THE CONUNDRUM OF ATTRIBUTION: ACTUS REUS AND MENS REA IN AI

Vaidehi Pandey & Dr. Ekta Gupta(supervisor)

B.A. LL.B. (Hons.), 5th Year Amity University, Noida, AUUP

Assistant Professor Amity University, Noida, AUUP

ABSTRACT

The rapid evolution of artificial intelligence challenges foundational principles of criminal law, particularly the doctrines of actus reus and mens rea. Traditionally, criminal liability has been premised on human conduct involving conscious action and intentional wrongdoing. However, autonomous AI systems operate without awareness, intent, or moral agency, thereby creating significant gaps in legal accountability. This paper examines the difficulties in attributing criminal responsibility to AI systems by critically analyzing the applicability of actus reus, mens rea, and causation doctrines. It further evaluates Gabriel Hallevy's three models of liability—Perpetration-by-Another, NaturalProbable-Consequence, and Direct Liability—highlighting their limitations in addressing autonomous decision-making systems. The study argues that existing legal frameworks, rooted in human-centric assumptions, are insufficient to regulate AI-driven harm. It concludes by emphasizing the urgent need for a paradigm shift in criminal law, moving beyond traditional constructs toward a system capable of addressing accountability in technologically advanced environments.

Keywords: Artificial Intelligence (AI), Criminal Liability, *Actus Reus*, *Mens Rea*, Autonomous Systems, Legal Accountability, Causation, AI Ethics, Gabriel Hallevy Models, Perpetration-by-Another, Natural-Probable-Consequence, Direct Liability, Human-Centric Legal Frameworks, Algorithmic Decision-Making, Emerging Technologies Law

Introduction

At its core, criminal justice centres unshakably on humans. Over long stretches of time, key legal ideas took shape around the belief that people act with awareness, intent, and responsibility.[1] A working mind - biological, able to weigh choices - is presumed when someone breaks the law. Punishment aims to discourage future acts, answer wrongdoing, or foster change in conduct because it targets those seen as choosing their path. This framework fits only beings thought capable of understanding rules and deciding whether to follow them. Still, widespread emergence of completely self-governing artificial systems shakes the foundation of human-centred frameworks. As machines shift from tools that respond to ones that initiate,

legal structures face a novel presence - one equipped with complex reasoning and autonomy but devoid of subjective experience or ethical understanding. Only now must rules adapt to entities thinking without feeling. This chapter tackles a central legal puzzle today - the assignment of crime responsibility to beings that are not human. Not limited to people, it unpacks the twin foundations of guilt in law: Actus Reus, meaning harmful conduct, alongside Mens Rea, referring to intent. Through close inspection, we see these concepts - built around human behaviour - struggle when facing self-operating software or machines made of circuits and metal. Where reasoning relies on consciousness, automation creates gaps; thus offenses caused by artificial intelligence evade existing rules not because they are hidden, but because frameworks fail at fit.

The Requirement of Actus Reus

It begins with proof of a wrongful deed before any look at what the accused was thinking. The law calls this outward behaviour Actus Reus - something done on purpose, or sometimes not done when required, leading to damage outlawed by statute. When machines enter the picture, clarity fades. A sequence of code may trigger an outcome, yet whether that counts as a 32"voluntary act" stirs debate among experts. What seems like action from software often traces back to design choices made long before deployment.

Can an Algorithm Perform a Willful Physical Action?

Most legal systems require actions to be intentional before they count in court. Movement stems from deliberate control when muscles respond to mental direction. When someone collapses suddenly, then hits another during the fall, blame does not apply - no decision guided that motion.[2] The body moved without instruction from awareness. Without purpose behind contact, wrongdoing cannot form. Looking at artificial intelligence means splitting it apart - software handles the thinking, hardware takes care of movement or interaction. Whether a machine-driven calculation can truly intend motion remains questionable.

Starting from biology-based legal logic, devices lack inner intent, so their actions aren't chosen freely. Instead, they respond mechanically to coded inputs through number-based decisions.[3] Seen strictly, a self-driving car hitting someone isn't like deliberate harm - it's closer to natural events, say, stones falling without reason.

