. When AI operates within complex, adaptive networks, the link between a person's actions and an unlawful outcome can appear indirect or fragmented. Courts may need to interpret "acting through another" more expansively, treating AI as an intermediary through which a person executes an unlawful act.

Second, foreseeability now plays a central role in determining liability. The more autonomous a system is, the greater the responsibility on developers[37] and deployers to anticipate potential

[36]Rebecca Wexler, “Life, Liberty, and Trade Secrets” 70 *Stanford Law Review* 1343 (2018).
[37]National Institute of Standards and Technology (NIST), *AI Risk Management Framework 1.0*, supra note 32.

misuse. Criminal law may increasingly require evidence of risk assessments, validation testing, and ethical safeguards to determine whether appropriate care was taken before deployment.

Third, many AI-related crimes occur in corporate or institutional settings rather than by individuals acting alone. In such contexts, doctrines of corporate or vicarious liability become crucial. Companies may face prosecution if their decision-making structures, risk management practices, or supervision systems contribute to the occurrence of offenses driven by AI. The emphasis shifts from the behaviour of a single wrongdoer to the moral accountability of organizations that fail to act responsibly.

AI cannot possess consciousness or moral judgment, and therefore cannot, by any definition, commit a crime. Yet it can amplify human wrongdoing in ways never seen before, turning individual acts into potent, large-scale offenses. As a criminal instrumentality, AI brings both opportunity and danger. It challenges legal systems to adapt, to ensure that accountability remains firmly tied to the humans who control, misuse, or fail to control these tools.

The task for modern criminal law is to maintain the moral focus of liability while recognizing the new realities of technology. Law must not treat machines as moral beings, nor allow human responsibility to be obscured by technical complexity. The future of justice in the age of AI depends on striking a balance between innovation and accountability, ensuring that progress never becomes an excuse for impunity.

Case Study: Deepfake Fraud Prosecutions

Deepfake fraud prosecutions reveal a simple yet important truth: the law still looks beyond the technology and asks, "Who is pulling the strings?" Scammers may now use AI-generated voices and videos to impersonate a CEO, a relative in distress, or a government officer, but courts focus on the individuals who choose to utilize these tools to trick others out of money or sensitive information. Even if the model "autonomously" creates the fake media, investigators trace accounts, devices, and communications to show that a human intentionally set the fraud in motion, so automation never really severs the link between human intent and the resulting harm.

Case Study: AI-Enabled Securities Fraud and Algorithmic Market Manipulation

In AI-driven securities and market manipulation cases, the pattern is similar: the trading algorithm might move faster than any person could, placing thousands of spoofing or layering orders in seconds; however, the law still treats it as an extension of its human designers and operators. If a team knowingly programs strategies that will mislead the market, or looks the

other way while safeguards are stripped out, responsibility stays with them, not with the code. These cases drive home a straightforward message: crimes committed at "machine speed" remain human crimes in the courtroom, and clever use of AI will not shield those who profit from manipulation and deceit.

II. How Is Criminal Intent Established in Algorithmic Decision-Making?

The rise of artificial intelligence in decision-making processes poses one of the most profound challenges to classical criminal law. Criminal liability is based on two pillars: the physical act (actus reus) and the mental element (mens rea). While the former can easily be identified through AI's output or effect, the latter is far trickier to locate. Machines, no matter how advanced, have no consciousness, emotions, or intentions. They do not desire outcomes or understand moral consequences. Yet, human actions mediated through those machines can still produce harm.

Therefore, the law must ask a difficult question: if an algorithm acts without a mind of its own, whose mind supplies the intent? In practice, criminal responsibility must still be attributed to identifiable human actors, who are the individuals responsible for designing, controlling, deploying, or benefiting from AI-driven processes. Establishing intent in these contexts does not mean redefining mens rea but adapting its application to environments where decisions are partly automated, distributed, and opaque.

