VOLUME 6 | ISSUE 2

INTERNATIONAL JOURNAL OF ADVANCED LEGAL RESEARCH

FROM MORAL FAULT TO ALGORITHMIC HARM - THE BREAKDOWN OF MENS REA IN CRIMES SHAPED BY ARTIFICIAL INTELLIGENCE

- Princy Verma[1] & Tarun Sharma[2]

ABSTRACT

The rapid integration of artificial intelligence (AI) into domains involving decision-making with profound legal consequences has exposed foundational fractures within criminal law, particularly in its reliance on *mens rea* as the primary marker of moral and legal culpability. Traditionally conceived as the mental element accompanying a criminal act, *mens rea* presupposes a human subject capable of intention, foresight, and moral judgment. However, contemporary harms increasingly arise from algorithmic systems whose operations are autonomous, opaque, probabilistic, and distributed across multiple human and non-human actors. This article critically interrogates the resulting doctrinal dissonance, conceptualized here as the shift from "moral fault" to "algorithmic harm," and examines how the classical architecture of criminal liability struggles to accommodate AI-shaped wrongdoing.

Adopting a comparative and critical legal methodology, the research analyzes how India, European Union (EU), & United States (US) confront the erosion of *mens rea* in crimes mediated or shaped by AI. It argues that existing criminal law frameworks inadequately address responsibility gaps created by algorithmic decision-making, wherein culpability is diffused among developers, deployers, data curators, corporate entities, and regulatory authorities, while the immediate harm appears to be caused by an ostensibly autonomous system. In the Indian context, the continued reliance on anthropocentric notions of intent under the Bhartiya Nyaya Sanhita, 2023 (which replaced Indian Penal Code, 1860) reveals a normative lag, with limited doctrinal tools to address algorithmic agency. In contrast, the EU's precautionary and regulatory

[1] Assistant Professor, Jaipur National University, Jaipur
[2] Student, B.A.LL.B., Jaipur National University, Jaipur (Rajasthan)

turn, exemplified by the proposed AI Act, reflects an emerging preference for ex ante governance over ex post criminal attribution. The US, meanwhile, demonstrates a fragmented approach characterized by functional liability, prosecutorial discretion, and a growing reliance on civil and corporate accountability mechanisms.

The research critiques the inadequacy of extending personhood or intent to AI systems, warning against both doctrinal fiction and accountability evasion. Instead, it proposes a reconceptualization of *mens rea* grounded in foreseeability, systemic risk creation, and culpable human design or deployment choices. By foregrounding structural power, technological opacity, and institutional responsibility, the research advances a normative framework that shifts criminal law's focus from individualized moral blame to collective and organizational culpability. Hence, this research contends that unless criminal law evolves to meaningfully address algorithmic harm, it risks losing both its moral coherence and its legitimacy in an increasingly automated society.

Keywords: Mens Rea, Artificial Intelligence, Algorithmic Harm, Criminal Liability, Moral Culpability, Responsibility Gap, Autonomous Systems.

BACKGROUND

Artificial intelligence has rapidly migrated from experimental research laboratories into the core infrastructures of governance, commerce, and security. Algorithmic systems now shape decisions in criminal justice (predictive policing and sentencing tools), transportation (autonomous vehicles), finance (high-frequency trading), healthcare (diagnostic algorithms), and national security (surveillance and threat assessment). Unlike earlier automated tools, contemporary AI systems, particularly those based on machine learning, operate through probabilistic inference rather than deterministic logic. They evolve over time, generate outputs not explicitly programmed by humans, and frequently resist transparent explanation.[3]

This transformation has profound implications for legal responsibility. Criminal law has historically functioned on the assumption that harmful conduct can be traced to a discernible

[3] V.A. Tyrranen, Artificial Intelligence Crimes, 3(17) Territory Dev. 10, (2019), https://doi.org/10.32324/2412-8945-2019-3-10-13.

human agent whose mental state justifies punishment. AI disrupts this assumption by inserting a non-human decision-making layer between human action and harmful outcome. As a result, harms occur without a clearly identifiable guilty mind, creating what scholars describe as a "responsibility gap."[4]

Philosophical Importance of Mens Rea in Criminal Law

The doctrine of *mens rea* occupies a foundational position in criminal jurisprudence. Rooted in the maxim *actus non facit reum nisi mens sit rea*, the act does not make a person guilty unless the mind is also guilty, it embodies the moral intuition that punishment is justified only where wrongdoing reflects culpable choice. Philosophers from Aristotle to Hart have emphasized that moral blame presupposes agency, rational deliberation, and the capacity to choose otherwise.

