

Artificial Intelligence and Criminal Responsibility: Reconstructing Mens Rea and Attribution for AI-Assisted Cybercrime in Pakistan

Syed Mudassir Shah

Lecturer, University of Southern Punjab (USP), Multan

Email: shahmudassir0@gmail.com

Abstract

Artificial intelligence is changing the scale, speed, personalization, and technical complexity of cybercrime without eliminating the need to attribute criminal responsibility to human actors. In Pakistan, the principal substantive framework remains the Prevention of Electronic Crimes Act 2016, as amended in 2025, supplemented by the Pakistan Penal Code, the Qanun-e-Shahadat Order 1984, the Electronic Transactions Ordinance 2002, and the National Forensics Agency Act 2024. These laws are largely technology-neutral, yet AI-assisted offending creates evidential and conceptual pressure on traditional categories of intention, knowledge, authorization, causation, participation, and digital attribution. A person may use a generative model to draft phishing messages, automate reconnaissance, generate malicious code, create deepfakes, or optimize fraud while delegating numerous operational choices to software. The central legal problem is therefore not whether an AI system itself possesses criminal mens rea, but how courts should infer the mental state and causal contribution of programmers, operators, deployers, account holders, collaborators, and beneficiaries when machine output intervenes between human instruction and prohibited result. This paper argues against premature recognition of AI as a criminal subject. Current systems lack the normative agency, legal status, and punishment-sensitive interests that make criminal blame meaningful. Pakistani law should instead adopt a human-centred attribution framework built around purposeful use, knowledge of capabilities, control over deployment, foreseeability of system behaviour, acceptance of risk, contribution to the offence, and corroborated digital evidence. The paper proposes a four-layer attribution model covering identity, control, mental state, and causal contribution, together with evidential safeguards for logs, prompts, model outputs, account records, device forensics, and chain of custody. Reform should clarify AI-assisted liability without creating strict criminal responsibility for developers or users merely because an unpredictable model produced harmful output.

Keywords: Artificial Intelligence; Cybercrime; Mens Rea; Attribution; PECA 2016; Digital Evidence

1. Introduction

Artificial intelligence does not create an entirely new category of crime. It changes how familiar crimes can be planned, executed, scaled, concealed, and personalized. Generative systems can produce persuasive phishing content, synthetic voices, false images, malicious scripts, reconnaissance summaries, and automated conversational fraud. Machine-learning tools can assist password attacks, vulnerability identification, evasion, and social engineering. The significance for criminal law lies in mediation: a human may set a goal, provide data or prompts, configure tools, and launch a process while software selects intermediate steps or generates content not specified word for word.

Traditional criminal responsibility is organized around human conduct and mental state. Liability ordinarily requires a prohibited act or omission attributable to the accused and the mens rea specified by the offence. Cybercrime already complicates attribution because conduct occurs through accounts, devices, networks, cloud services, intermediaries, and pseudonymous infrastructure. AI adds another layer by making outputs probabilistic, adaptive, and sometimes only partly foreseeable.

Pakistan's principal cybercrime statute is the Prevention of Electronic Crimes Act 2016, significantly amended in 2025. PECA criminalizes forms of unauthorized access, copying or transmission of data, interference with information systems or data, critical-infrastructure offences, electronic fraud, identity-related offences, cyberstalking, and other electronic misconduct. The 2025 amendments altered institutions and additional regulatory provisions, including the investigative architecture involving the National Cyber Crime Investigation Agency (Prevention of Electronic Crimes Act, 2016, as amended 2025). Nothing in the core logic of these offences requires the offender to type every command manually. A person can "cause" unauthorized interference through an automated tool just as one can cause it through conventional malware.

The harder issues are evidential and normative. If a model produces harmful code after ambiguous prompting, what must prosecutors prove about the user's intention? If an autonomous agent chooses a target beyond the operator's specific instruction, is the result attributable to the operator? If a developer releases a dual-use model later misused by criminals, when, if ever, does knowledge of misuse become criminal participation? If several persons control different parts of infrastructure, how should courts distinguish technical association from culpable contribution?

This paper argues that Pakistani law should resist two opposite errors. The first is anthropomorphism: treating AI output as though the machine itself had intention, motive, or blameworthiness. The second is automation exceptionalism: assuming that human responsibility disappears whenever software makes intermediate choices. A better approach reconstructs mens rea and attribution around the human-AI relationship. Courts should examine what the person intended, knew, controlled, authorized, monitored, accepted, and contributed, while using reliable digital evidence to connect the person to the relevant system and prohibited result.