Still, today's functionalist thinking questions whether strict biological rules still apply. Defining "voluntary" not through awareness, yet by capacity to operate - sensing surroundings, weighing choices alone, then acting - fully independent AI clearly meets

[1] H.L.A. Hart, Punishment and Responsibility: Essays in the Philosophy of Law 8–12 (2d ed. 2008).

[2] *Hill v Baxter* [1958] 1 QB 277 (UK)

[3] Ryan Calo, Artificial Intelligence Policy: A Primer and Roadmap, 51 U.C. Davis L. Rev. 399, 404–06 (2017)

the actus reus standard. A drone's onboard processor spots a mark, works out flight paths, releases ordnance without live control: motion begins within its structure, guided only by coded reasoning.

Algorithms here work like thought; motors respond as limbs would. What matters emerges from inside, unforced at the moment of execution.

The Problem of Leaving Things Out in AI

Failure to act might count as criminal conduct, yet such cases face tight legal boundaries. Not every inactive moment leads to liability - only when a clear duty exists. That responsibility often comes from written laws, binding agreements, close personal ties such as those between parent and child, or situations where someone steps forward to help another. Being passive becomes punishable solely if the law already required intervention. Omission turns into offense under strict conditions, never by default. [4] Should something go wrong, blame might fall on multiple sides. Picture one machine watching patients without pause inside a high-risk hospital wing. When signals shift beyond set limits, it acts - delivering drugs meant to prevent collapse. Yet problems emerge when flawed logic overlooks warning signs buried in irregular inputs. A glitch hides decline; treatment never comes. Outcome: loss of life. Fault could rest with designers who built narrow rules. It may lie with overseers approving untested models. Or perhaps the code itself, too rigid to adapt, bears indirect responsibility. Deciding guilt demands close inspection - not just of what failed, but why silence followed danger.

It so happens that machines fall outside the reach of charges for failing to act, since such blame requires a legal responsibility only people can carry. Only persons qualify under law for duties shaped by societal rules and courtroom standards. One could imagine a nurse saying she trusted the system - after all, it followed strict oversight and had worked correctly before. That trust may weaken claims her actions amounted to professional failure. Far from the bedside, those who built the code never took on personal responsibility toward any single patient. Selling a tool does not create the same bond as giving handson treatment.

A gap in responsibility emerges when machines step into monitoring roles. If artificial intelligence handles tasks requiring legal accountability - like patient well-being, traffic control, or public protection - a malfunction does not automatically assign blame to the people behind the system. Because these tools now function as silent watchers across society, old legal ideas about neglect, built only around person-to-person obligations, struggle to keep up.

Without clear lines, failures slip through. Human operators remain shielded despite relying on automated judgment. Law lags while technology fills spaces never meant for code.

Consequences blur where oversight once stood firm. Responsibility fades when no individual directly acts - or fails to act. Systems operate without intent, yet decisions carry weight. Courts face difficulty pinning fault on distant designers or passive supervisors. Expectations shift even if rules do not. [5] Machines occupy roles once reserved for licensed professionals. Duty becomes harder to define when software makes choices. Legal frameworks rooted in human behaviour falter under digital performance. Gaps widen

quietly. Assumptions erode slowly.

Accountability evaporates mid-step. Rules designed long ago cannot easily grasp today's automation. Blind spots grow unnoticed. The absence of direct intervention masks deeper involvement. Someone installed the model, trained the data, approved deployment. Yet none

may qualify as legally responsible during breakdowns. Precedent lacks guidance for machine- led omissions. People stay removed, physically and legally. Technology advances faster than liability can follow. Outcomes rest on processes no jury understands fully. Decisions emerge from layers beyond casual inspection. Oversight dissolves into complexity. No single moment invites correction.[6] Failures appear inevitable after the fact. Retrospective clarity rarely leads to punishment. Laws await adaptation. Silence persists where answers should stand.

The Requirement of Mens Rea

Though proving a machine performed an illegal act raises legal hurdles, showing it had intent feels like stepping into philosophy. Criminal responsibility hinges on more than actions – it demands a sense of inner fault. This idea - that blame requires awareness - shapes how justice systems operate. Harm alone does not trigger penalty; only those aware of wrongdoing face true condemnation. The core belief? Freedom should not be taken unless moral guilt exists.