Modern criminal law operationalizes this moral insight through graded mental states: intention, knowledge, recklessness, and negligence. These categories allow courts to distinguish between accidental harm and blameworthy conduct, ensuring proportionality in punishment. Without *mens rea*, criminal law risks collapsing into a purely consequentialist system that punishes harm irrespective of moral fault.[5]

Tensions Between AI Automation and Criminal Law

AI-mediated harm challenges each of these assumptions. Algorithmic systems do not "intend" outcomes in the human sense, nor do they possess awareness or moral understanding. Yet they generate decisions that can cause death, discrimination, or massive economic loss. When an autonomous vehicle kills a pedestrian, or a predictive policing algorithm disproportionately targets marginalized communities, the harm is real, but the mental element is elusive.[6]

Traditional criminal law responds poorly to this scenario. Prosecuting individual programmers often fails due to lack of direct intent or foreseeability. Corporate liability may dilute moral blame, while strict liability risks unjust punishment. This tension exposes a structural mismatch

[4] Ibrahim Suleiman Al Qatawneh et al., Artificial Intelligence Crimes, 12 Acad. J. Interdisc. Stud. 143, (2023), https://doi.org/10.36941/ajis-2023-0012.
[5]Artificial Intelligence Crimes Internationally, 07 RIMAK Int'l J. Humans. & Soc. Scis., (2025), https://doi.org/10.47832/2717-8293.36.28.
[6] Id.

between nineteenth-century doctrines of culpability and twenty-first-century technological realities.[7]

THEORETICAL FRAMEWORK

Criminal liability traditionally rests on two pillars: *actus reus* (the guilty act) and *mens rea* (the guilty mind). The actus reus establishes external harm, while *mens rea* provides the moral justification for punishment.[8] The mental element is typically classified into four categories:

- *Purpose (Intention)* – where the actor consciously aims to bring about a prohibited result.
- *Knowledge* – where the actor is aware that the result is virtually certain.
- *Recklessness* – where the actor consciously disregards a substantial and unjustifiable risk.
- *Negligence* – where the actor fails to perceive a risk that a reasonable person would have foreseen.

These categories presuppose a unitary human decision-maker. They are ill-suited to systems where decisions emerge from complex interactions between data, models, and institutional choices.[9]

Beyond its doctrinal function, *mens rea* embodies the moral architecture of criminal law. Punishment is not merely a response to harm but a communicative act that condemns wrongful choice. This moral dimension distinguishes criminal law from tort or regulatory regimes. However, AI-mediated harm undermines this moral narrative. When outcomes are produced by opaque algorithms trained on vast datasets, identifying a wrongful choice becomes difficult. The moral blame traditionally attached to the actor dissipates, raising concerns about the legitimacy of punishment.[10] AI systems implicated in criminal harm include machine learning models, predictive algorithms, and autonomous agents. These systems operate through pattern recognition rather than rule-following, often generating outputs that even their creators cannot

[7] Dr Mabroka Abdalsalam Mhajer Aqrira, Criminal Liability for Artificial Intelligence Crimes, 2025 ARID Int'l J. Soc. Scis. & Humans. 1, https://doi.org/10.36772/arid.aijssh.2025.6151.
[8] Id.
[9] Yang Zhao &Huiqin Zhu, On the Regulation of Artificial Intelligence Crimes, 27 J. Legal Stud. 47, (2019), https://doi.org/10.35223/gnulaw.27.2.3.
[10] Mingguo Huangfu & Jiahui Gao, On the Regulation of Artificial Intelligence Crimes, 27 J. Legal Stud. 195, (2019), https://doi.org/10.35223/gnulaw.27.2.9.

fully explain. Their adaptive nature means that harmful behavior may emerge over time, without any single human decision directly causing it. The concept of the "moral crumple zone" captures how humans nearest to AI systems, often operators or low-level employees, absorb blame when systems fail, even if they lack meaningful control. This phenomenon illustrates how AI redistributes responsibility without corresponding doctrinal adjustments, leaving criminal law conceptually strained.

AI-MEDIATED HARMS AND THE FRACTURING OF MENS REA

The challenge posed by artificial intelligence to criminal law is not merely theoretical; it emerges most starkly through the concrete harms produced by algorithmic systems operating in real-world contexts. These harms often resemble conventional criminal wrongs in their gravity and social impact, yet they arise through causal mechanisms that fundamentally diverge from the assumptions underpinning classical doctrines of culpability. The difficulty is not that AI creates entirely new forms of harm, but rather that it produces familiar harms through unfamiliar architectures of agency, decision-making, and control.[11]