2. Methodology and Conceptual Scope

The paper uses doctrinal and comparative legal analysis. It examines PECA, related evidentiary and electronic-transactions legislation, the institutional role of the National Forensics Agency, and general principles of criminal responsibility. Because Pakistani reported case law specifically addressing generative AI-assisted cybercrime remains limited, comparative scholarship and international cybercrime materials are used to identify conceptual models rather than to import foreign law mechanically. Relevant sources include Europol's analysis of large language models, the United Nations Office on Drugs and Crime's work on AI-enabled criminal innovation, NIST's adversarial machine-learning taxonomy, Council of Europe materials on cybercrime and electronic evidence, and recent scholarship on AI-related criminal responsibility.

The term "AI-assisted cybercrime" refers to offences in which artificial intelligence materially assists planning, execution, targeting, deception, adaptation, concealment, or exploitation. It does not require a highly autonomous artificial general intelligence. Most current cases involve human use of AI as a tool or semi-autonomous agent. The legal questions therefore concern ordinary persons and organizations operating through technically advanced systems.

Three categories should be separated. In tool-use cases, the human directs the essential criminal objective and AI improves efficiency. In delegated-execution cases, the human sets goals and parameters while the system makes meaningful operational choices. In unexpected-output cases, the system produces harmful conduct that the human did not specifically request and may not have foreseen. Liability is easiest in the first category and most difficult in the third. A sound doctrine must preserve this gradient rather than impose one rule on all AI involvement.

3. Pakistan's Substantive Cybercrime Framework

PECA is drafted primarily in terms of conduct directed at information systems, data, identity, communications, and electronic deception. Its technology-neutral language gives it considerable capacity to address AI-assisted offending. Unauthorized access remains unauthorized whether achieved by a human typing commands, a script, or an AI-enabled agent. Interference with data or

systems remains legally significant whether the code was handwritten or machine-generated. Electronic fraud can be committed through synthetic media or automated persuasion if the statutory elements are satisfied.

This point is important because public debate can overstate the need for a new offence labelled "AI crime." Novel technology does not necessarily create a lacuna in substantive criminalization. The first question should always be whether existing offences already capture the conduct. Overlapping statutes can create complexity, but unnecessary AI-specific offences risk duplication, vague drafting, and punishment based on technology rather than harm and culpability.

PECA's mental elements vary across offences and must be read carefully. Terms such as intentional, dishonest, fraudulent, unauthorized, knowingly, or without authorization perform different functions. AI does not lower the prosecution's burden to prove the mental state applicable to the accused. If an offence requires fraudulent intent, the state cannot substitute evidence that an AI system generated deceptive material for evidence that the accused intended the deception.

The Pakistan Penal Code may also apply where cyber conduct forms part of cheating, forgery-related activity, criminal intimidation, conspiracy, abetment, or other general offences, subject to statutory overlap and specific legal provisions. The key principle is complementarity: cyber statutes define technology-mediated wrongs, while general criminal law supplies broader participation and responsibility concepts where applicable.

The 2025 PECA amendments are relevant institutionally as well as substantively. They recognize the National Cyber Crime Investigation Agency as the principal investigation agency for offences under the Act, subject to the amended statutory scheme (Prevention of Electronic Crimes Act, 2016, s. 30, as amended 2025). AI-assisted cases will test the technical capacity of investigators to preserve model interactions, cloud records, prompts, API logs, virtual assets, and cross-border service-provider data. Substantive adequacy is therefore inseparable from evidential capability.

4. Why Mens Rea Becomes Harder with AI

Mens rea is not a metaphysical inquiry into what a machine "thought." It is a legal inquiry into the mental state of the accused human actor at the relevant time. AI complicates that inquiry because intention may be expressed indirectly through prompts, system configuration, target selection, model choice, tool permissions, feedback loops, and acceptance of generated outputs.

In a simple phishing case, a user may instruct a model to draft messages impersonating a bank, provide stolen customer data, request urgency language, and send the messages through an automated campaign. The user's fraudulent intention is readily inferable from the overall conduct even though the exact wording came from the model. The model functions as an instrument.

A harder case arises where a user asks an autonomous security agent to "gain access" to a competitor's network and authorizes it to discover vulnerabilities and deploy code. The agent selects a particular exploit and causes damage not specifically contemplated. The user may still possess the requisite intent for unauthorized access and knowledge that interference is a likely consequence, depending on the statutory offence and evidence. Delegation does not erase culpability where the criminal objective is clear.

The most difficult case is an apparently legitimate instruction that produces unexpected harmful output. Suppose a developer deploys an AI agent to test the developer's own systems, but a configuration error causes scanning or exploitation of third-party systems. Criminal liability should not arise merely from causation. The prosecution must establish the mental element required by the relevant offence. Negligence, recklessness, knowledge, and intention are not interchangeable, and courts should resist hindsight reasoning.