A. The Mens Rea Problem in AI Contexts

Traditionally, mens rea captures a defendant's mental state at the time of the offense: intention, knowledge, recklessness, or negligence. Each level expresses a degree of moral blameworthiness. Intent signifies purposeful wrongdoing; knowledge reflects awareness; recklessness involves conscious risk-taking; and negligence denotes failure to act with due care. These categories assume a conscious actor making deliberate or careless choices[38].

AI systems break that assumption. A machine learning model does not "intend" to harm, nor does it "know" it is causing loss, discrimination, or injury. It operates by processing data and applying programmed logic or probabilistic inference. The absence of subjective awareness or free will leaves no room for mens rea in the machine itself. Hence, when harm occurs, the court cannot treat the AI as a culpable entity; the analysis must return to the humans behind it.

This creates what scholars often describe as the "mens rea gap." It is a gap not in the law's moral framework but in its object of attribution. When an artificial intelligence system makes a

[38]Rebecca Wexler, “Life, Liberty, and Trade Secrets” 70 *Stanford Law Review* 1343 (2018).

decision that leads to unlawful or harmful consequences, such as erroneously blocking financial transactions, misidentifying individuals in crucial situations, recommending discriminatory sentences in the justice system, or unintentionally instigating cyberattacks, questions arise regarding accountability. Whose intent should be considered in such scenarios? Is it the developer, the individual who designed and programmed the system, who bears responsibility? Or is it the executive? This decision-maker endorsed implementing this technology for organizational use. Perhaps it is the operator, the person tasked with monitoring the system's performance and outcomes, who overlooked critical alerts. Alternatively, could it be a combination of all these actors? Such complexities underscore the need for a nuanced understanding of accountability in the face of rapidly advancing technology[39].

The modern approach suggests that mens rea should be reconstructed across the chain of human decisions surrounding the AI's creation and deployment. Responsibility is assessed by examining knowledge, foreseeability, and control at each stage of the process. This approach maintains fidelity to classical principles while recognizing that culpable decision-making can now occur indirectly through machines that act on human instructions and data.

B. Attribution of Criminal Responsibility

To ensure that automation does not weaken accountability, the law attributes criminal responsibility to specific human actors throughout the AI lifecycle: developers, corporate officers, deployers, and operators. Each plays a distinct role in shaping the system'sbehavior and bears different duties of care.

1. Developers

Developers occupy the earliest stage of responsibility. They design, program, and train AI systems. Their choices define the system's capabilities, limitations, and embedded risks. When developers act with reckless disregard for foreseeable harm, they may incur criminal liability. Suppose engineers build a facial recognition algorithm known to produce racially biased results, yet they release it for use in law enforcement without safeguards or disclaimers. If wrongful arrests or discrimination result, the foreseeability of harm could ground a finding of recklessness. Developers may also be liable[40]for deliberately embedding harmful functionalities or ignoring obvious safety flaws. For example, training an autonomous vehicle on insufficient or biased

[39]Danielle Keats Citron, "Sexual Privacy" 128 *Yale Law Journal* 1870 (2019).
[40]National Institute of Standards and Technology (NIST), *AI Risk Management Framework 1.0*, supra note 32.

datasets could lead to foreseeable accidents. Similarly, writing code that enables large-scale data theft or market manipulation may cross the boundary from negligence to intentional facilitation of such activities. The key legal question is whether the harmful outcome was both predictable and preventable through the application of reasonable design precautions: the more obvious the danger, the stronger the case for criminal recklessness or wilful disregard of the law.

Moreover, as AI development becomes increasingly collaborative, dividing responsibility is complex. A software engineer might write safe code, but another team could later integrate that code into a harmful application. This diffusion of responsibility challenges courts to collectively evaluate knowledge and foreseeability. Documentation, risk assessments, and internal warnings become evidential tools in reconstructing mens rea within distributed design processes[41].