AI-mediated harm frequently results from the interaction of autonomous or semi-autonomous systems with complex social environments. Autonomous vehicles, for instance, may cause fatal accidents not because of reckless driving in the human sense, but due to probabilistic misclassification, sensor failure, or flawed training data. The resulting harm—loss of life—is indistinguishable in consequence from negligent homicide, yet the absence of a conscious decision-maker complicates attribution of criminal fault. The system does not "choose" to act dangerously; it executes statistical correlations derived from past data. Responsibility, therefore, does not reside in a single moment of culpable choice but is diffused across the lifecycle of design, deployment, maintenance, and regulatory oversight.[12]

Similar complications arise in the financial sector, where algorithmic trading systems can precipitate market crashes or manipulate prices at speeds beyond human intervention. These systems may technically comply with regulatory parameters while producing outcomes that

[11] Makayla Beitler & Eric Talbot Jensen, Battlefield Artificial Intelligence and War Crimes Prosecutions, 2024 SSRN Elec. J., https://doi.org/10.2139/ssrn.4881578.
[12] Id.

destabilize markets and harm millions. Traditional offences such as fraud or market manipulation rely heavily on proof of intent or knowledge. However, when harm emerges from emergent behavior, unanticipated interactions between multiple self-learning systems, the mental element becomes fragmented and elusive. No single human actor may possess sufficient foresight to satisfy classical standards of mens rea, even though the systemic risk was foreseeable at an institutional level.

Surveillance and predictive policing systems further illustrate the structural tension between AI and criminal culpability. Algorithmic risk assessment tools used to predict criminal behavior or allocate policing resources often reproduce historical biases embedded in training data. The resulting discriminatory outcomes, over-policing of marginalized communities, wrongful suspicion, or deprivation of liberty, mirror intentional discrimination in effect, yet they lack a clearly identifiable discriminatory intent. Criminal law, which has historically treated discrimination as an intentional wrong, struggles to address harms that are statistically produced rather than consciously willed. This exposes a doctrinal blind spot where structural injustice escapes criminal accountability because it does not conform to individualized mental states.[13]

What unites these diverse harms is the way in which AI destabilizes the foundational assumption that criminal wrongdoing can be traced to a human subject possessing a guilty mind. The opacity of algorithmic decision-making exacerbates this problem. Many advanced AI systems function as black boxes, generating outputs that cannot be meaningfully explained even by their creators. This epistemic opacity undermines the evidentiary logic of criminal law, which relies on reconstructing mental states through inference from conduct. If the causal pathway from input to output is inscrutable, proving foreseeability, recklessness, or negligence becomes extraordinarily difficult.[14]

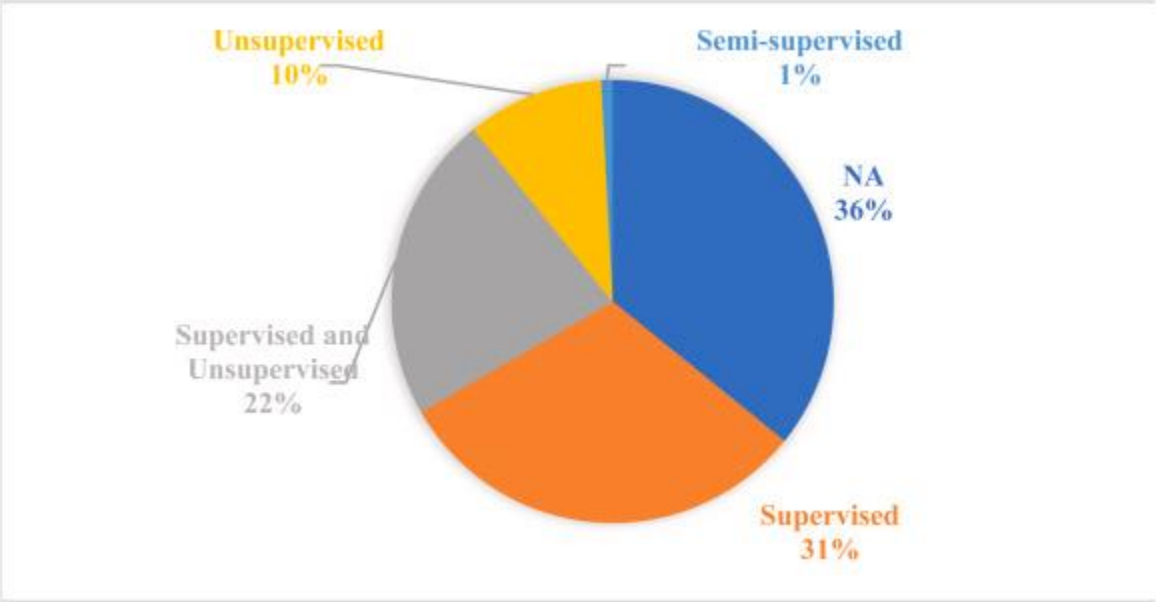
Equally significant is the phenomenon of distributed agency. AI systems are not authored by isolated individuals but emerge from complex organizational ecosystems involving programmers, data scientists, corporate executives, third-party vendors, and regulatory bodies.