Recent scholarship on "hard AI crime" highlights cases in which harmful AI conduct cannot fairly be linked to a culpable human state of mind (Lagioia & Sartor, 2024). Such cases reveal the limits of criminal law rather than justify fictional machine *mens rea*. Civil, administrative, product-safety, or regulatory mechanisms may address harms where human criminal culpability cannot be proved. The principle of legality is better served by tolerating some non-criminal harmful outcomes than by converting serious offences into strict liability without legislative authorization.

5. AI as Instrument, Agent, or Autonomous Actor

Three conceptual models appear in the literature. The first treats AI as an instrument or innocent agent through which the human perpetrates the offence. This model works well when the person intentionally uses AI for a criminal purpose. It resembles responsibility for using malware, bots, or another non-responsible mechanism.

The second model attributes consequences to a human where the AI's harmful conduct was a natural, probable, or sufficiently foreseeable consequence of the person's deployment. This model is more controversial because foreseeability alone does not establish every form of *mens rea*. It may support recklessness or negligence where those mental states are legally sufficient, but it cannot automatically substitute for specific intent.

The third model proposes direct criminal liability of AI. That approach is currently unsuitable for Pakistan. Existing AI systems do not possess legal personality in criminal law, do not experience punishment in the sense relevant to desert, and lack the social and moral agency on which blame ordinarily depends. Assigning them *mens rea* would risk metaphor being mistaken for doctrine.

The distinction between technical autonomy and legal agency is critical. A system may autonomously select actions within a task, but autonomy in engineering does not prove capacity for criminal responsibility. A thermostat can operate without moment-to-moment human instruction; that does not make it a legal subject. Contemporary AI is vastly more complex, yet the normative gap remains.

Pakistani law should therefore treat AI as part of the causal and evidential architecture of human conduct. The degree of autonomy matters because it affects what a human could foresee and control, not because it transforms software into an offender. This framing keeps criminal responsibility anchored to persons while allowing doctrine to respond to increasingly delegated execution.

6. Reconstructing Intention and Knowledge

Intention in AI-assisted cybercrime should be inferred from the accused's objective, instructions, configuration choices, target selection, and subsequent use of output. Prompts are potentially powerful evidence, but they should not be overvalued. A single prompt may be ambiguous, satirical, exploratory, or part of defensive research. Investigators should examine the full interaction history, surrounding communications, files, financial flows, and actual deployment.

Repeated refinement can be especially probative. A sequence requesting more convincing impersonation, evasion of security filters, credential harvesting, or concealment may show purposeful progression. Likewise, directing an agent toward a particular victim, supplying stolen credentials, or connecting output to exploit infrastructure can demonstrate intention beyond mere curiosity.

Knowledge requires similar contextual reasoning. A user may not know every technical detail of generated malware yet can know that the program is designed to access systems without authorization. Criminal law routinely distinguishes knowledge of essential facts from mastery of mechanism. A person who hires another to commit an offence need not understand the precise technique. AI delegation should be treated similarly where the accused knows the essential unlawful character of the objective.

However, courts should avoid inferring knowledge solely from technological sophistication. A cybersecurity professional may possess advanced tools and offensive knowledge for legitimate

reasons. Dual-use capability is not *mens rea*. The prosecution must connect expertise to the alleged criminal purpose through evidence.

Model warnings can also be relevant but are not dispositive. If a system warns that a requested action would be unauthorized or harmful and the user deliberately circumvents safeguards, that conduct may support an inference of knowledge. Conversely, absence of a warning does not establish innocence where the criminal nature of the act is otherwise obvious.

7. Recklessness, Negligence, and Foreseeability

AI systems can behave unpredictably, which makes foreseeability central to borderline cases. Criminal law should distinguish conscious risk-taking from mere failure to predict. Recklessness ordinarily requires awareness of a substantial or unjustified risk, depending on the governing doctrine, while negligence concerns failure to meet a required standard of care. PECA offences should not be silently converted to negligence offences when their text requires intention or knowledge.

Foreseeability is nevertheless important in several ways. It can help establish what risks an operator consciously accepted, whether a harmful output was within the system's known capabilities, and whether the operator's safeguards were genuine or deliberately inadequate. A person who configures an agent to exploit "any vulnerable host" on the public internet cannot plausibly rely on the agent's target selection to deny awareness of unauthorized intrusion.

By contrast, a developer who deploys a defensive model within a controlled environment and encounters a novel emergent failure may lack criminal *mens rea* even if better engineering could have prevented harm. Regulatory negligence may exist, but serious criminal responsibility should not be inferred merely because harm was technically foreseeable in hindsight.

NIST's adversarial machine-learning taxonomy demonstrates the range of attacks and misuse scenarios affecting predictive and generative systems, including evasion, poisoning, privacy, and misuse attacks (Vassilev et al., 2025). Such technical taxonomies can inform what risks professionals reasonably understand, but they do not themselves define criminal standards. Legal culpability must remain offence-specific.