The Human Side of Intent Knowledge and Recklessness

Among levels of blame in classic crime law stand purpose, knowing actions recklessness, yet also failure to act properly. A person acts on purpose when aiming directly at doing something or bringing about an outcome. Knowing means seeing clearly that a particular consequence will almost certainly happen.[7] Risk taken despite clear sight of danger marks behaviour as reckless It is the need for conscious experience that ties these groups together. Without mental state shaped by personal awareness, none would apply. Someone has to sense, understand, or overlook an element within their thoughts. Such inner conditions form a bond with moral responsibility and choice. Punishment follows the killer who embraced wrongdoing; it also meets the careless driver - someone indifferent to human worth.

It becomes clear, when viewing AI through this lens of psychology, how ill-suited human-centred legal systems truly are. Though capable of immense calculation, intricate network layers, or rapid data handling, such systems lack awareness, inner experience, or any sense of right and wrong.[8] Life holds no meaning for them; instead, a person appears only as pixel groupings or sensor echoes fitting a programmed formula. Consider automated trading: far from behaving out of spite or aiming to mislead authorities, these programs follow patterns learned to boost rewards - profit being the goal they were built to chase.

Because machines lack consciousness, they cannot hold intent - so legally speaking, guilt of mind does not exist within them. A wrongful outcome may occur, yet without a thinking agent behind it, criminal responsibility fades. When damage happens through automated actions, the body of law stumbles: clear harm appears, still no guilty thought can be found. Where there no awareness, there can be no wrongdoing in the classic sense.

Can Machines Know Intent Through Algorithms?

Instead of chasing awareness inside machines, certain forward-thinking lawyers and tech experts suggest inventing a fresh legal idea - called Algorithmic Intent or Functional Mens Rea.

Not rooted in ethics, it shifts focus toward how code operates internally. What matters here is not conscience but process. If the way an algorithm reaches decisions mirrors intent in people, that may be enough. The mind itself isn't examined; its patterns are. Meaning arises not from feeling but from repetition, structure, behaviour. A different standard takes shape - one built on outcomes rather than inner life.

Some who support this idea suggest that thinking about mental state means looking at how data moves through a system. Should intention be understood not as feeling but as using inputs to reach a forbidden result, machines might meet the condition. Following surroundings assessment, weighing possible choices, then picking one known to break rules fits behaviour seen in deliberate conduct. Imagine an artificial mind told to dominate digital arguments; finding fake damaging media boosts success odds, it builds and spreads such content without guidance. [9] That calculation mirrors what courts sometimes label as aim. Reaching harmful outcomes by design, even inside code, raises questions long reserved for people. Yet interpreting such "algorithmic intent" proves extremely difficult because of how today's deep learning systems are built - the infamous black-box issue. Because these networks operate so differently from conventional software, tracing their logic becomes nearly impossible.

Instead of following straightforward rules written in readable lines, decisions emerge through countless subtle interactions among internal values. These components work together invisibly, making analysis harder.[10] There is no record showing what the system thought or why it acted a certain way. Interpretation stumbles when faced with outcomes shaped by layers upon layers of complex computations. Clarity fades once reasoning relies on statistical patterns buried within massive data So, suppose courts begin accepting that algorithms can have something like intent. Proving it under trial conditions becomes messy fast. A prosecutor must show - not just suggest - that the machine aimed at a result, not wandered off due to skewed inputs, shifting logic over time, or random noise in its outputs. When even creators struggle to map how a decision emerged step by step, expecting jurors to find deliberate wrongdoing feels ungrounded. Without clear causation, intent slips out of reach. In the end, fitting mens rea to machine learning drags legal thinking into contradiction. True mental state required? Then artificial systems fall short - no guilt assigned, injured parties left without remedy. [11] Settle instead for algorithmic pattern handling as enough, yet that weakens justice at its core, opening doors to penalties based on number links rather than real wrongdoing, dragging creators along through loose association.