[13] Sahar Fouad Majeed Al-Najjar, Criminal responsibilities arising from artificial intelligence crimes, 4 Imam Ja'afar Al-Sadiq U.J. Legal Stud., (2025), https://doi.org/10.64682/3104-9419.1094.

[14] Mohammad Nematei et al., Artificial Intelligence Economic Crimes: Threats and Solutions, 7 Compar. Stud. Juris. L. & Pols. 302, (2025), https://doi.org/10.61838/csjlp.308.

Each actor exercises partial control, yet none exercises total control. Criminal law, however, remains structurally oriented toward identifying singular culpable agents. This mismatch produces what scholars describe as a "responsibility gap," where serious harm occurs without a legally satisfactory locus of blame. In practice, this gap is often filled through the scapegoating of proximate human actors, operators, supervisors, or low-level employees, who lack meaningful authority over system design or deployment. This phenomenon, frequently referred to as the "moral crumple zone," reveals how AI redistributes risk downward while insulating powerful institutional actors from accountability.

The adaptive nature of machine learning systems further complicates temporal aspects of *mens rea*. Unlike static tools, AI systems evolve through continuous learning, meaning that harmful behavior may emerge long after deployment. Criminal law traditionally ties culpability to a specific moment of action or omission. When harm arises months or years later through system adaptation, establishing a temporal nexus between human mental states and harmful outcomes becomes tenuous. This temporal dislocation undermines doctrines of causation and culpability that presuppose relatively linear chains of action and intent.[15]



Source - https://www.sciencedirect.com/science/article/pii/S2590291122000961

[15] Chaima Atailia, Criminal Liability Arising From Artificial Intelligence Crimes, 2025 مجلة المفكر 577, https://doi.org/10.37136/0516-020-001-026.

CRIMINAL LAW RESPONSES ACROSS JURISDICTIONS

The erosion of mens rea in AI-mediated harm is not addressed uniformly across jurisdictions. Comparative analysis reveals divergent strategies shaped by legal culture, institutional capacity, and regulatory philosophy. In India, criminal law remains deeply rooted in classical culpability theory. The replacement of the Indian Penal Code with the Bhartiya Nyaya Sanhita, 2023 reflects continuity rather than rupture in the treatment of mens rea. The new code retains intention, knowledge, recklessness, and negligence as the organizing principles of criminal liability. While this continuity preserves doctrinal coherence, it also reveals a normative lag in addressing algorithmic harm. AI-mediated wrongdoing does not comfortably fit within these categories, as they presuppose human cognition and choice.[16]

Indian statutory responses to technology-related harm are primarily routed through the Information Technology Act, 2000, which criminalizes unauthorized access, data theft, and digital fraud. However, these offences assume direct human misuse of technology. Harm arising from lawful deployment of AI systems that subsequently behave unpredictably often falls outside the scope of these provisions. As a result, prosecutors are forced to rely on expansive interpretations of negligence or conspiracy, which courts have been reluctant to endorse without explicit legislative guidance. Judicial caution in India, while grounded in rule-of-law concerns, exacerbates the accountability gap. Courts have traditionally resisted the creation of new categories of criminal liability through interpretation alone. In the absence of statutory recognition of algorithmic risk or systemic foreseeability, AI-related harms remain under-criminalized. This restraint, though normatively defensible, leaves victims without meaningful recourse and allows powerful actors to externalize algorithmic risk.[17]

By contrast, the European Union has adopted a markedly different strategy. While EU criminal law continues to recognize *mens rea* as a foundational principle, the Union has increasingly embraced regulatory governance as the primary mechanism for managing technological risk. The proposed Artificial Intelligence Act exemplifies this shift. Rather than attempting to retrofit

[16] Rosario Girasa, AI and Crimes, in Artificial Intelligence as a Disruptive Technology 247, (2025), https://doi.org/10.1007/978-3-032-02827-3_7.
[17] Mayada Moustafa El-Mahrouki, Legislative Industry Challenges in Confronting Artificial Intelligence Crimes, 12 J.L. & Sustainable Dev. e3566, (2024), https://doi.org/10.55908/sdgs.v12i4.3566.

criminal doctrines to AI, the Act classifies AI systems based on risk and imposes stringent ex ante obligations on developers and deployers of high-risk systems. This regulatory turn does not eliminate criminal liability but reshapes its foundations. By embedding duties of transparency, risk assessment, and human oversight into law, the EU effectively relocates culpability from subjective intent to objective risk creation. Non-compliance with regulatory obligations may later ground criminal or quasi-criminal sanctions, allowing mens rea to be reconstructed through institutional negligence or reckless disregard of systemic risk. This approach reflects a precautionary philosophy that prioritizes harm prevention over moral blame.[18]

The US occupies an intermediate position characterized by fragmentation and pragmatism. US criminal law lacks a comprehensive framework for AI accountability and relies heavily on prosecutorial discretion and analogical reasoning. Courts have addressed algorithmic issues primarily through constitutional litigation, particularly in challenges to risk assessment tools used in sentencing and bail decisions. However, criminal liability for algorithmic harm remains rare.