8. Attribution: From Device to Human Actor

Digital attribution is often the decisive issue in cybercrime. An IP address, device, account, wallet, or API key can connect infrastructure to activity, but it does not automatically identify the human who controlled the relevant act. Shared devices, compromised accounts, remote access, VPNs, cloud services, botnets, and credential theft can produce misleading associations.

AI makes attribution more complex because critical evidence may be distributed across several layers: the user's device, a cloud model provider, an API gateway, an orchestration platform, a code repository, command-and-control infrastructure, and the victim system. Investigators must reconstruct the chain rather than rely on a single artefact.

A robust attribution analysis should separate identity from control. Identity asks who is associated with an account or device. Control asks who actually directed the relevant session or deployment. A registered account holder may not be the operator; an operator may act through another person's credentials. Evidence of login patterns, multi-factor authentication, device identifiers, payment records, session tokens, time correlations, and local artefacts can support or weaken the inference.

Attribution should also separate control from culpability. A system administrator may control infrastructure used by others without sharing their criminal purpose. A model provider may technically host a service without knowing a particular user's conduct. Criminal responsibility requires evidence linking the person not only to infrastructure but to the prohibited act and required mental state.

The distinction between technical, operational, and legal attribution used in cyber-security scholarship is useful. Technical attribution identifies infrastructure and methods; operational attribution links activity to an actor or group; legal attribution asks whether admissible evidence proves a particular accused's responsibility to the required standard. Courts should not permit high-confidence technical assessment to substitute automatically for legal proof.

9. A Four-Layer Human Attribution Model

Pakistan can develop a coherent approach through four cumulative layers: identity, control, mens rea, and contribution. The model does not create new offences; it organizes proof under existing law. Layer one is identity. The prosecution should show why the accused is connected to the accounts, devices, credentials, financial instruments, or infrastructure involved. Authentication evidence, subscriber information, device seizures, account recovery data, and communications may be relevant. Layer two is control. The prosecution should establish that the accused controlled or materially directed the relevant AI interaction or deployment. Relevant evidence includes prompts, configuration files, API calls, access logs, scheduling tasks, code commits, agent permissions, and instructions to collaborators.

Layer three is mens rea. The court should identify the mental state required by the charged offence and connect it to evidence. The question may concern intention to obtain unauthorized access, knowledge that data would be copied, fraudulent purpose, or another statutory state of mind. Model output cannot supply a missing human mental element. Layer four is contribution. The prosecution should prove that the accused's conduct materially contributed to the prohibited result or attempt. A person who merely downloaded a publicly available model but never used it in the offence is not equivalent to a person who configured and deployed the operation. Causation, attempt, abetment, conspiracy, and joint liability should be analysed according to ordinary principles. The layers prevent common shortcuts. Identity without control is insufficient. Control without mens rea may describe an accident. Mens rea without contribution may amount to fantasy or preparation rather than the charged offence. Contribution without the required mental state may support another form of liability, but not necessarily the offence alleged.

10. Digital Evidence and the Qanun-e-Shahadat Framework

AI-assisted cases will depend heavily on electronic evidence. Pakistan's evidentiary framework permits modern-device evidence, and Article 164 of the Qanun-e-Shahadat Order has long provided a route for evidence made available through modern devices or techniques, subject to authenticity and reliability. The Electronic Transactions Ordinance 2002 reinforces legal recognition of electronic records and provides criteria concerning integrity, retention, origin, destination, date, and time (Electronic Transactions Ordinance, 2002, ss. 3-6).

Prompts and model outputs should therefore be treated as electronic records requiring authentication, not as self-proving transcripts. Investigators should document how records were obtained, whether they came from the device or provider, how timestamps were generated, whether content can be edited, and what account or session identifiers connect the record to the accused.

Logs can be especially valuable because they may record API requests, tool calls, execution steps, model versions, system instructions, and outputs. Yet logs are not infallible. Providers may retain limited data; users may delete local artefacts; timestamps may use different time zones; automated agents may create events without direct human input. Expert interpretation is often necessary.

Chain of custody remains fundamental. A sophisticated AI case can still fail if devices are improperly seized, forensic images are not verified, or extracted data cannot be shown to remain intact. The National Forensics Agency Act 2024 creates an institutional framework that can support standardized forensic practice and capacity for digital evidence. Coordination between NCCIA investigators, prosecutors, and forensic experts will be essential.

Deepfakes create a special authentication problem. Audio or video that appears to show a person speaking may be synthetic. Courts should avoid intuitive reliance on realism and instead consider metadata, source provenance, compression history, device evidence, expert analysis, and corroborating circumstances. The growth of synthetic media makes authenticity a central issue for both prosecution and defence.

