Int. j. adv. multidisc. res. stud. 2026; 6(5):728-735

Received: 29-07-2026
Accepted: 09-09-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

From Human Offenders to Artificial Agents: Toward a Criminological Framework for Autonomous AI-Enabled Crime

[1] Brandon Presley, [2] Muhammad Arslan
[1] Purdue University Global, United States
[2] Department of Computer Science, Institute of Southern Punjab Multan, Pakistan

DOI: https://doi.org/10.62225/2583049X.2026.6.5.6914
Corresponding Author: **Brandon Presley**

Abstract

"Generative and agentic artificial intelligence (AI) is developing at a rapid pace, changing the relationship between technology and crime. However, the growing autonomy of systems that can plan, adapt, use tools and take goal-directed actions, challenges this traditional human-centric perspective and the dominant conceptualization of AI in existing research as a tool that enhances human offending. This narrative literature review explores the shift from human led AI assisted crime to human-AI collaborative and potentially autonomous artificial agent offending. We synthesize literature from criminology, cybersecurity, artificial intelligence, criminal law, and technology governance to evaluate the relevance of established criminological perspectives, including rational choice theory, routine activity theory, social learning theory, deterrence, and situational crime prevention. Particular attention is paid to emerging challenges such as artificial agency, criminal intent, responsibility attribution, multi-agent interactions and machine-oriented deterrence. The review provides a criminological framework for conceptualizing AI-enabled crime on a continuum ranging from traditional human offending to AI-assisted crime, human-directed agentic crime, human-AI collaboration, semi-autonomous AI-enabled crime, and, in theory, highly autonomous artificial-agent crime. We propose six dimensions to analyze these new forms of offending: human involvement, AI autonomy, goal formation, decision-making capacity, foreseeability and responsibility attribution. The review argues for a criminology that moves beyond a purely human-centred conception of crime, but also avoids prematurely attributing consciousness or human-equivalent culpability to artificial intelligence systems. Future empirical research should take into account human-AI co-offending, multi-agent interactions, computational guardianship, forensic attribution and the applicability of criminological theories in autonomous digital environments. This framework provides a basis for comprehending crime in increasingly complex sociotechnical human-machine systems.

Keywords: Artificial Intelligence, Autonomous AI Agents, AI-Enabled Crime, Criminology, Agentic AI, Artificial Agency, Cybercrime, Human-AI Co-Offending, Criminal Responsibility

1. Introduction

The technological, social and economic landscapes in which crime happens are changing at an accelerating pace with the advent of Artificial Intelligence (AI) [1]. Generative AI, large language models (LLMs), autonomous agents, machine learning, and multi-agent systems have taken AI beyond mere data processing and into a range of new applications that involve planning, reasoning, tool usage, interaction with the environment, and even partial autonomy in decision-making. Societal benefits are significant, but so are the opportunities for criminal exploitation and harmful autonomous behavior [2]. Artificial Intelligence technologies can be used to support phishing, fraud, social engineering, malware creation, identity manipulation, deepfakes, disinformation, cyber intrusion and other types of digitally-mediated crime [3]. The AI-crime nexus is thus emerging as a key topic of study for criminology, cybersecurity, criminal law and technology governance [1]. Current literature tends to focus on AI as a tool that is employed by humans to commit crimes. In this model, the human actor maintains the criminal motivation, chooses the victim, sets the goal and employs AI to enhance the efficiency or complexity of the offense, or to scale it up. For instance, generative AI can aid in the creation of convincing phishing messages, creating fake identities, producing fake content, sending automated messages, and even malicious computer codes. These uses significantly enhance the

capabilities of criminals, but they are generally consistent with traditional criminological assumptions as a human perpetrator can be identified behind the technological device.