Instead, the US has increasingly channeled AI-related harms into civil liability, regulatory enforcement, and corporate criminal responsibility. This approach prioritizes deterrence and compensation while avoiding the doctrinal difficulty of attributing intent in AI-mediated harm. While pragmatic, this strategy risks hollowing out the expressive function of criminal law by treating serious algorithmic harms as regulatory infractions rather than moral wrongs.[19]

COMPARATIVE SYNTHESIS - STRUCTURAL RESPONSES TO THE EROSION OF MENS REA

A comparative examination of India, EU, & US reveals that while all three jurisdictions confront the same underlying phenomenon, algorithmic harm without clear human intent, their legal responses diverge significantly in structure, emphasis, and normative ambition. These divergences are not accidental; they reflect deeper jurisprudential commitments concerning the

[18] Won Sang Lee, A Study on Hacking Crimes Using Artificial Intelligence, 57 Yeungnam U. L.J. 1, (2023), https://doi.org/10.56458/yulj.2023.57.1.

[19] Sergej Zuev, Artificial Intelligence Internet Monitoring To Detect And Solve Crimes, 2020 SSRN Elec. J., https://doi.org/10.2139/ssrn.3719780.

purpose of criminal law, the legitimacy of punishment, and the role of the state in managing technological risk.[20]

At a structural level, all three systems continue to affirm mens rea as a foundational principle of criminal liability. None has formally abandoned the requirement of culpable mental states. However, the ways in which each jurisdiction attempts to preserve, reinterpret, or bypass mens rea in the face of AI-mediated harm reveal contrasting strategies for resolving the responsibility gap. India remains the most doctrinally conservative. Its criminal law framework continues to prioritize individualized culpability grounded in human cognition and choice. This approach reflects a strong commitment to moral blameworthiness as the justification for punishment. However, this commitment also constrains legal adaptability. AI-mediated harm does not easily map onto traditional categories of intention, knowledge, or recklessness, particularly when harmful outcomes arise from lawful deployment of complex systems rather than from misuse or malicious intent. As a result, Indian criminal law exhibits what may be described as normative inertia; an inability to meaningfully engage with algorithmic harm without legislative intervention.[21]

The EU, by contrast, has adopted a structurally preventive orientation. Rather than attempting to stretch criminal doctrines to fit AI, the EU has shifted the locus of accountability upstream. The proposed Artificial Intelligence Act exemplifies a regulatory philosophy that treats AI as a systemic risk requiring governance before harm materializes. This approach implicitly acknowledges the limits of mens rea-based criminal law in addressing algorithmic harm. By imposing mandatory obligations of risk assessment, transparency, and human oversight, the EU constructs a framework in which culpability arises from failure to manage foreseeable risk rather than from subjective intent.

The US occupies a hybrid position. While formally committed to intent-based criminal liability, US law has increasingly relied on civil enforcement, administrative penalties, and corporate criminal liability to address AI-related harms. This functional pragmatism allows for flexibility but risks doctrinal fragmentation. Without a coherent theory of algorithmic culpability,

[20]Yaumi Ramdhani et al., Countering Artificial Intelligence Crimes in a Criminal Law Perspective, 9 RSCH. REV. Int'l J. Multidisciplinary 167, (2024), https://doi.org/10.31305/rrijm.2024.v09.n04.020.
[21]Id.

accountability becomes inconsistent, dependent on prosecutorial discretion rather than principled standards. Moreover, the relegation of serious harms to civil or regulatory domains weakens the expressive function of criminal law as a mechanism for moral condemnation.

Available empirical data confirms that AI-related harm is already widespread. The OECD AI Incidents Monitor recorded over 3,800 documented AI-related incidents globally by 2025, spanning sectors such as law enforcement, transport, healthcare, and finance. In the United States, algorithmic risk assessment tools are used in over 25 states for bail or sentencing decisions, directly influencing millions of defendants each year, while studies on facial recognition systems have reported error rates exceeding 30–40% for certain racial and gender groups. These figures demonstrate that although individual algorithmic decisions may be probabilistic, the systemic risk of harm is empirically established and foreseeable.[22]

During *State v. Loomis*,[23] judges first use the COMPAS risk assessment tool, and though the algorithm is a 'black box' product, the judges do not concern themselves with its innate logic and thus sever the link between sentencing and reason. This case exemplifies the criminal law consequences of systems that do not contain even the most rudimentary forms of mens rea, which effectively dislocates blame from the intentional exercise of judicial discretion.

