CRIMINAL INTENTION IN THE AGE OF ARTIFICIAL INTELLIGENCE: CAN A MACHINE HAVE MENS REA

Yashwardhan Rai*

ABSTRACT

This paper explores the complex question of whether artificial intelligence (AI) can possess 'mens rea'—the "guilty mind" central to criminal liability. As AI systems increasingly perform autonomous actions, incidents of harm or crime prompted by AI advice or malfunction raise pressing legal dilemmas. The traditional foundation of criminal law, which pairs a wrongful act (Actus reus) with a culpable mental state (mens rea), is strained when the actor is a non-human entity. The paper argues that AI lacks consciousness, intention, and moral understanding, making it conceptually impossible to attribute mens rea in its human sense. It examines doctrinal challenges of actus reus, causation, and mental state attribution in AI-related crimes and evaluates three main legal responses: human-focused liability, strict or regulatory liability, and the idea of limited legal personhood for AI. Comparative perspectives from the EU, the United States, and India are analysed to highlight emerging frameworks. The study concludes that while machines cannot bear criminal intent, accountability must rest with humans and organisations responsible for their creation, deployment, or misuse. It advocates a hybrid approach integrating human culpability, regulatory oversight, and governance standards to balance justice, deterrence, and innovation.

Keywords: Artificial Intelligence, Mens Rea, Justice, Actus Reus.

INTRODUCTION

In the current scenario, AI has become an essential part of human life. From taking help for small projects, assignments, research, and to find answers easily for questions, AI becomes an important part. The use of AI is not limited to work and projects, it is also used for opinion, talking, and to find answer for personal questions. It will be no wrong to say that the whole

*BA LLB, SECOND YEAR, PRESTIGE INSTITUTE OF MANAGEMENT AND RESEARCH, GWALIOR.

world nowadays revolves around AI. But sometimes this became a big problem. Recently, many cases have been reported where a crime has been committed on the advice of an AI system. In the year 2021, a generative AI chatbot, Replika App advice a guy named Jaswant Singh to plan an assassination of Queen Elizabeth II by the use of a crossbow. Later, he got arrested and pleaded guilty to treason. The use of AI is not limited to physical crime. It is also used for online fraud and scamming. The AI tool, such as deep fake, is used in AI generated photo and videos to impersonate family members or individuals to take a large sum of money. A Hong Kong-based employee of a multinational firm was tricked into transferring HK$200 million after attending a video call with an AI deep fake of his company's CFO and other employees.

Criminal law rests on a moral arithmetic: we punish not just harmful results, but culpable perpetrators. A foundation principle of criminal law is that a person is not guilty of a crime unless both a wrongful act (actus Reus) and a guilty state of mind (mens rea) are present. The wrongful act (actus reus) is paired with the mental state (mens rea) to establish criminal culpability. When a self-sustaining system, not a human, determines to bring about the harm, the doctrinal framework connecting wrongdoing and culpability is stretched taut. The issue is not abstract. Self-driving cars, recommendations of algorithms that provoke violence, algorithmic trading robots that corrupt markets, and generative AI that creates libellous deep fakes all cause concrete harms. Criminal law must react when the "actor" is software, hardware, or a networked machine.

This article does three things:

- Explains what mens rea is in criminal doctrine and why it is important;
- Explores whether and how classic mental-state notions can be mapped onto artificial agents; and
- Considers practical legal solutions and policy choices, providing a framework for legislatures, judges, and regulators.

Throughout the article contend that mens rea in the traditional, subjective sense is not meaningfully ascribable to present AI systems; thus, a hybrid framework, integrating human responsibility, graduated strict liability, compulsory design and governance standards on high-risk AI, and restricted regulatory personhood where fitting ,more supports justice, deterrence, and social wellbeing.

WHAT IS MENS REA AND WHY DOES IT MATTER?