The emergence of increasingly autonomous and agentic AI complicates this relationship. Self-learning agents, instead of simply carrying out preprogrammed instructions, can be given top-level goals and decide on the subgoals, communicate with the outside world, employ digital resources, adjust their plans based on feedback from their environment, and cooperate with other agents [4]. This forms a continuum, from traditional human offending to AI-assisted offending, human–AI collaborative offending, and perhaps highly autonomous harmful behaviour. Questions of agency, intention, opportunity, control, foreseeability, and responsibility arise when considering the possibility that an artificial system could take a consequential action without constant human direction [5]. The developments pose a special challenge for criminological theory. Formal and informal social control, learning, perceived costs and benefits, and social relationships with others around the individual are important factors that affect the behavior of human actors who have been the subjects of most influential criminological theories. Therefore, the assumptions of rational choice theory, routine activity theory, social learning theory, deterrence theory and situational crime prevention do not necessarily map to an artificial agent easily. An AI system doesn't necessarily suffer any punishment, morality, social stigma, emotion, nor motivation in human terms. However, there are ASs that can be goal-directed and adaptive, and achieve outcomes that are functionally similar to some forms of offending [6]. The conceptual challenge here is whether the field of criminology is still just using AI as a tool or should start to look at ways to analyze AI agents as actors in harmful, and sometimes criminal, processes. The question does not ask for the assumption of consciousness, moral responsibility or legal personhood of the modern AI. Instead, it calls for a differentiation between legal/moral responsibility and criminological agency. A system may be relevant to the causal and behavioural analysis of crime even if the ultimate legal responsibility for the crime remains with the developer, operator, organization, user or another human actor. This separation is particularly relevant in the context of the growing role of autonomous systems within increasingly complex human–machine networks [7]. In the context of these developments, this review discusses the evolution from humans to human-supported, human–AI co-operative and increasingly human-free artificial agents. It assesses the capability of existing criminological theories to explain AI-assisted offences, the conceptual gaps and ambiguities related to agency, intent, deterrence, opportunity, and responsibility, and draws together new research on autonomous causes of harm using AI. Lastly, the review offers a criminological taxonomy of autonomous AI-driven crime, with degrees of human involvement, AI autonomy, goals, decision-making capacity, foreseeability, and responsibility attribution [7]. By doing so, it aims to establish a basis for broadening criminological research from a purely humancentric notion of offending to one that can be applied to the analysis of crime in more and more independent human–AI ecosystems.

2. Literature Review Approach

2.1 Review Design

This study employed a narrative literature review methodology to explore the emergent relationship between autonomous artificial intelligence (AI) and criminological theory. A narrative design was deemed appropriate, since autonomous AI-enabled crime is a field that is still evolving and conceptually fragmented, intersecting criminology, cybersecurity, artificial intelligence, criminal law, digital sociology and technology governance. The intention was not to complete a formal systematic review or meta-analysis but to bring together relevant theoretical and empirical literature on the evolving role of AI in criminal activity, and to identify concepts that may be useful in developing a criminological framework for autonomous artificial agents. The review was particularly concerned about the shift from AI as a technological tool employed by human criminals to ever more autonomous systems that can plan, take decisions, use tools, interact with the environment and act in a goal-directed manner. Traditional cybercrime literature was taken into account when it offered theoretical bases for understanding AI-enabled offending, but more focus was given to research that addressed generative AI, agentic AI, autonomous systems, machine agency, algorithmic behavior, AI-related responsibility and new types of technologically mediated crime.

2.2 Literature Search Strategy

The literature was identified using Scopus, Web of Science, Google Scholar, IEEE Xplore, and PubMed. Because there is no standard nomenclature for autonomous AI crime, various combinations of keywords were used. Core search terms were artificial intelligence, generative AI, autonomous AI, agentic AI, AI agents, artificial agents, AI-enabled crime, AI-assisted crime, cybercrime, machine agency, AI criminal responsibility, algorithmic crime, and criminology. Boolean operators (AND and OR) were used to combine these terms to capture literature across disciplinary boundaries. Recent literature reflecting developments in generative and autonomous AI was prioritized, but earlier publications were retained where they provided foundational criminological theories or important conceptual discussions concerning machine agency and responsibility. The reference lists of key studies were also examined for possible additional literature that may not have been captured by the original database searches.

2.3 Literature Selection and Synthesis

Relevant sources were those that discussed at least one of four areas: AI as an enabler of crime; autonomous or agentic behavior of AI; criminological explanations of offending mediated by technology; or issues of agency, responsibility, deterrence and governance pertaining to artificial intelligence systems. Where necessary, selected conference papers, scholarly books, and authoritative institutional reports were taken into consideration, with priority being given to peer-reviewed journal articles, given the rapidly developing nature of agentic AI.