11. Provider Records, Privacy, and Cross-Border Evidence

Many AI services are hosted outside Pakistan. Relevant evidence may therefore be held by foreign companies in multiple jurisdictions. Traditional mutual legal assistance can be slow relative to the volatility of digital data. Preservation requests, lawful production mechanisms, and international cooperation are increasingly important.

The Budapest Convention on Cybercrime provides a widely used framework for substantive cybercrime offences, procedural powers, and international cooperation, and its Second Additional Protocol expands mechanisms concerning electronic evidence. Pakistan is not presently listed as a Party or observer country with accession status on the Council of Europe's current Convention status materials, although Council of Europe profiles recognize PECA as addressing many Convention offences (Council of Europe, 2026). Pakistan can still use the Convention as a comparative model and consider the benefits and obligations of deeper cooperation.

Cross-border evidence powers must be balanced with privacy, due process, and legality. AI-service records can reveal not only criminal prompts but extensive lawful research, personal data, professional communications, or privileged material. Production requests should therefore be targeted and linked to the investigation rather than functioning as broad fishing expeditions.

Data minimization also protects evidential quality. Collecting enormous volumes of irrelevant model history can obscure the interactions that matter. Investigators should define time periods, accounts, sessions, and offence-related terms carefully, while preserving enough context to prevent misleading excerpts.

12. Developers, Model Providers, and the Limits of Secondary Liability

The most controversial attribution question concerns developers and providers whose systems are misused by others. Criminal liability should not arise merely because a model is capable of harmful use. General-purpose software, encryption, programming languages, cloud hosting, and communication platforms are all dual-use technologies. Imposing criminal responsibility for foreseeable misuse without a sufficiently culpable contribution would chill legitimate innovation and distort mens rea.

Liability becomes more plausible where a provider intentionally designs, markets, customizes, or supplies a system for criminal use, knowing the specific unlawful purpose and materially facilitating it. Existing doctrines of abetment, conspiracy, aiding, or participation may then be relevant depending on statutory requirements and evidence. The key is purposeful or knowing assistance, not mere availability.

A provider's failure to prevent all misuse should normally be addressed through regulation, contractual controls, safety standards, or civil liability unless a criminal statute specifically creates a duty and mental threshold. Criminal law is a blunt mechanism for governance of systemic risk.

The same principle applies to open-source developers. Publication of code with legitimate uses should not itself establish participation in later offences. Evidence of direct coordination with criminals, tailored modifications for an attack, profit-sharing, or deliberate concealment could change the analysis. Context, not technology label, should determine responsibility.

13. AI-Enabled Fraud, Deepfakes, and Social Engineering

AI's most immediate criminal impact is likely to occur through scale and persuasion rather than autonomous hacking. Europol has identified the capacity of large language models to facilitate phishing, social engineering, disinformation-related abuse, and criminal knowledge acquisition (Europol, 2023). UNODC has similarly documented how generative AI can increase the scale, linguistic reach, and credibility of social-engineering operations (United Nations Office on Drugs and Crime, 2025). FATF's 2026 work on cyber-enabled fraud highlights AI-enabled deepfakes as part of evolving fraud methods (Financial Action Task Force, 2026).

For Pakistani criminal law, these uses often fit familiar categories. A synthetic voice used to impersonate a bank official can support fraud if the accused acts dishonestly to obtain property or data. AI-generated identity documents or images may engage forgery or identity-related offences depending on the facts. Automated multilingual messages do not alter the mental element when the human operator intentionally orchestrates the scheme.

The challenge is proof at scale. A fraud operation may involve one person acquiring victim data, another configuring a model, another controlling payment accounts, and another laundering proceeds. Prosecutors should avoid attributing every act to every participant without evidence of common purpose or legally recognized participation. The four-layer model can identify each person's identity, control, mens rea, and contribution.

Victim deception evidence should also be preserved carefully. Screenshots alone may omit headers, account identifiers, or context. Original message exports, platform records, transaction trails, and device data can establish provenance. As synthetic media becomes easier to generate, provenance will often be more probative than visual or auditory plausibility.

14. AI-Generated Malware and Automated Exploitation

AI-generated code raises a common misconception: that machine authorship weakens responsibility for deployment. If a user intentionally requests malicious code, tests it, modifies it, and deploys it against unauthorized targets, authorship is legally secondary. The human's conduct and mental state remain central.

More difficult issues arise when an agent autonomously selects vulnerabilities or payloads. Here, courts should examine the scope of the instruction. An operator who authorizes an agent to penetrate systems without permission has adopted the unlawful objective even if the agent chooses technical means. An operator who authorizes testing only within a defined laboratory but encounters an unexpected escape presents a different case.