2. Executives and Corporate Officers

Liability does not end at the engineering level. Corporate executives and managers often make the ultimate decisions about deploying AI technologies. Under doctrines of corporate criminal liability, organizations can be held accountable for the actions of their leaders, particularly when those leaders authorize dangerous deployments, disregard expert warnings, or engage inwilful blindness toward risk.

Executives bear legal duties to ensure compliance, safety, and ethical oversight. When they knowingly proceed with AI implementation despite evidence of potential harm, for instance, releasing a financial algorithm likely to engage in illegal market behaviours, they may be implicated in offenses ranging from regulatory violations to fraud or negligent endangerment. Courts frequently assess internal communications, risk reports, and governance structures to determine whether leadership acted responsibly.

A growing concern is the phenomenon of "ethical outsourcing," where executives rely on algorithmic opacity to deflect responsibility. Some claim ignorance of specific AI operations, arguing that decisions were made by the system rather than by any person. Courts are increasingly unwilling to accept such claims. The principle of willful blindness, where actors intentionally avoid learning inconvenient truths, prevents leaders from hiding behind complexity. Where risk is foreseeable, ignorance is no defense.

[41]OECD, *OECD Principles on Artificial Intelligence*, supra note 9.

Furthermore, corporate culture plays a crucial role. If executives cultivate an environment that prioritizes profit and speed over safety, they may face liability not for direct actions but for the predictable consequences of systemic neglect. In this way, corporate accountability ensures that oversight responsibilities cannot be delegated away to lines of code.

3. Deployers and Operators

Once AI systems are placed into active use, responsibility also falls on those who deploy and supervise them. Deployers, such as government agencies, private companies, or individual professionals, are often the final human link in decision-making chains. They decide how the AI interacts with real-world data and people[42].

Criminal liability may arise when deployers use systems outside their intended or tested scope, neglect supervision duties, or deliberately misuse them. For example, a bank using an AI credit assessment tool may become liable if it knowingly overrides fairness constraints or allows discriminatory scoring to persist. A police department could face accountability for deploying predictive policing software known to reinforce racial profiling[43].

Operators, the individuals directly managing AI systems, also bear duties of vigilance. Failing to monitor outputs, ignoring warnings, or continuing to rely on clearly defective systems can amount to criminal negligence. The essence of liability here lies not in the complexity of the technology but in the failure of human supervision. As courts have noted in similar contexts, automation does not excuse carelessness.

C. Standards of Culpability

To translate traditional mental states into AI contexts, courts apply existing mens rea categories, recklessness, negligence, and willful blindness, according to the degree of awareness and control each actor possessed. In certain regulated domains, strict liability may also apply.

1. Recklessness and Negligence

Recklessness is established where an actor consciously disregards a substantial and unjustifiable risk. In AI development, this scenario can occur when a developer or executive is aware that an algorithm poses safety hazards but chooses to proceed without implementing mitigation measures. For instance, releasing an untested autonomous weapon system despite known targeting uncertainties reflects a conscious disregard for harm.

[42]Jennifer Arlen, *Corporate Criminal Liability: Theory and Evidence* (Edward Elgar Publishing 201
[43]Virginia Eubanks, *Automating Inequality*, supra note 6.

Negligence, by contrast, applies where the actor ought to have known the risk but failed to take reasonable care. A company that deploys a decision-making system without proper auditing of bias or accuracy, when such evaluation is a standard practice, may be found negligent. The legal test centres on foreseeability regarding whether a reasonable person in the same position could have anticipated the harm and acted differently[44].

These standards encourage proactive risk management. In fast-moving AI industries, where innovation often precedes regulation, courts view the omission of basic safeguards as evidence of indifference rather than ignorance. Documentation of testing, review, and human oversight thus becomes essential for demonstrating due diligence.

2. Willful Blindness

An increasingly important concept in AI criminal law is willful blindness, also known as deliberate ignorance. It addresses situations where actors suspect the existence of wrongdoing but intentionally avoid confirming it. In AI contexts, willful blindness may be inferred when a company or individual deliberately avoids investigating how their system operates or what harm it causes[45].