Mens rea means "guilty mind" in Latin it defines the state of mind necessary for most criminal acts. Based on the crime, the intent, knowledge, recklessness, or negligence may be required by the law. Whether present or absent, the defendant's state of mind usually decides whether culpable wrongdoing or excusable accident is at hand. To cite esteemed classic American criminal law jurisprudence, mens rea differentiates blameworthy conduct from harmful conduct.

Why is mens rea important? First, punishment can be morally justified only if agents had control and knowledge adequate to render them blameworthy. Second, mens rea has utilitarian purposes — it assists deterrence, incapacitation, and expressive denunciation by directing punishment at those who could have avoided and foreseen harm. Third, the mental-state requirement safeguards individuals from punishment for accidents or situations outside their control of their mind. Therefore, any suggestion to make machines criminally responsible will have to take account of these moral and policy roles.

In criminal law, R v Prince (1875),[1] it is believed that this is a landmark case where the mens rea principle is clarified and established in law.

WHY ASCRIBING SUBJECTIVE MENTAL STATES TO MACHINES IS CONCEPTUALLY PROBLEMATIC?

At first blush, one might say: if a system acts in ways that satisfy actus Reus (the physical element of a crime), why not hold it liable under the same rules as a person?

The problem is that mens rea presupposes capacities that current AI lacks: belief, intention, understanding of moral reasons, and normative awareness.

Intentionality and Comprehension: Normal human intention involves an agent creating a goal and acting to ensure the realisation of that goal. Modern AI systems — even sophisticated machine-learning systems — work through optimisation processes, pattern matching, and probability estimates; they do not create intentions in the phenomenological or normative

[1] https://www.casebriefs.com/blog/law/criminal-law/criminal-law-keyed-to-kadish/defining-criminal-conduct-the-elements-of-just-punishment/regina-v-prince/

sense. They have no consciousness, subjective reasons, or sense of moral or legal norms. Charging an algorithm with "intending" threatens anthropomorphism[2] and category error.

Epistemic States and Knowledge:[3] Knowledge-based legal mens rea (e.g., "knowingly") assumes an epistemic state; the agent believes facts rendering the act wrongful. AI systems have internal representations and confidence measures, but these are computational states, not beliefs in the normative sense. The mapping from system internals to the legal notion of "knowing" is contested and underdetermined.

Negligence and Recklessness: Recklessness demands conscious "disregard" of a significant risk; negligence entails inability to act with reasonable care. Although AI conduct can be adjusted to risk levels, imputing a conscious "disregard" to a machine is metaphysically questionable. Still, the predictable conduct of an AI based on how it was designed and trained can generate risk, and human beings who design, send forth, or neglect to monitor that AI could be criminally negligent or reckless.

Scholars thus mostly conclude that attributing subjective mens rea mirroring human psychological states to AI is unrealistic for current and near-future systems. Hallev,[4] in his paper “The criminal liability of Artificial Intelligence entities,[5] describes three models of the criminal liability of artificial intelligence. Similarly, Benoit Dupont, Yuan Y. Stevens and Hannes Westermann in their work “Artificial Intelligence in the context of Crime and Criminal Justice”[6] wrote about Artificial Intelligence and its work in the context of crime and criminal law.

DOCTRINAL ISSUES: ACTUS REUS, CAUSATION, AND MENS REA IN AI SITUATIONS

Three interlocking doctrinal problems arise when AI causes harm: establishing the physical act (actus Reus), proving causation, and identifying the relevant mental state.

Actus Reus: Actus Reus typically requires a voluntary act or omission. If an AI system’s sensor triggers and it moves a robotic arm that injures someone, a human-centred reading can treat the

[2] The attribution of human characteristics or behaviour to a god, an animal or an object
[3] Reality to knowledge or to the degree of its validation
[4] An Israeli professor of criminal law
[5] https://ideaexchange.uakron.edu/cgi/viewcontent.cgi?article=1037&context=akronintellectualproperty
[6] https://www.cicc-iccc.org/public/media/files/pro

machine's movement as a physical act. But most statutory schemes take it for granted that the "actor" is a natural person. Some statutes actually criminalise conduct by "persons" or "individuals" and possibly do not envision machines as immediate perpetrators. Where statutes are indifferent, courts must then decide whether to broaden the category of "actor" or to consider the machine's motion as the proximate physical act executed by its owner or controller, a human being.