Attempt liability may become important where AI assistance is used to prepare or launch an attack that fails. Courts should distinguish remote preparation from conduct sufficiently proximate to the offence according to applicable law. Generating code or asking a hypothetical question may be lawful; deploying exploit traffic against a target may cross the line.

Security research requires particular caution. Penetration testing, malware analysis, vulnerability research, and red-team exercises often use techniques resembling criminal attacks. Authorization is therefore a critical fact. Written scope, contracts, bug-bounty terms, lab configuration, and communications can demonstrate legitimate purpose. Criminal investigation should not treat possession of offensive tools as equivalent to unauthorized use.

15. Corporate and Organizational Responsibility

AI-assisted cybercrime may occur within companies, political consultancies, financial networks, or other organizations. Attribution then extends beyond identifying the employee who pressed a button. The law must ask whether managers directed, authorized, knowingly tolerated, or benefited from the conduct and whether corporate liability is available under the applicable statute.

Organizational evidence can include internal approvals, budgets, performance targets, compliance warnings, audit logs, emails, and instructions concerning model deployment. A senior official who deliberately orders an AI-driven fraud campaign cannot avoid responsibility by delegating technical implementation. Conversely, executive status alone should not create criminal liability for a rogue employee's conduct.

Compliance systems may have evidential significance. Genuine restrictions, access controls, monitoring, and documented responses to misuse can support the inference that unlawful conduct was unauthorized. Sham controls combined with incentives for unlawful results may support the opposite inference. Criminal courts should focus on actual knowledge and contribution rather than treating compliance documents as conclusive.

16. International and Comparative Developments

International institutions increasingly treat AI as both a cybercrime enabler and an investigative challenge. Europol's 2023 report on large language models warned that criminal actors could use them for phishing, impersonation, and technical assistance. UNODC's 2025 work describes AI-driven social engineering and automation as amplifiers of criminal scale. Europol's 2026 assessment of threatening criminal networks identifies AI among technologies used to scale operations and reduce risk. These materials support the proposition that AI changes operational capability even when the underlying legal offences remain familiar (Europol, 2023, 2026; United Nations Office on Drugs and Crime, 2025).

Recent academic work also resists simplistic answers. Nerantzi and Sartor (2024) examine "hard AI crime" scenarios in which no human actor may satisfy existing criteria for responsibility. Other scholarship on AI-powered cyber-aggression emphasizes the difficulty of connecting autonomous conduct to intent and knowledge and debates whether recklessness, *dolus eventualis*, or other models are appropriate. These debates are useful to Pakistan because they expose the danger of forcing every harmful result into intentional liability.

The Council of Europe has established a working group on artificial intelligence within the Cybercrime Convention Committee and has emphasized preservation, chain of custody, responsible investigative use of AI, and international cooperation. Pakistan should monitor these developments even without current Convention Party status. Cybercrime evidence is inherently transnational, and technical standards increasingly influence practical access to proof.

17. A Reform Agenda for Pakistan

Pakistan does not need to criminalize "AI" as an abstract object. Reform should instead clarify how existing offences apply when AI mediates conduct. An interpretive amendment or official prosecutorial guidance could state that an offence may be committed through automated or AI-enabled means while preserving the mental element required by the offence. This would confirm technology neutrality without lowering culpability thresholds.

Second, investigators and prosecutors need specialized protocols for AI evidence. Protocols should cover prompt and output preservation, model and version identification, system instructions, agent tool calls, API logs, provider requests, account authentication, cloud evidence, cryptographic verification, and chain of custody. Forensic reports should distinguish observed facts from inferential conclusions.

Third, Pakistan should develop judicial guidance on digital attribution. Courts should require a chain from infrastructure to human control and from control to *mens rea*. An IP address or account registration should rarely be treated as sufficient by itself where plausible alternative users exist.

Fourth, legislation should avoid broad developer liability based solely on capability or foreseeability. If policymakers wish to impose duties on providers of high-risk AI services, those duties should be

stated prospectively with clear standards and proportionate regulatory sanctions. Criminal liability should attach only where the statute and evidence establish the required fault.

Fifth, international cooperation should be strengthened. Pakistan should evaluate accession pathways and compatibility with the Budapest Convention and its electronic-evidence mechanisms, alongside other bilateral and multilateral arrangements. Rapid preservation and lawful access to provider records will often determine whether AI-assisted cases can be proved.

Sixth, safeguards must accompany law-enforcement use of AI. Investigative models used for triage, facial recognition, anomaly detection, or intelligence analysis should not become opaque substitutes for admissible proof. Human review, validation, disclosure where required, and the ability to challenge methodology are necessary to protect fair trial.