For example, an executive who chooses not to audit an algorithm because revealing bias could damage reputation or profits may be criminally accountable under this doctrine. Similarly, a developer who ignores red flags in model performance to meet a product launch deadline acts with willful blindness. Courts apply this principle to prevent actors from escaping liability by hiding behind technical complexity or outsourcing to third parties[46].

This approach is vital in machine-learning systems that evolve unpredictably. Because such systems can modify themselves, continuous monitoring is essential. Failing to track outputs or refusing to understand updates that cause harm cannot absolve responsibility. Willful avoidance of knowledge is treated as equivalent to actual knowledge when determining mens rea.

3. Strict Liability

In some regulatory regions, the law imposes liability regardless of intent. Data protection, environmental safety, consumer protection, and financial compliance are prime examples. When

[44]Andrew Ashworth, *Principles of Criminal Law*, supra note 36.
[45]*Global-Tech Appliances, Inc. v. SEB S.A.*, 563 U.S. 754 (2011).
[46]Jennifer Arlen, *Corporate Criminal Liability*, supra note 44.

offenses are classified as strict liability, actors can be convicted simply because a prohibited outcome occurred, even if they lacked the specific intent to commit the offense.

This approach reflects a policy judgment that some domains are so sensitive that responsibility must be automatic. An AI system that breaches personal data, manipulates markets, or causes environmental damage may trigger strict regulatory penalties. Companies in these sectors must therefore implement robust compliance mechanisms, given that "I didn't know" or "the AI decided" will not constitute a defense.

Strict liability also serves a deterrent purpose. It pressures industries to build safety and accountability into the design and deployment of AI from the outset. By equating oversight failure with liability, the law encourages an anticipatory approach rather than a reactive blame-shifting mentality.

D. Judicial Illustration: Algorithmic Risk Assessment

Judicial reliance on algorithmic risk assessment tools in criminal justice systems provides a powerful case study of these challenges. Such tools are often used to predict the likelihood of reoffending, inform sentencing, or guide parole decisions. They operate through proprietary algorithms trained on historical data, which may encode systemic biases.

Courts have increasingly recognized the due process risks associated with such systems. Defendants have argued that reliance on opaque algorithms[47] violates their right to challenge evidence, as neither they nor the court fully understands the tool's decision-making logic. When such tools produce discriminatory or inaccurate outcomes, determining responsibility becomes a complex task. The question that often arises is whether it is the software vendor, the court that adopted the tool, or the official who relied on its outputs.

Judicial decisions across jurisdictions have consistently affirmed one principle: technological opacity does not negate accountability. When human liberty is at stake, decision-makers cannot defer to algorithms without scrutiny. Judges, parole boards, and administrators are expected to understand the limitations of the systems they use and to retain ultimate responsibility for outcomes. In essence, AI may inform judicial decisions, but it cannot replace judicial judgment.

This perspective reinforces a broader legal truth: the use of AI does not alter the chain of accountability; it only changes its form. Even when algorithms act autonomously, the humans who create or rely upon them remain the locus of legal and moral responsibilities.

[47] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016).

Criminal intent in cases involving AI is not a property of machines but of people. The law must therefore reconstruct mens rea by examining human knowledge and control throughout the AI lifecycle. From developers who create algorithms to executives who authorize deployment and operators who oversee use, each stage involves opportunities for choice, awareness, and intervention

III. WHAT ARE THE KEY CONSIDERATIONS FOR PROSECUTORS AND DEFENSE COUNSEL?

As artificial intelligence becomes deeply embedded in decision-making, commerce, finance, and governance, its intersection with criminal law creates new practical and ethical challenges for both prosecutors and defense counsel.