Causation: Causation raises thorny issues when behaviour results from complex interactions: training data, emergent model behaviour, third-party model fine-tuning, and online updates. Determining the chain from human inputs/design decisions to harmful output is frequently technically challenging and may confound traditional proximate cause analysis. This practical difficulty frequently redirects legal attention toward the regulation of the human co-developers, deployers, or operators instead of the machine itself.

Mens Rea: Even if actus Reus and causation are demonstrated, mens rea is generally absent. For most crimes fraud, murder, theft subjective purpose is key. If the autonomous agent has no beliefs and intentions, criminal law is faced with the challenge of either (a) analogising machine states to mens rea; (b) assigning the necessary mental state to some human participant (designer, operator, owner); or (c) reconfiguring the law to allow for liability without classical mens rea for some harms. Each path has trade-offs.

THREE DOCTRINAL RESPONSES

Legal frameworks are converging on three general responses: (1) Assign Culpability to Humans (direct or accessorial liability); (2) Place strict or regulatory liability for harm caused by AI; and (3) Explore limited legal personhood for specific AI entities.

Human-Focused Attribution (Direct and Accessory Liability): This maintains typical mens rea by allocating responsibility to humans, such as programmers, developers, operators, and corporate actors.

Human Direct Liability: Where a human utilised AI as an instrument in committing a crime (e.g., programming an AI to send phoney emails for phishing purposes), the human may be held liable for intent and Actus Reus. The fact that there is AI as the "instrument" does not supersede human mens rea.

Accessorial and Omission Liability: Humans who facilitate, neglect to monitor, or are careless/careless in deployment can be held liable under doctrines such as negligence, recklessness, or accomplice liability. This applies to most cases of harm that are foreseeable from bad design, insufficient testing, or known model flaws.

This human-focused path supports orthodox criminal law and is commonly preferred as it maintains moral blame-worthiness: humans are the creators of the risks and can be held to take measures to minimise them. Most commentators prefer enlarging current doctrines — such as corporate liability — instead of inventing machine liability.

Regulatory Offences and Strict Liability: Where subjective intent is impractical to prove or excessively burdensome, legislatures may enact strict or regulatory offences for some high-risk activities with AI. Some examples include criminalising the use of AI in safety-critical applications without required protections, or for data-protection breaches resulting in harm.

Strict liability lightens the cognitive load of establishing intent and enhances deterrence through less complex enforcement. But sparingly applied, it must be: criminal sanction without mens rea is morally suspect and can be unjust where participants had no realistic capacity to avoid harm. A reasonable response might be to employ administrative penalties and regulatory fines for most offences and reserve criminal sanctions for gross negligence or systematically reckless, blind behaviour.

Scholars such as Abbott,[7] in his book "The Reasonable Robot: Artificial Intelligence and the Law,[8] have made the argument that civil/regulatory regimes ought to bear the primary burden of responsibility for AI, since these regimes can more easily be made adaptable and technologically tailored than criminal measures. This is a consideration for policy design: criminal law need not be the sole tool.

Limited Juristic Personhood for AI: A still more extreme idea is to give AI systems a kind of juristic personhood so that they can be held (and penalised) as entities. That could facilitate direct imposition of liability and make asset forfeiture or mandatory shutdown possible. But Juristic personhood to AI is in serious theoretical and normative trouble.

[7] Author of "The Reasonable Robot: Artificial Intelligence and the Law"
[8] https://share.google/JevTEQx5HyS6NXKeM

Moral Desert and Retribution: Punishment requires the capacity to understand and respond to sanctions. Punishing a machine (e.g. dismantle the hardware of the machine) does not express moral blame in the criminal-law sense.

Practical Enforcement: AI systems neither have assets nor moral psychology, and enforcement can simply pass on the externalities back to individuals. Personhood might give rise to perverse incentives for authors to split duties.