18. Critical Synthesis

AI exposes a basic truth about cybercrime law: technical causation and criminal responsibility are different questions. A model may generate code that causes intrusion, but the court must still determine who deployed it, under what authority, with what knowledge, and toward what purpose. The more autonomous the system, the more carefully those questions must be separated.

Pakistan's existing law is capable of addressing many AI-assisted offences because PECA is concerned with unauthorized access, interference, fraud, identity, and related conduct rather than the brand of tool used. The principal vulnerability lies in attribution. Investigators may be tempted to infer guilt from account ownership, device possession, or suspicious model output. Courts should demand a fuller evidential chain.

A human-centred approach also prevents overcriminalization. Developers should not become guarantors against all misuse, and users should not face intentional-offence liability for genuinely unpredictable outputs. The criminal law's legitimacy depends on matching punishment to culpability. Regulatory law can address broader safety failures where criminal *mens rea* is absent.

At the same time, automation should not provide a loophole. A person who deliberately delegates a criminal objective to an AI agent remains responsible for using that agent as a means, even if the system selects operational details. The decisive issue is not manual control over each step but adoption of the unlawful objective and material contribution to its execution.

A further issue is the danger of evidential circularity when investigators use AI to analyze AI-assisted crime. A law-enforcement model may rank suspects, cluster accounts, translate chats, detect malware similarity, or flag deepfake indicators. These outputs can be useful investigative leads, but they should not become self-validating proof. The prosecution should disclose the human-verifiable evidence that independently supports the inference and, where an AI-generated analytical result is relied upon materially, the methodology, limitations, and validation process should be capable of challenge. Otherwise, a defendant may face an opaque accusation produced by a system whose error rate, training data, or assumptions are unknown. Article 10A of the Constitution makes this concern more than technical. Fair trial requires a meaningful opportunity to test the evidence used to establish criminal responsibility.

This concern is especially acute in attribution. Suppose an investigative model concludes that a suspicious cluster of messages resembles the linguistic style of the accused or that malware resembles code found on the accused's laptop. Such similarity may justify further investigation, but stylistic or code resemblance is probabilistic and can be affected by common templates, copied repositories, AI generation, or deliberate imitation. Courts should insist on corroboration through account access, device artefacts, financial records, communications, or other evidence linking the accused to control of the operation. The availability of powerful analytic tools increases the need for evidential discipline because technical sophistication can create an undeserved aura of certainty.

Pakistan should also distinguish preservation duties from content-monitoring duties. Requiring a provider to preserve identified evidence after lawful notice is conceptually different from imposing a general obligation to inspect all user prompts for potential crime. Broad monitoring may create privacy, expression, cybersecurity, and feasibility concerns and may encourage providers to over-report ambiguous activity. Targeted preservation, lawful production, and risk-based platform safeguards are more compatible with due process than indiscriminate surveillance.

Finally, reform must anticipate evidential disappearance. AI services can change models, retention periods, interface features, and logging practices rapidly. A prompt-output pair may lose meaning if investigators cannot identify the model version, system instructions, tool permissions, or external data available at the time. Preservation protocols should therefore capture contextual metadata early. Investigators should record not only what output was visible but the environment in which it was generated. This contextual record is essential when the defence argues that harmful output was unexpected, when the prosecution alleges deliberate prompt engineering, or when expert witnesses attempt to reproduce the system's behaviour. Reproducibility may never be perfect because probabilistic models can generate different outputs, but careful preservation can narrow uncertainty and prevent courts from treating a later reconstruction as though it were the original event.

The same discipline should apply to defence access. Where feasible and consistent with security, privacy, and privilege claims, the defence should receive sufficient information to test whether logs are complete, whether timestamps are reliable, whether an account could have been compromised, and whether model behaviour supports alternative explanations. Equality of arms does not require disclosure of irrelevant proprietary material, but trade secrecy should not become a blanket answer to a criminal accusation. Courts can use protective orders, expert access, or limited disclosure where necessary. The objective is procedural fairness: a conviction should rest on evidence that can be meaningfully examined rather than on inaccessible technical authority.

19. Conclusion

Artificial intelligence complicates cybercrime by widening the distance between human instruction and technical execution, but it does not require abandonment of human-centred criminal responsibility. Pakistan should continue to ask the conventional questions of *actus reus*, *mens rea*, participation, causation, and proof while adapting those questions to AI-mediated facts.

The appropriate doctrinal reconstruction has four layers. The prosecution should prove identity, control, the offence-specific mental state, and material contribution. Prompts, logs, model outputs, account data, device forensics, financial records, and provider evidence can support these elements, but no single artefact should automatically establish guilt. Electronic records must be authenticated and preserved through reliable chain of custody.

AI should not presently be treated as an independent criminal subject. Technical autonomy is not equivalent to legal or moral agency, and fictional machine *mens rea* would obscure the responsibility of real persons. Nor should criminal law impose strict responsibility on developers or operators merely because harmful output was possible. Where intention or knowledge cannot be proved, regulatory or civil mechanisms may be more appropriate.