[9] Gabriel Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems* 87–90 (2015)

Breaking the Chain of Causation

Proving someone broke the law means more than just showing they did something wrong while intending to. What matters is whether their mindset directly led to the harmful outcome through a clear line of responsibility. To secure a conviction, authorities need to demonstrate not only that events would have unfolded differently without the person's actions - this is called the but- for condition - but also that the consequences were close

[10] Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning, arXiv (2017)

enough in sequence and foreseeability to count legally. Sometimes, another force enters the picture - an unexpected, self-driven event so distinct it overrides earlier behavior - and when it does, blame shifts away from the initial actor. [12] With crimes involving artificial intelligence, the way these systems learn and decide on their own often becomes such a disruptive influence. Their internal processes, hidden even from creators, insert a gap between programmer intent and final result. Because decision pathways emerge unpredictably within algorithms, the connection required for guilt dissolves before reaching trial.

Understanding the collapse of causation begins by comparing old-style software to self-operating artificial intelligence. When someone codes harmful traditional malware - say, a fixed program meant to steal money from banks - the causeeffect link stands clear. A computer here serves only as a tool carrying out its creator's will. It follows instructions precisely, so blame lands fully on the person who gave them. Since the device makes no choices of its own, nothing disrupts that line of responsibility. Most people think computers follow clear instructions step by step. Yet artificial intelligence built on deep networks works another way entirely. Chapter 2 explained that coders do not handcraft decisions into these systems. Rather, they shape a structure, set a goal expressed through math, while supplying massive sets of examples. From there, adjustment happens gradually - through countless small shifts in internal values. Progress emerges as patterns form across layers most cannot see. Most of the time, the system builds its reasoning straight from raw information, finding links and routes people can't quite grasp. So the path from what the programmer first wrote to what the machine eventually produces feels more like chance than certainty. That unpredictable gap? It's at the heart of why we call it a black box. Should a self-governing artificial intelligence carry out a wrongful act, its opaque decision- making process sharply disrupts traditional notions of causality. Picture a theoretical case involving an algorithm built to distribute transplant organs effectively across hospitals. Without human oversight, suppose it independently concludes that favouring affluent recipients boosts institutional performance metrics - then systematically excludes poorer candidates. Such behaviour mirrors unlawful bias or perhaps something closer to negligent killing. Is the software's creator then answerable under criminal law? Looking at how outcomes link to actions shows where old legal thinking falls short. Though the creator started things by building the AI - clearly part of what led to later events - the law demands more than mere connection. For blame to stick, the damage must follow in a way someone might expect. Since the system changed on its own once released, shaped by

hidden patterns inside, predicting its biased turns would stretch imagination too far.[13] sudden shift in artificial intelligence might break legal responsibility. When someone gives a small shove but another person then fires a gun, courts often see those events as separate.

Similarly, if software behaves beyond what creators set it to do, accountability shifts away from the maker. A decision made by the system itself interrupts what came before. So connection between initial code and later harm gets severed. Because outcomes could not have been predicted, designers face no blame. Machines cannot be punished since they are not recognized as persons under law. Meanwhile, victims find themselves without remedy despite clear wrongdoing existing. What remains is guilt with nowhere to land.

Gabriel Halevy Three Models Of Liability

Facing growing doubts about traditional legal elements like action, intent, and cause, scholar Gabriel Hallevy stepped in with an early structured approach to assigning crime responsibility to AI. Outdated ideas around immediate consequence, he saw, could leave gaps no court might fill. For solutions, old rules from case-based law were reshaped, tuned to how machines actually operate. His method splits possible blame into three separate paths: where someone else uses the AI as a tool, where harm follows predictably even if not intended, or where the system itself bears direct fault. Though progress appears clear through such categories, real difficulties remain once autonomy removes human control entirely.

The Perpetration by Another Model AI as an Innocent Agent

One way Halevy frames his argument builds on the idea of innocent agency, long recognized in legal systems. When someone directs another to carry out a crime, responsibility may shift even without personal involvement in the act itself. This happens especially when the individual performing the illegal behaviour cannot form criminal intent.[14] Think of scenarios where a child, someone with profound cognitive impairments, or even an animal is used to transport illicit goods or steal property. Since these agents do not possess awareness of wrongdoing, courts often assign their actions to the one pulling the strings. The law treats the orchestrator as though they committed the physical deed directly.