Regulatory Alternatives: Many scholars are convinced that existing corporate or organisational liability can play the exact safeguard function without granting personhood to software. Because of this second thought, juridical personhood remains controversial and is not yet a generally accepted legal remedy. Nevertheless, hybrid proposals — wherein some legally defined "autonomous systems"[9] assume limited responsibilities and are required to carry mandatory insurance and be monitored — are under debate in academic and policymaking communities.

COMPARATIVE AND INNOVATIVE LEGAL MODELS

Various jurisdictions are testing models that combine the above categories.

European Union and Regulatory Focus: The EU's AI Act (planned and discussed over 2023–2025) prioritises a risk-based regulatory framework: high-risk AI systems are subject to mandatory design, transparency, and conformity obligations, enforced through administrative sanctions. The EU model aims to avoid harm by imposing pre-deployment controls and obligations rather than solely by way of conventional criminal prosecutions.

United States: Mix of extant doctrines and sectional regulation. In the United States, courts and prosecutors had mainly applied existing criminal laws against humans who exploit AI. Where the law is silent on machines, the law is enforced on fraud, negligent use and computer-crime statutes. The U.S. approach is civil-sanctions oriented and human-prosecution oriented for technologically based harms, and agencies operated by the government turn to lawmaking in specific sectors (financial markets, transport).

India: In India, there is doctrinal uncertainty and a need for reform. Indian scholarship and recent policy debates have highlighted gaps in the criminal law of India with respect to AI

[9] Self-governing or Decision-making entities, such as robots or software programs

harms. Suggestions include updating statutes to cover automated decision-making, enhancing corporate liability, and imposing regulatory compliance duties for high-risk AI applications. New scholarly articles indicate India might embrace a hybrid approach: regulatory measures complemented by criminal enforcement of severe forms of negligence or malevolent exploitation.

Throughout the jurisdictions, there is a common theme: scholars and legislators alike opt for preventive and regulatory solutions over comprehensive doctrinal reengineering of mens rea.

PRACTICAL EXAMPLES: AUTONOMOUS VEHICLES, DEEPFAKES, AND MARKET-MANIPULATING BOTS

Three practical examples demonstrate the tensions in doctrine and the policy responses they elicit.

Autonomous Vehicles: Autonomous cars can be used in accidents that would classically involve vehicular homicide or reckless driving laws. But who is the "driver"? Most models hold the manufacturer, computer programmer, fleet owner, or the human operator responsible based on the control and design of operations. In situations where vehicle action stems from unpredictable emergent characteristics of machine learning, it is hard for prosecutors to establish human mens rea unless there is proof of negligent testing, disregarded warnings, or willful disregard of safety procedures. For run-of-the-mill regulatory offences, administrative sanctions and product liability tend to take centre stage.

Deepfakes and Harms of Expression: Generative models make it possible to produce realistic deep fakes employed to extort, defame, or urge violence. If AI independently creates a fake video leading to public harm, prosecutors will usually target the human(s) who commissioned the dissemination, requested the material, or benefited from it.

MARKET-MANIPULATING ROBOTS

Pricing robots that collude on prices or implement manipulative tactics give rise to concerns such as fraud and distrust. Courts argue that whether the program can intrinsically conspire or intend market manipulation. Practically, enforcement usually deals with human traders, compliance specialists, and entities that direct the bots and regulate the use of market

monitoring, the obligation of record keeping and possible sanctions for criminal knowledge of manipulation.

POLICY RECOMMENDATIONS

Specific suggestions for consistency, social protection and innovation, which are government-friendly, are:

Retain Mens Rea Centred to Human as the Norm: Criminal liability still demands human mens rea (intent, knowledge, recklessness) for central offences unless the legislature provides specifically. It helps in retaining moral legitimacy and does not punish representatives who are incapable of moral agency.

Establish Targeted Strict-Liability/Regulator for High-Risk AI Deployment: Categories such as safety or critical transportation, healthcare, critical infrastructure; the legislature ought to suggest administrative and criminal sanctions graded by culpability. For example; Fines and orders by administration to correct failures, criminal penalties for gross negligence, reckless deployment or intentional disregard of safety rules.