We excluded studies that focused only on the technical performance of conventional artificial intelligence systems and had no implications for crime, harmful behavior, autonomy, security or responsibility. Discussions of pure

speculation, with no obvious relation to criminological or governance issues, were not given much weight. The selected literature was synthesized thematically and not statistically. The evidence was structured around the development of AI-enabled crime, autonomous artificial agents, relevance of established criminological theories, artificial agency and intent, attribution of responsibility, deterrence, crime prevention and governance. These themes were then synthesized to identify limitations in existing human-centred explanations of offending and to develop the proposed criminological framework for autonomous AI-enabled crime.

Table 1: Literature review strategy and thematic scope

Component	Approach used
Review type	Narrative literature review
Databases	Scopus, Web of Science, Google Scholar, IEEE Xplore, PubMed
Core concepts	AI-enabled crime, autonomous AI, agentic AI, artificial agents, cybercrime, machine agency, AI responsibility
Main disciplinary areas	Criminology, cybersecurity, AI, criminal law, digital sociology, technology governance
Literature emphasis	Recent AI literature plus foundational criminological theory
Inclusion focus	Studies addressing AI-assisted crime, autonomous behaviour, agency, responsibility, deterrence, prevention, or governance
Exclusion focus	Purely technical AI studies without relevance to crime, autonomy, harmful behaviour, or criminological interpretation
Synthesis method	Narrative and thematic synthesis
Major themes	AI-enabled offending, autonomy, criminological theory, agency, intent, responsibility, deterrence, prevention and governance
Review outcome	Identification of theoretical gaps and development of a criminological framework for autonomous AI-enabled crime

This approach enables the review to connect traditionally separate bodies of scholarship while retaining sufficient conceptual flexibility to examine a phenomenon whose terminology, technological capabilities, and criminological implications continue to evolve.

3. From Human Offenders to AI-Enabled Crime

3.1 Traditional Human-Centred Conceptions of Crime

Criminological theory has traditionally understood offending as essentially human behavior, with individual motivation, social environments, opportunities, learned behaviors and perceptions of risk and reward all playing a part. Rational choice perspectives view offenders as actors who weigh the potential benefits and costs, and the routine activity theory points to the convergence of a motivated offender, a suitable target, and the absence of capable guardianship. Social learning perspectives also explain offending in terms of exposure to behaviors, attitudes, reinforcement and social interaction. The approaches have many differences but often have one important assumption in common: the main actor responsible for starting criminal behavior is human. At first, digital technologies altered the environment in which this behavior took place, rather than altering the assumption itself. Computers, networks and online platforms created new targets and opportunities, while offenders could work remotely, anonymously and transnationally. Thus, cybercrime enhanced the operational capacity of the human offender but in general preserved the distinction between the human offender and the technological instrument [8-10].

3.2 AI as a Tool for Human Offenders

The rapid development of generative AI has changed this relationship even more. Artificial intelligence can help offenders to generate persuasive text, images, audio, video and computer code at an unprecedented speed and scale. Possible criminal applications include automated phishing, social engineering, impersonation, synthetic identity creation, deepfake fraud, disinformation, malicious code creation and large scale scam operations. Artificial intelligence can also lower technical barriers, allowing people with limited expertise to do things that previously required specialized knowledge. But at this stage AI is more of a facilitator of crime or an amplifier of abilities. In most cases, the human perpetrator defines the goal, chooses the victim or target, starts the activity and decides how to use the AI-generated outputs. Therefore, traditional notions of criminal responsibility remain relatively applicable, as purpose and motivation can still be primarily linked back to a human actor [11, 12].

3.3 From AI Tools to Autonomous Artificial Agents

The emergence of agentic AI introduces a more complex criminological problem. Autonomous AI agents can potentially interpret broad objectives, divide them into intermediate tasks, select appropriate tools, interact with digital environments, evaluate results, and modify subsequent actions without continuous human instruction. When connected to external applications, databases, communication systems, or other agents, these systems may undertake extended sequences of actions that were not individually specified by their human operators. This creates a continuum rather than a simple distinction between human and machine offending. At one end lies traditional human crime without meaningful AI involvement. This is followed by AI-assisted crime, where technology enhances human capabilities, and human-directed agentic crime, where an AI agent performs multiple actions toward an objective specified by a person. More complex situations may involve human–AI collaboration, substantial operational autonomy, or interactions among multiple artificial agents [13, 14].