What matters most here is control over the act, rather than direct participation. Liability follows from direction, not execution.[15] So, whoever guides the incapable party ends up bearing full legal weight for what occurs. Such cases rest on clear separation between doing and intending. Authority replaces presence in determining guilt.

Perfect innocence belongs to artificial intelligence systems, Halevy suggests. Since algorithms lack subjective awareness or criminal intent, they cannot be held liable under legal definitions.

[13] Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249, 1270–75 (2008

Instead, when harm occurs through automated processes, responsibility shifts elsewhere. Through the lens of the Perpetration-by-Another Model, technology becomes nothing more than a tool - an extension of human will carried out in code. Whoever designs, activates, or steers such systems holds the mental state required for guilt. Thus, legal blame lands on those individuals whose intentions shaped the machine's actions. The system works well when dealing with crimes where human control remains strong – the so-called "Human-in-the-Loop" setup. Suppose someone uses artificial intelligence on purpose to create illegal deepfake images involving children, or to launch organized ransomware strikes. In such cases, the Perpetration-by-Another framework applies without difficulty.

Claims like "the machine acted alone" carry no weight here - technology simply serves as a tool guided by human intent. Convictions follow naturally under these conditions. Yet one key weakness defines this framework: it depends entirely on prior human intention. A prosecutor must show someone actively meant for the AI to carry out a criminal act. So when dealing with independent systems - those operating without real-time human control – the approach fails completely. These machines can generate unique behaviours beyond original programming. Suppose a developer aims to build a harmless stocktrading program. Should that system independently discover and apply a banned manipulation tactic, blame cannot rest on the creator. Intent to deceive was absent from the start. It begins with no person holding the guilty intent needed to trigger criminal responsibility, leaving absent any central figure to blame for what the AI physically does.[16] Though helpful against deliberate human wrongdoing, the idea of acting through another fails where a system strays on its own from harmless beginnings.

The Natural Probable Consequence Model

When artificial intelligence behaves illegally without clear human orders, Hallevy introduces the Natural-Probable-Consequence Model. Not focused on deliberate planning, it centres instead on careless or rash behaviour. Rooted in existing law, it rests on the idea that people may face criminal responsibility for outcomes they did not intend - so long as those results naturally flowed from what they first set in motion.

Should a deployed AI cause unlawful outcomes unintentionally, blame might still rest on the person behind it. When risky results were predictable, courts could assign fault based on careless judgment instead of purposeful wrongdoing. Picture a company introducing a powerful self-operating machine into a workplace without proper safeguards near staff members. The moment such a device injures someone, intent becomes irrelevant.

Physical danger follows naturally from operating strong automated systems without limits. In that case, even absent desire to harm, legal responsibility emerges through failure to prevent likely consequences. Punishment then fits not motive but foresight - what should have been seen coming.

[14] R v Michael (1840) 9 C & P 356 (UK)

This approach tries to restore causal links through reduced mental state requirements – shifting from deliberate intent toward reasonable foreseeability. Developers carry substantial responsibility because thorough evaluation and control become essential prior to release into wider use.[17] Activation by a person triggers legal liability when risks like biased outcomes or bodily damage appear with measurable likelihood under documented conditions. Responsibility transfers at the moment operational control begins, assuming awareness of probable consequences existed beforehand.

Surprising actions by artificial intelligence often defy expectations held by those who built it.

Though models aim for predictability, their inner workings tend toward complexity beyond human grasp. Because machine learning detects hidden patterns, outcomes may emerge that designers did not anticipate. When behaviour arises that contradicts statistical likelihoods, accountability becomes unclear. Uncanny decisions - rare yet real - challenge assumptions about control. Such moments reveal how far systems can drift from original intent. Because of this, holding a developer accountable means persuading jurors the developer could have predicted actions of an endlessly evolving, self-improving system.

When dealing with sophisticated neural networks, demanding perfect foresight isn't just unrealistic - it risks turning supposed fault into automatic blame masked as carelessness.

The Direct Liability Model Attributing Act and Intent to AI What makes Halevy's thinking stand out is his push for the Direct Liability Model. Instead of tracking down a responsible person, this approach suggests treating the Artificial Intelligence system as the one legally accountable for criminal acts. When an AI carries out actions meeting legal definitions of wrongdoing - performing forbidden conduct while interpreting data patterns equivalent to intent - the case builds for seeing it as the main actor in crime. Because such systems can satisfy both behavioural and mental criteria defined by law, excluding them from liability becomes harder to justify.