Regulatory regimes (model-based conformity, incident reporting, compulsory audits) should go alongside such liability provisions. The regulatory strategy within the EU's AI Act offers a crucial model for risk-based measures and pre-market requirements.

Enforce Corporate and Organisational Liability: Organisational liability principles (vicarious, corporate criminal liability) need to be brought up to date to accommodate harms from AI use. Companies designing, marketing, or profiting from high-risk AI should be held accountable when organisational culture, lack of sufficient compliance, or incentives yielded foreseeable harm.

Enforce Design, Test, and Transparency Requirements: Legislators ought to require minimum standards for high-risk AI:

- Thorough pre-deployment testing and validation.
- Explain ability standards and availability of logs for incident analysis.
- Insurance or financial responsibility schemes to ensure victim compensation.

These requirements make it more practicable to assign responsibility to human agents and minimise the necessity to attribute mens rea to machines.

Establish an Incident Reporting and Monitoring Regime: A public repository of severe AI malfunctions, as with aviation or medical device reporting systems, would enhance system learning, enable the regulators to identify perilous patterns, and facilitate civil or criminal investigations when human fault seems evident.

Apply criminal law judiciously and openly spare criminal punishment for performers who demonstrate states of moral fault intent, knowledge, recklessness, or gross negligence. Refrain from imposing criminal penalties on mere negligence or technical mishaps without evidence of human fault.

Promote Standards and Industry Regulation: Harmonised technical standards (testing thresholds, data provenance, safety guarantees) minimise uncertainty. Public–private collaboration may facilitate the deployment and revision of standards quickly as AI continues to develop.

REFUTING COUNTERARGUMENTS

Counterargument 1: Since harm is done by AI systems and humans abdicate responsibility by resorting to concealment behind complexity, criminal law doesn't deter; machines must then be punished criminally.

The threat of human abdication is genuine, but the answer lies not in penalising machines but in strengthening human and institutional responsibilities (compulsory logging, rigid product liability, regulatory penalties). Penalising machines is symbolically inadequate and effectively useless; it could also allow human perpetrators to go scot-free by taking cover behind an artificial "entity" that cannot be ethically corrected.

Counterargument 2: Certain AI systems have emergent behaviour unexpectedly no human reasonably anticipated the harm. Making anyone criminally responsible is unfair.

Answer: Where no human anticipated the harm, and all that could have been reasonably anticipated was done, criminal liability is unsuitable. Insurance, civil remedies and regulatory accident responses should cover the costs. Criminal law should be aimed at cases where human actors defaulted on their responsibilities, were negligent, or intentionally abused AI.

SPECULATIVE POTENTIAL FOR FUTURE AI

The above analysis is based on the capabilities of current and near-future AI. Assuming that hypothetical future systems might have consciousness, self-awareness, and normative comprehension like human mental states, the moral and legal environment would shift. Questions of literal machine culpability, rehabilitation, and moral desert would then need new analysis. For the time being, these hypotheses remain philosophical: legal and policy efforts today should concentrate on possible technological paths and on accountability models assigning responsibility to human beings and organisations.

CONCLUSION

Can a machine have mens rea? On current legal, philosophical, and technological understanding, the answer is: not in any sense that justifies direct criminal punishment. Mens rea assumes capacities intention, understanding, and moral agency that current AI systems lack. Therefore, effective and justifiable legal measures focus on human responsibility, targeted regulation, strict liability or administrative punishment in well-defined high-risk situations, and governance structures that encourage safe design and deployment. Criminal law should not be discarded; it should instead be honed to aim at human culpability where it is present. Regulatory and civil measures should assume the majority of AI-specific regulation due to their higher degree of flexibility, technical remediation suitability, and resistance to crossing over moral desert principles. Through doctrinal commitment to mens rea coupled with pragmatic regulatory creativity, the law may handle AI hazards without compromising fundamental principles of justice.