What once existed as a linear human act of crime and a straightforward investigation now unfolds across digital ecosystems involving machine learning, algorithmic predictions, and automated processes. The goal of both prosecution and defense remains the same: to uphold justice while ensuring accountability and fairness. Yet, how each side approaches AI-driven cases will increasingly determine whether the justice system maintains its credibility in the age of intelligent machines.

A. Prosecutorial Considerations

Prosecutors play a crucial role in translating technical misconduct into legally cognizable offenses. In cases involving AI, the challenge is not merely to prove that a crime occurred, but to establish a connection between human responsibility andthe algorithm's actions. This requires navigating technical uncertainty, corporate complexity, and evidentiary opacity while respecting fairness and due process.

1. Charging Decisions

The first and most fundamental task for prosecutors is deciding whom to charge. When a harmful act involves AI, there may be several potential defendants, such as an individual operator, a programmer, a corporation, or all of the above. Determining responsibility depends on three central factors: control over the system, foreseeability of harm, and the legal or regulatory context in which the deployment occurs.

The question arises as to who built the system, who launched it, and who had the authority or obligation to monitor its use. Foreseeability is considered to be equally crucial. If the harmful consequence was an unforeseeable anomaly, criminal sanctions may be inappropriate. But if

risks were apparent, such as using untested AI for medical diagnoses or deploying predictive policing tools with known racial bias, then actors who proceeded regardless may face liability for reckless or negligent conduct.

Context holds an equivalent relevance. In sectors such as finance, data processing, or environmental management, regulations require proactive measures to prevent harmful outcomes. Prosecutors will therefore assess whether compliance obligations were breached and whether any lapses demonstrate intentor gross negligence.

Increasingly, prosecutors are bringing hybrid charges that combine traditional criminal statutes with regulatory offenses. For example, a financial corporation whose AI trading bot engages in spoofing could face both fraud charges and violations under market conduct regulations. In essence, prosecutorial strategy must integrate technical understanding with legal principles to locate the human intention behind automated harm.

2. Evidence Preservation

Evidence in AI-related crimes differs from that in conventional cases. It exists in the form of digital datatraining sets, which are spread across servers and often located in multiple jurisdictions.

Prosecutors must therefore preserve evidence before it is encrypted or altered. AI systems are dynamic, meaning they update themselves with new data; therefore, a single dataset may no longer exist after significant events. The challenge lies in ensuring that what investigators present in court accurately represents the AI's state at the time of the offense.Practical evidence preservation includes securing the training data (which influences bias and output), system logs (which record decisions and commands), model versions (as algorithmic models evolve), and output records. Chain-of-custody protocols must ensure the integrity and reproducibility of evidence. A single omission, such as failure to capture version histories, can render the evidence unreliable[48].

Prosecutors are increasingly collaborating with digital forensic experts who specialize in reconstructing AI behaviour. They must also comply with discovery obligations, ensuring that potentially exculpatory data, including biases or software limitations, are preserved and disclosed

[48]*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).

to defense teams. For both legal fairness and scientific reliability, careful documentation is no longer a procedural formality but a constitutional necessity[49].

3. Admissibility of AI Evidence

Even when evidence is preserved, the next hurdle is admissibility. Courts traditionally require scientific or technical evidence to meet specific reliability standards before it can be presented to a jury. In common-law jurisdictions, this principle is guided by standards such as the Daubert or Frye tests, which examine whether the evidence rests on a valid scientific methodology, has known error rates, and is generally accepted within the relevant expert community.

For AI evidence, these questions are highly complex. Machine learning systems are not static formulas; they are adaptive models that make probabilistic judgments based on data correlations. Prosecutors must demonstrate that the system's processes are reliable, validated, and applicable to the case at hand.

Courts may inquire into the algorithm's error rates, training quality, and validation studies. Moreover, proprietary algorithms present an additional challenge: the vendors often claim trade secret protection, restricting access to key design details. Yet, due process and transparency demand that the accused have sufficient opportunity to challenge the evidence. Consequently, prosecutors must be prepared to describe, in clear and accessible terms, how the AI functions, its limitations, and its relationship to the alleged offense.