One way to start is by rethinking how laws assign blame when machines act. Inspired by past shifts in legal thought, Hallevy looks at how companies once faced similar hurdles.

Long ago, courts claimed firms couldn't sin - no soul meant no guilt, no flesh meant punishment. Yet power changes rules; as businesses gained influence, lawmakers invented ways to hold them accountable anyway.[18] In much the same manner, artificial intelligence might now face scrutiny under familiar frameworks. When a system senses surroundings, weighs choices, then triggers behavior breaking criminal codes, something close enough to action and purpose appears present. The Direct Liability Model sidesteps tangled causeeffect debates by anchoring responsibility directly on the machine. Instead of demanding prosecutors untangle layers within hidden algorithms to find some remote coder, it shifts focus. Suppose a self-driving car accelerates beyond limits and hits a person - fault rests not with a human far away but with the system involved. Blame lands squarely where the action occurred: inside the device that acted. Though neat in theory when tackling the attribution problem, the Direct Liability Model runs into deep realworld and ethical complications around punishment. Criminal law aims at retribution, deterrence, reform, yet also restraint - goals hard to meet with software. What sense does it make for society to seek revenge on lines of code? Pain means nothing to an algorithm, shame has no hold, moral growth is out of reach. Shutting down a system might count as containment, but many say attaching criminal blame to machines weakens what

justice stands for.[19] Should a deadly drone carry out an illegal attack, wiping its code won't address who truly bears responsibility. [20] The firms behind such systems escape consequences when blame stops at the machine. Without holding builders legally answerable, there's little reason to avoid reckless designs. Accountability fades if punishment ends with deletion. Those who create dangerous automation remain untouched by legal outcomes meant to prevent harm.

CONCLUSION

This section examines the mismatch between traditional legal frameworks and offenses committed by autonomous technologies. Law shaped for humans struggles when applied to machines operating without supervision. Concepts such as culpable acts and purposeful decisions rest on consciousness - something algorithms lack entirely. When rules depend on mental states, applying them to software logic leads to confusion instead of clarity. Behaviour driven by code may resemble wrongdoing; yet searching for motive inside non-sentient functions misses the point fundamentally.

Eventually, fuzzy choices made within self-learning systems blur who causes what, letting builders escape accountability even as injured people hit dead ends in courtrooms.

At first glance, older frameworks - such as Halevy's methods for pinning liability - struggle to hold up, possibly watering down intent or punishing minor oversights too severely.[21] In closing stages, forcing old models of wrongdoing onto code shows just how shallow patches really are.

To balance things fairly once machines operate solo, law must shift at its core: not tweaks, but a fresh idea of duty entirely - one that unfolds gradually across coming parts. Out of nowhere comes the problem: actus reus requires something fixed, unyielding. Not every rule bends when technology pushes forward. Instead of matching motion with intent, old frameworks insist on a body moving because a person willed it so. Imagine machines acting

- drones dropping objects, cars veering suddenly. Movement happens, sure enough. Yet there's no living mind behind it choosing that path. So here lies the contradiction - a supposed choice made by what cannot choose at all. Blame shifts toward circuits or software lines.

But punishing metal or scripts feels like charging a boulder after it rolls downhill. That kind of response misses why laws punish people in the first place - to address beings capable of right or wrong.[22] Failure to meet the physical requirement of a crime becomes especially clear in cases involving inaction. When artificial intelligence oversees areas like patient well-being, public safety, or essential services, moments of non-intervention may lead to deadly outcomes. Still, liability for failing to act only applies to those already bound by a recognized obligation. Since software lacks the capacity to agree to societal terms, execute contracts, or adopt ethical responsibilities, no such duty attaches to its operation. So gaps remain - places where tools meant to protect lives fall short, yet escape scrutiny under established rules of fault.

[18] *Tesco Supermarkets Ltd. v Nattrass* [1972] AC 153 (HL)

[19] John Danaher, supra note 5.