The incident involving the death of a pedestrian struck by an Uber vehicle in 2018 operating in autonomous mode is an example of diffuse blame across software engineering, corporate governance, and human supervision. The only (safety) driver of the vehicle in which the AI system malfunctioned was charged. This is an example of how AI diffuse mens rea into a lack of willful wrongdoing across multiple entities.[24]

What emerges from this comparison is a shared recognition that classical mens rea doctrine is ill-equipped to address AI-mediated harm, coupled with divergent strategies for managing this inadequacy. India preserves doctrinal purity at the cost of effectiveness; the EU sacrifices

[22] E.G. Ageev et al., Artificial Intelligence as a Subject of Crimes in Medicine, 4 Crim. L. 133, (2025), https://doi.org/10.31085/2949-138x-2025-4-208-133-138.
[23] 881 N.W.2d 749.
[24] National Transportation Safety Board. (2019). Collision between vehicle controlled by developmental automated driving system and pedestrian (Accident Report No. NTSB/HAR-19/03).

individualized moral blame in favor of systemic prevention; and the US prioritizes practical enforcement while tolerating theoretical incoherence.

DATA, EMPIRICAL INDICATORS, & SCALE OF ALGORITHMIC HARM

Although criminal law analysis is often normative, the scale and frequency of AI-related harm underscore the urgency of doctrinal reform. Empirical data increasingly demonstrates that algorithmic systems are not marginal tools but central actors in high-stakes decision-making. Studies conducted by EU institutions indicate that over 60% of large enterprises in the Union deploy some form of AI in decision-making processes, with a significant proportion classified as "high-risk" under proposed regulatory standards. In the criminal justice domain, algorithmic risk assessment tools are now used in sentencing or bail decisions in more than half of US states, affecting millions of defendants annually.[25] In India, while official data remains limited, government initiatives under Digital India and smart policing programs have rapidly expanded the use of predictive analytics and facial recognition technologies.

In *NJCM v. Netherlands*,[26] court ruled that an algorithmic fraud detection system violated privacy law. The system indeed discriminated against the most disadvantaged, though there was no proof of deliberate discrimination. The destructive consequences were a consequence of an algorithm, without an element of intentional bias, which illustrates the challenges of intent liability.

Crucially, empirical research consistently reveals error rates and bias differentials in algorithmic systems that exceed acceptable thresholds in criminal justice contexts. Facial recognition technologies, for example, have been shown to exhibit significantly higher false-positive rates for women and minority groups. Autonomous vehicle incident reports demonstrate that a substantial proportion of fatal accidents involve edge-case scenarios not anticipated during system training. These figures matter not merely as statistics but as indicators of systemic risk

[25] Fatima Dakalbab et al., Artificial intelligence & crime prediction: A systematic literature review, 6 Soc. Scis. & Humans. Open 100342, (2022), https://doi.org/10.1016/j.ssaho.2022.100342.
[26] ECLI: NL: RBDHA:2020.

creation. They demonstrate that algorithmic harm is neither hypothetical nor exceptional; it is structural and foreseeable at an institutional level.[27]

From the perspective of criminal law, this data undermines claims that AI-related harms are unforeseeable accidents. While individual outcomes may be unpredictable, the risk of harm is increasingly well-documented. This distinction between outcome unpredictability and risk foreseeability is central to rethinking *mens rea* in the algorithmic age.

COMPARATIVE TABLE - APPROACHES TO AI AND MENS REA

Dimension	India	European Union	United States
Core criminal law model	Intent-based, anthropocentric	Preventive, risk-based	Fragmented, pragmatic
Treatment of *mens rea*	Central and largely unmodified	Indirectly reconfigured through regulation	Preserved formally, bypassed in practice
Primary accountability mechanism	Individual criminal liability	Ex ante regulatory obligations	Civil, regulatory, and corporate liability
Approach to AI opacity	Largely unaddressed	Transparency and audit mandates	Case-by-case judicial scrutiny

[27] Gabriel Hallevy, External Element Involving Artificial Intelligence Systems, in Liability for Crimes Involving Artificial Intelligence Systems 47, (2014), https://doi.org/10.1007/978-3-319-10124-8_3.