The strongest reform agenda is therefore modest in form but demanding in implementation: preserve technology-neutral offences, clarify that AI can be an instrument of commission, build specialist forensic capacity, develop structured attribution standards, strengthen cross-border electronic-evidence cooperation, and protect fair-trial rights when investigative AI is used by the state. Such an approach can address the operational reality of AI-assisted cybercrime without sacrificing the culpability principle that gives criminal punishment its legitimacy.

References

- Abbott, R., & Sarch, A. F. (2019). Punishing artificial intelligence: Legal fiction or science fiction. *UC Davis Law Review*, 53, 323–384.
- Council of Europe. (2001). Convention on Cybercrime (ETS No. 185).
- Council of Europe. (2022). Second Additional Protocol to the Convention on Cybercrime on enhanced co-operation and disclosure of electronic evidence (CETS No. 224).
- Council of Europe. (2025). Working Group on Artificial Intelligence: Terms of reference on cybercrime, electronic evidence and artificial intelligence. Cybercrime Convention Committee.
- Council of Europe. (2026). Pakistan: Octopus Cybercrime Community—Country profile. Council of Europe Cybercrime Programme Office.
- Electronic Transactions Ordinance, 2002 (Ordinance LI of 2002) (Pakistan).
- European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
- Europol. (2023). ChatGPT: The impact of large language models on law enforcement. Publications Office of the European Union.
- Europol. (2026). Decoding the EU's most threatening criminal networks—Issue 2: The blueprint of criminal opportunism. Publications Office of the European Union.
- Financial Action Task Force. (2026). Cyber-enabled fraud: Digitalisation and money laundering, terrorist financing and proliferation financing risks. FATF.
- Hallevy, G. (2010). The criminal liability of artificial intelligence entities—From science fiction to legal social control. *Akron Intellectual Property Journal*, 4(2), 171–201.
- Hallevy, G. (2013). When robots kill: Artificial intelligence under criminal law. Northeastern University Press.
- King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics*, 26, 89–120. <https://doi.org/10.1007/s11948-018-00081-0>
- Lagioia, F., & Sartor, G. (2020). AI systems under criminal law: A legal analysis and a regulatory perspective. *Philosophy & Technology*, 33, 433–465. <https://doi.org/10.1007/s13347-019-00362-x>
- Mangi, D. B., Butro, I., & Shaikh, A. (2025). The impact of artificial intelligence on data protection laws in Pakistan: A critical analysis. *ACADEMIA International Journal for Social Sciences*, 4(2), 1653–1663. <https://doi.org/10.63056/ACAD.004.02.0279>
- Ministry of Information Technology and Telecommunication, Government of Pakistan. (2025). National Artificial Intelligence Policy 2025.

- National Forensics Agency Act, 2024 (Act XXXVII of 2024) (Pakistan).
- National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). U.S. Department of Commerce.
- Nerantzi, E., & Sartor, G. (2024). 'Hard AI crime': The deterrence turn. *Oxford Journal of Legal Studies*, 44(3), 673–701. <https://doi.org/10.1093/ojls/gqae018>
- Pakistan Digital Authority. (2026). Islamabad AI Declaration on sovereign, responsible and capability-driven artificial intelligence. Government of Pakistan.
- Pakistan Penal Code, 1860 (Act XLV of 1860) (Pakistan).
- Prevention of Electronic Crimes Act, 2016 (Act XL of 2016), as amended by the Prevention of Electronic Crimes (Amendment) Act, 2025 (Act II of 2025) (Pakistan).
- Qanun-e-Shahadat Order, 1984 (President's Order No. 10 of 1984) (Pakistan).
- Sattar, S. (2025). The next frontier of cybercrime law for artificial intelligence and criminal liability. *International Journal of Law and Policy*. <https://doi.org/10.59022/ijlp.355>
- Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- United Nations Interregional Crime and Justice Research Institute, Europol, & Trend Micro. (2020). Malicious uses and abuses of artificial intelligence. UNICRI.
- United Nations Office on Drugs and Crime. (2024). Transnational organized crime and the convergence of cyber-enabled fraud, underground banking and technological innovation in Southeast Asia: A shifting threat landscape. United Nations.
- United Nations Office on Drugs and Crime. (2025). Emerging threats: The intersection of criminal and technological innovation in the use of automation and artificial intelligence. United Nations.
- United Nations. (2024). United Nations Convention against Cybercrime: Strengthening international cooperation for combating certain crimes committed by means of information and communications technology systems and for the sharing of evidence in electronic form of serious crimes.
- Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2e2025). National Institute of Standards and Technology.