As a result, evidentiary admissibility is no longer purely a legal issue but a technical narrative one. Prosecutors, working with experts, must convey the story of the algorithm, its design, intended operation, and observable outputsin language that judges and juries can understand.

4. Expert Testimony

Given the complexity of AI systems, prosecutors increasingly rely on expert testimony. Expert witnesses, often data scientists, forensic analysts, or systems engineers, play a vital role in explaining how the algorithm works and why its outputs should be trusted.

However, prosecutors must carefully select and manage these experts. Credibility and independence matter enormously. A court is less likely to give weight to an expert who is financially tied to the corporation that developed the AI or who fails to explain findings

[49]*Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

impartially. The most successful expert witnesses are those who can translate technical detail into legal relevance, bridging the gap between code and culpability.

Experts must also be transparent about error margins, bias risks, and interpretive limits. Overstating certainty can backfire, especially when defense counsel exposes weaknesses on cross-examination. Prosecutors thus face the delicate task of balancing persuasive advocacy with scientific humility. They must educate without oversimplifying, ensuring that juries comprehend probabilities and uncertainties inherent in algorithmic reasoning.

5. Transnational Challenges

AI-enabled crimes rarely respect national borders. Cyberattacks, money laundering rings, intellectual property theft, and financial manipulations often operate through networks that span multiple jurisdictions. For prosecutors, this creates a logistical and legal labyrinth: evidence may be stored in foreign data centres, perpetrators may reside in different countries, and cooperation may depend on complex extradition treaties or mutual legal assistance frameworks.

International cooperation becomes indispensable. Prosecutors increasingly rely on mechanisms such as INTERPOL exchanges, joint investigation teams, and cross-border warrants. Yet, disparities in privacy law, evidentiary rules, and corporate secrecy can obstruct progress. A country may refuse to share algorithmic data on grounds of national security or trade secrecy, leaving prosecutors without vital evidence.

In this environment, future-focused prosecutors must cultivate both technical and diplomatic capacity. Task forces combining cybercrime experts, legal scholars, and policy negotiators are emerging as crucial tools to navigate the global nature of AI. Effective cross-border prosecution will depend not only on the law but also on diplomacy, trust, and standardized international protocols for sharing digital evidence.

B. Defense Considerations

If prosecutors must prove responsibility, defensecounsel's role is to test that claim rigorously. In AI-driven cases, defense strategies focus on exposing uncertainty, bias, and gaps in causation. Defense lawyers act as the system's moral safeguard, ensuring that technological mystique does not override fundamental rights to a fair trial, confrontation, and due process.

1. Challenging Algorithmic Reliability

Defense counsel often begins by questioning the reliability and accuracy of the AI system at the heart of the prosecution's case. They may argue that the algorithm is biased, that its training data was flawed, or that its analytical methods lack scientific validity.

For example, if predictive models classify certain behaviours as suspicious based on biased datasets, defense lawyers may argue that the AI merely reproduces societal prejudices in digital form. In such cases, the reliability of any conclusion drawn from the algorithmand thus the prosecution's evidenceis undermined.

Defense teams may request full access to the model's architecture, training data, and output history to verify results independently. This process is complex, as companies often invoke intellectual property protections. Yet from a defense perspective, withholding such information prevents meaningful cross-examination. Courts are increasingly sympathetic to these arguments, acknowledging that fairness demands transparency even when trade secrets are involved.

2. Discovery and Transparency

Discovery is a cornerstone of criminal justice: the defendant must have access to the evidence used against them. In AI-related cases, discovery becomes both a technical and a legal matter. Defense counsel may seek disclosure of the algorithm's source code, data inputs, and validation processes.