Risk of responsibility gap	High	Moderate (shifted upstream)	High but mitigated through enforcement flexibility
Normative focus	Moral blameworthiness	Harm prevention and systemic safety	Deterrence and compensation

This comparative overview illustrates that no jurisdiction has fully resolved the tension between mens rea and algorithmic harm. Each model involves trade-offs between moral coherence, practical effectiveness, and institutional legitimacy.[28]

EMERGING GLOBAL TRENDS AND LESSONS

Despite jurisdictional differences, several convergent trends are discernible. First, there is a growing recognition that AI-related harm cannot be addressed solely through ex post criminal punishment. Preventive governance, risk assessment, and institutional accountability are increasingly viewed as essential complements to criminal law. Second, responsibility is gradually shifting from isolated individuals to organizations and systems. This reflects an acknowledgment that algorithmic harm is produced by structures of power and decision-making rather than by rogue actors.

Third, there is an emerging consensus that transparency and auditability are prerequisites for any meaningful attribution of responsibility. Without access to algorithmic decision pathways, criminal law cannot even begin to reconstruct culpability. Finally, jurisdictions are grappling with the expressive role of criminal law in the AI context. Treating algorithmic harm as a mere regulatory failure risk normalizing injustice, while insisting on traditional *mens rea* risks impunity. These trends suggest that the future of criminal liability in the age of AI lies neither in

[28] Gabriel Hallevy, Artificial Intelligence Technology and Modern Technological Delinquency, in Liability for Crimes Involving Artificial Intelligence Systems 1, (2014), https://doi.org/10.1007/978-3-319-10124-8_1.

abandoning mens rea nor in preserving it unchanged, but in reconstructing it around concepts of foreseeability, systemic risk creation, and institutional control.[29]

RECONSTRUCTING MENS REA FOR THE AGE OF ALGORITHMIC HARM

The preceding analysis demonstrates that the crisis of *mens rea* in AI-mediated harm is not a temporary anomaly but a structural challenge to the moral and doctrinal foundations of criminal law. Attempts to resolve this crisis by either extending legal personhood to artificial intelligence or abandoning mental fault in favor of strict liability are both normatively unsatisfactory. The former relies on doctrinal fiction that severs culpability from moral agency, while the latter risks collapsing criminal law into a purely regulatory instrument devoid of ethical meaning. What is required instead is a principled reconstruction of the *mens rea* that remains faithful to criminal law's moral commitments while acknowledging the realities of algorithmic governance.[30]

A reconfigured conception of *mens rea* must begin by decentering the search for intention at the moment of harm and instead focus on culpable human choices embedded in the lifecycle of AI systems. This approach recognizes that while AI systems lack consciousness, they are nonetheless shaped by human decisions concerning design architecture, training data selection, deployment contexts, and oversight mechanisms. Criminal culpability should therefore attach not to the algorithmic output itself but to the human and institutional decisions that created and sustained unreasonable risks of harm.

The Flash Crash, 2010 was a phenomenon triggered by a complex interplay of autonomous high-frequency trading (HFT) algorithms. Many of such HFT algorithms were not designed or overseen by any human traders. The consequences were devastating, however, and a deliberate attempt at market manipulation was never established. Such and similar events draw attention to the phenomenon of 'unexpected damages' caused by autonomous AI systems, where systems

[29] Kong Yuchen, Safeguarding the Future: Legal Frontiers in Preventing Artificial Intelligence Crimes, 38 Lecture Notes Educ. Psych. & Pub. Media 192, (2024), https://doi.org/10.54254/2753-7048/38/20240646.
[30] Asri Gresmelian Eurike Hailtik&Wiwik Afifah, Criminal Responsibility of Artificial Intelligence Committing Deepfake Crimes in Indonesia, 2 Asian J. Soc. & Humans. 776, (2024), https://doi.org/10.59888/ajosh.v2i4.222.

satisfy the requirements of advanced AI, while the human operators (if any) do not exhibit any intention or concept of damages, or the classical criminal 'mens rea' requirements.[31]

Foreseeability emerges as a central organizing principle in this reconstructed framework. Traditional criminal law already recognizes that moral blame may arise from conscious risk-taking or negligent failure to anticipate harm. In the context of AI, the relevant inquiry is not whether a specific harmful outcome was intended, but whether the risk of harm was foreseeable given the known limitations, biases, and failure modes of the system. Where developers or deployers proceed despite documented risks, liability may be grounded in a form of systemic recklessness. Where they fail to conduct reasonable risk assessments or implement safeguards, liability may arise from institutional negligence.[32] This reconceptualization also requires rethinking causation. In AI-mediated harm, causation is rarely linear. Instead, it is distributed across time and actors. Criminal law must therefore adopt a more nuanced understanding of causal contribution, recognizing that responsibility may be shared among multiple actors who collectively create conditions for harm. This does not necessitate collective punishment, but rather calibrated attribution of liability based on degrees of control, knowledge, and advantage.