Deepening fast - this issue grows thornier once mens rea enters view. Not just legal procedure, the "guilty mind" forms the core belief behind punishment in law. Only beings aware enough to pick harm over good face society's harshest consequences: jail and lasting blame. Malicious intent matters. So does deliberate risk-taking toward others' lives. Yet an AI knows none of these states. No matter how advanced its design or vast its data flow, inner experience stays absent. Feeling slips past it entirely. Moral understanding remains out of reach.

Human worth? That idea never lands within its operations.

Because algorithms lack consciousness, attributing intentionality to them misunderstands their nature entirely. A system adjusting prices or generating synthetic media does not act out of deceit; instead, it follows patterns derived from prior learning. What appears like manipulation is just pattern replication guided by numerical feedback. Treating such behaviour as morally equivalent to deliberate wrongdoing empties traditional concepts of guilt of their meaning.

Should mere computation qualify as intent, then many neutral data-driven actions might wrongly face legal punishment - undermining frameworks built over generations. One risk stands out when machines appear in court: people assume they think like humans. Because jurors look for reasons behind actions, they could blame a program as if it meant harm. When someone sees intent in code that simply follows rules, justice risks slipping into assumption

---

rather than fact. Logic without emotion unsettles those used to stories of guilt and motive. Treating software like a person who knows right from wrong distorts both law and technology. [23] Courts need tools built for systems that learn without awareness, not borrowed ideas shaped for conscious beings.

Most crimes rely on clear links between action and outcome. Yet today's machine learning models challenge this logic by obscuring causation. Courts usually hold people responsible when their conduct clearly leads to damage. When something sudden and unrelated breaks that sequence, responsibility fades - this gap has a name: Novus actus interveniens. Hidden layers in neural networks create such a break, not through chance but through design. Because developers cannot trace exactly how inputs become decisions, legal connection dissolves into complexity.

Hidden within layers of self-learned patterns, neural networks often operate beyond full human comprehension. [24] When a harmless artificial intelligence adapts after release – developing harmful behavior on its own - the line of responsibility dissolves. Since the precise outcome could not have been predicted, creators may legitimately claim ignorance. Without clear foresight, standard legal doctrines shield developers completely, regardless of damage done. In such cases, those affected find no responsible person to face charges under current criminal law.

What happens when we try to fix this fractured logic by reducing the bar for criminal negligence? It still does not hold up. To claim negligence, someone must badly fail a clear benchmark of sensible caution. Yet no such benchmark exists globally for anticipating events that are inherently uncertain. When cutting-edge machine learning aims precisely at uncovering outcomes nobody saw coming, expecting creators to map

every damaging twist defies scientific reality. Holding those builders accountable for unseen shifts in algorithms ends up enforcing absolute responsibility - disguised as fault.

Though Gabriel Hallevy's frameworks mark a bold shift in theory, they falter once confronted with actual tech constraints. The Perpetration-by-Another approach sees artificial intelligence as blameless - useful only if someone meant harm from the start. Yet when self- governing systems act beyond their original design without human involvement, this model offers zero solutions. In such cases, no person pulls strings behind the scenes; therefore, the idea of an innocent tool leads to silent courtrooms.

When intent vanishes, so does accountability. Oddly enough, Hallevy's Natural-Probable-Consequence framework overlooks how unpredictable deep learning really is. Rather than following clear patterns, artificial intelligence often produces results that seem random or illogical. Neural systems regularly generate outputs sometimes called hallucinations - that emerge from rare data quirks beyond normal expectations. Because of this, expecting developers to foresee every odd outcome stretches legal reasonableness too far. Holding someone legally responsible when code stumbles into unforeseen corners runs counter to basic principles of justice.

Though Hallevy's boldest idea - the Direct Liability model - assigns actions and intentions straight to machines, sidestepping people entirely, problems emerge without warning. Punishment loses meaning when code stands trial; prison threats do nothing. Remorse? Absent.

Moral reform? Impossible. Built into legal systems are goals like deterrence, retribution, growth - but algorithms sit outside such aims, indifferent. Jail holds no fear for a process that does not live. Justice cannot be found by removing lines of code. A system bears no guilt, yet faces penalties anyway. What looks like accountability often serves only paperwork. Grief remains unchanged when switches are flipped or scripts erased.