However, proprietary AI presents new tensions. Companies argue that releasing source code would violate trade secret law, exposing valuable intellectual property. Defense teams counter that without such disclosure, the accused cannot effectively challenge the evidence. Courts are beginning to experiment with compromise orders, allowing independent experts to review sensitive material under confidentiality restrictions.

Effective defense advocacy requires an understanding not only of the law but also of the relevant scientific principles. Attorneys must collaborate with data scientists capable of performing counter-analyses, verifying biases, and identifying algorithmic errors. Increasingly, major defense firms are building interdisciplinary teams that combine legal reasoning with machine learning literacy to challenge algorithmic evidence on its own technical grounds[50].

3. Constitutional Challenges

[50]*Crawford v. Washington*, 541 U.S. 36 (2004).

The use of AI-generated evidence and automated surveillance raises profound constitutional questions. Three issues are particularly prominent: due process, confrontation rights, and privacy.

Due process requires that defendants understand and challenge the evidence against them. When decisions are made by opaque algorithms, known as "black boxes," this right is compromised. A defendant cannot meaningfully challenge a prediction or risk score if neither they nor the court knows why or how it was produced. Consequently, defense lawyers often argue that algorithmic opacity itself violates due process by substituting mechanistic judgment for reasoned legal accountability.

The right to confrontation poses another challenge. Traditionally, defendants have the right to confront and cross-examine their accusers. But an AI system cannot testify, explain, or be questioned. This creates a conceptual dilemma: how can cross-examination occur when "witnesses" are lines of code? Courts have begun resolving this issue by allowing cross-examination of the humans involved, the designers, operators, and interpreters of the AI system. They become, in effect, the voice of the machine.

Finally, algorithmic surveillance raises privacy concerns. AI-driven facial recognition, predictive policing, and data aggregation blur the boundary between lawful investigation and intrusive monitoring. Defense counsel increasingly questions whether AI-based evidence was obtained in violation of constitutional protections against unreasonable searches. As algorithms expand the state's capacity to observe, these questions will shape the new frontier of digital rights jurisprudence.

4. Cross-Examination of AI Evidence

Cross-examination has long been one of the defense's most potent tools, exposing inconsistencies and testing the credibility of witnesses. With AI evidence, confrontation is impossible; instead, defense counsel must redirect scrutiny toward the humans who built, maintained, and interpreted the systems.

Skilled defense lawyers probe not only "what" the AI concluded but "how" it reached that conclusion. They challenge the assumptions underlying the training data, expose previously ignored errors, and question whether the outputs were correctly contextualized. An AI risk score may seem objective, but under scrutiny, it may rely on discriminatory predictors or improperly weighted factors.

Through expert collaboration, defense counsel can reframe seemingly authoritative algorithms as fallible human constructs. This approach emphasizes to juries that AI results are not divine truth but evidence requiring the same scepticism applied to any witness testimony. Cross-examination thus remains a vital safeguard against the danger of "automation bias," the tendency to treat computer outputs as inherently accurate or impartial.

IV. CONCLUSION

Criminal law does not collapse in the face of artificial intelligence, but it is being tested by it. AI systems cannot be moral agents or criminals, yet they profoundly transform how offenses occur and how accountability is determined. Prosecutors must learn to build legally coherent cases out of technical data, establishing human intent and causation behind algorithmic acts. Defense lawyersmust develop strategies to work withthe ever-evolving technology and defend constitutional rights in decision-making.

Ultimately, the central task of criminal law in the age of AI is not to humanize machines, but to preserve human accountability. Automation must not become a shield against responsibility or a tool for coercion. Just as juries once learned to evaluate fingerprints, DNA, and digital forensics, they must now learn to evaluate algorithmsscrutinized through fair, transparent, and adversarial processes.

Justice in the algorithmic age will not depend solely on technical mastery but on moral clarity,the unwavering insistence that every machine's actions are traceable to human judgment, and that no system, however intelligent, stands above the rule of law[51].

[51]John Rawls, *A Theory of Justice*, supra note 26.