OPERATIONALIZING ACCOUNTABILITY - LEGAL AND INSTITUTIONAL MECHANISMS

Reconstructing mens rea in theory must be accompanied by practical mechanisms capable of operationalizing accountability. One such mechanism is the mandatory adoption of AI impact assessments prior to deployment in high-risk domains such as criminal justice, transportation, and healthcare. These assessments would document foreseeable risks, mitigation strategies, and residual uncertainties. Failure to conduct or heed such assessments should constitute evidence of culpable disregard for risk.[33] Another critical mechanism is the creation of algorithmic audit trails. Criminal investigation relies heavily on reconstructing past events and mental states. Without access to decision logs, training data histories, and model updates, meaningful

[31] U.S. Securities and Exchange Commission. (2010). Findings regarding the market events of May 6, 2010.https://www.sec.gov/news/studies/2010/marketevents-report.pdf.

[32] S. Maltseva & V. Geranin, The Use of Artificial Intelligence in the Investigation of Crimes, 11 Bull. Sci. &Prac. 356, (2025), https://doi.org/10.33619/2414-2948/113/47.

[33]Criminal Liability for Crimes Committed by Artificial Intelligence Devices (Robots), 8 Twejer 228, (2025), https://doi.org/10.31918/twejer.2584.eli.11.

accountability is impossible. Legal mandates requiring auditability and explainability are therefore not merely regulatory preferences but prerequisites for criminal responsibility.[34]

Standards of liability must also differentiate between actors based on their functional roles. Developers who design core architectures, corporations that profit from deployment, and public authorities that mandate or authorize AI use qualitatively different forms of control. A reconstructed *mens rea* framework should reflect these distinctions, avoiding both blanket immunity and indiscriminate punishment. Corporate criminal liability, in particular, offers a promising avenue for aligning responsibility with organizational power, provided it is accompanied by meaningful sanctions and compliance obligations.

ETHICAL LEGITIMACY AND THE EXPRESSIVE FUNCTION OF CRIMINAL LAW

Beyond doctrinal coherence, any reform of the mensrea must preserve the ethical legitimacy of criminal law. Punishment is not merely instrumental; it is expressive. It communicates societal condemnation of wrongful conduct and reaffirms shared moral values. If criminal law fails to meaningfully address algorithmic harm, it risks appearing complicit in technological injustice, undermining public trust.[35] At the same time, ethical legitimacy requires restraint. Punishing individuals who lack meaningful control over AI systems would be unjust and counterproductive. This underscores the importance of aligning liability with power and knowledge. Those who shape technological systems, set incentives, and reap benefits must also bear responsibility for the harms those systems produce. Human oversight remains an ethical cornerstone. Meaningful human control over AI systems is not only a technical safeguard but a moral imperative. Where humans retain the ability to intervene, override, or halt harmful processes, failure to exercise such control should attract heightened scrutiny. Conversely, deploying systems that preclude meaningful human intervention may itself constitute culpable risk creation.[36]

CONCLUSION

[34] Dr Mabroka Abdalsalam Mhajer Aqrira, Criminal Liability for Artificial Intelligence Crimes, 2025 ARID Int'l J. Soc. Scis. & Humans. 1, https://doi.org/10.36772/arid.aijssh.2025.6151.
[35] Alexander Babuta & Marion Oswald, Machine learning predictive algorithms and the policing of future crimes, in Predictive Policing and Artificial Intelligence 214, (2021), https://doi.org/10.4324/9780429265365-11.
[36] Id.

The rise of artificial intelligence exposes a profound tension at the heart of criminal law. The doctrine of mens rea, long regarded as the moral anchor of punishment, presupposes a human subject capable of intention and moral judgment. AI-mediated harm disrupts this presupposition, producing serious wrongdoing without a clearly identifiable guilty mind. This research has argued that the appropriate response is neither to anthropomorphize machines nor to abandon culpability altogether, but to reconstruct *mens rea* around concepts of foreseeability, systemic risk, and institutional responsibility.

Comparative analysis reveals that jurisdictions are already moving, albeit unevenly, toward this reconstruction. India's adherence to classical culpability underscores the need for legislative innovation. The European Union's regulatory turn demonstrates the value of preventive governance but risks sidelining criminal law's moral voice. The United States' pragmatic reliance on civil and corporate liability offers flexibility but lacks normative coherence. Together, these approaches illustrate both the necessity and the difficulty of reform.

Hence, the legitimacy of criminal law in an automated society depends on its capacity to evolve without abandoning its moral foundations. By shifting the focus from individualized moral fault to culpable human choices embedded in technological systems, criminal law can continue to serve as a meaningful instrument of justice. Failure to do so would render it increasingly irrelevant in a world where harm is no longer the product of human intention alone, but of algorithms acting within structures of human power.