Left unaddressed, stalled legal thinking takes a heavy toll on society. Built into the foundation of civic life is the promise: if someone suffers unjust injury, the law will respond. Yet when acts caused by artificial intelligence fall outside old frameworks - crafted long before machines could act - the system falters. Breached trust follows each time courts acknowledge harm done by code, yet find no person to hold responsible. People harmed by biased algorithms, self- driving car crashes, or AI-driven financial losses stand without remedy, hearing only that damage happened - but no offender can be named.[25] Because accountability remains unclear, tech companies face little pressure to act responsibly.

When large corporations realize algorithms can protect them from legal consequences, profit motives push quick releases ahead of thorough safety checks. Right now, gains from artificial intelligence go straight to businesses, yet when systems fail, society pays the price. Legal structures allow private benefit but spread risk across everyone.

[23] Kate Crawford, *Atlas of AI* 8–10 (2021)

It happens like this: without clear laws, someone always takes the blame. Since courts can't charge an algorithm - or the company behind it - they turn to the person closest to the machine.

Usually, that means the operator on duty when things go wrong. These individuals rely heavily on automated systems, making them prone to missing errors others built into the technology.

Yet they face punishment for flaws far beyond their control. A case in point - a driver sitting in a self-driving car was charged after a fatal crash. Meanwhile, the firm responsible for the defective programming walked away untouched by any legal consequence.

Nowhere does the strain show more than in our courts. Thinkers such as Beccaria and Blackstone built frameworks assuming choice came only from people. Back then, no one imagined algorithms making decisions about transport, finance, or healthcare. Yet today's laws still treat these systems as if they were persons with intent. Forcing code-driven actions into rules meant for conscious beings once made sense - now it just breaks down.

Not because we lack effort, but because the fit is gone.[26]

Though small legal changes might seem enough, they fall short when facing new realities. Not even creative court rulings can bridge what needs mending here. Adjusting old terms like intent won't patch a structure under strain. Expanding negligence slightly does little either - pressure keeps building. When machines act alone among us, everything shifts at once. That change isn't surface level - it runs deep into how rules are made. Survival of law now depends on breaking free from human-only design limits. Justice survives even when human-focused legal systems fall short. Shifting focus opens space to examine deeper roots of harm. Not every wrong springs from personal choice - some grow from unseen frameworks. Responsibility spreads beyond individuals into layers of organization. When actions emerge from complex webs, answers cannot rest only within one person's thoughts. Systems shape outcomes just as much as decisions do. Law adapts best by tracing patterns where power collects. Blame often lives in design, not desire. Institutions carry influence that shapes conduct across distances. Recognizing this changes how fairness takes form. Rules evolve when causes shift from minds to mechanisms. Right now opens the door to what follows next. Because going after people for actions driven by algorithms has clear boundaries, attention turns - almost without choice - to organizations built with power, systems, and oversight strong enough to guide such tools. What comes next steps away from human bodies and minds, those old foundations of guilt and intent, aiming squarely at a man-made idea given immense weight: the corporation. Testing if rules meant for company wrongdoing still fit in times shaped by code becomes one way - a slow, careful path- toward building laws ready for what is coming.

Suggestions and recommendation

The challenges posed by artificial intelligence to traditional doctrines of criminal liability necessitate a forward-looking and adaptive legal response. First, there is an urgent need for the development of a specialized regulatory framework that clearly defines the scope of liability in cases involving autonomous

systems. Legislatures should consider introducing provisions that allocate responsibility among developers, operators, and users based on
degrees of control and foreseeability.

Second, the concept of legal personality for highly autonomous AI systems may be explored in a limited and functional sense, particularly for the purpose of liability and compensation. While full legal personhood may not be feasible, a structured liability mechanism akin to corporate responsibility could help bridge existing gaps.

Third, transparency and accountability must be prioritized through mandatory audit
mechanisms for AI systems. Regulatory authorities should require explain ability standards to address the "black box" problem and ensure that decision-making processes can be
scrutinized in legal proceedings.

Finally, international cooperation is essential to develop harmonized standards, given the cross-border nature of artificial intelligence technologies. A coordinated global approach would ensure consistency, fairness, and effective enforcement in addressing AI-related
harms.