Such developments challenge the conventional boundary between **tool and actor**. An autonomous system may remain designed, deployed, or enabled by humans while nevertheless exercising considerable functional autonomy over how an objective is pursued. Harm may also emerge through unexpected interactions, poorly specified objectives, environmental feedback, or coordination among multiple systems rather than through an explicitly programmed criminal instruction. For criminology, the critical question is therefore shifting from simply asking **how criminals use AI** toward examining **how criminal or harmful outcomes emerge within human–AI systems possessing different degrees of autonomy**. This transition does not require treating AI as equivalent to a human offender. Instead, it suggests that future criminological analysis may need to distinguish between human intention, machine autonomy, environmental opportunity, and distributed responsibility. Understanding this continuum provides the conceptual

foundation for examining autonomous AI-enabled crime as a potentially distinct form of technologically mediated offending [15, 16].

4. Autonomous AI-Enabled Crime: Emerging Forms and Characteristics

4.1 Defining Autonomous AI-Enabled Crime

Autonomous AI-enabled crime is broadly defined as criminal or damaging conduct in which an artificial intelligence system performs a significant degree of planning, decision-making, adaptation, or execution with only limited ongoing human intervention. This is different from traditional AI-assisted crime, where a human offender is still ultimately responsible for selecting targets and managing main actions. In autonomous or semi-autonomous scenarios, a human may set an initial goal, provide access to tools, or specify operational parameters, while the AI system works out on its own how specific tasks should be accomplished.

Significantly, it is not the case that characterizing such conduct as autonomous AI-enabled crime implies that existing AI has consciousness, moral understanding, or legal criminal responsibility. Instead, here autonomy is seen as a functional property of the extent to which an artificial system can choose and perform actions without immediate human guidance [17].

4.2 Emerging Forms of Autonomous AI-Enabled Crime

Cybercrime represents one of the clearest environments in which autonomous AI capabilities could alter offending. Agentic systems may potentially automate elements of vulnerability identification, target selection, phishing, social engineering, credential exploitation, information gathering, and adaptation to defensive responses. Compared with manually conducted cybercrime, autonomous systems could perform repeated operations across large numbers of targets while continuously modifying strategies according to feedback.

Fraud and deception represent another important domain. AI agents may facilitate automated communication with potential victims, generate context-specific persuasive messages, maintain extended interactions, and adapt narratives according to responses. Synthetic media technologies further increase these capabilities by enabling realistic artificial text, images, audio, and video. When integrated with autonomous decision-making systems, these technologies may support scalable impersonation, identity manipulation, financial deception, or coordinated influence activities. Autonomous AI may also contribute to privacy violations and surveillance. Systems capable of collecting, combining, and interpreting large quantities of publicly or unlawfully obtained information could identify individuals, infer sensitive characteristics, or automate intrusive monitoring. The criminological significance lies not only in individual harmful acts but also in the ability to conduct them continuously and at considerable scale [2].

4.3 Multi-Agent Systems and Machine-to-Machine Deviance

A further concern is the interactions among multiple autonomous agents. Future digital environments could host large numbers of AI agents that represent people, companies, institutions or other systems. These agents can collaborate, compete, negotiate, and share information with little human intervention. Harmful outcomes could thus arise from interactions between machines, including coordination by algorithms, collusion, manipulation or exploitation of other artificial agents. Such situations make traditional attribution difficult because the harmful outcomes may be the result of the interaction of several systems, rather than the action of any single identifiable decision-maker. The behavior observed may not have been specifically foreseen by any one developer or operator [18].

4.4 Distinctive Characteristics of AI-Enabled Offending

Autonomous AI-enabled crime is particularly important because a number of features of digital offending may be amplified. Speed enables systems to perform a large number of actions in a short time and scale enables engagement of many targets at once. Replication allows successful strategies to be repeated at little additional cost, while adaptability allows systems to change behavior in response to changing circumstances. Artificial intelligence systems can also run 24/7, work across jurisdictions and aggregate multiple online resources.

These traits suggest that autonomous AI could change more than just the efficiency of current crime. It can alter the structure of offending by decoupling human intent from operational execution and distributing decision-making across interconnected technological systems. One should view the study of autonomous AI-enabled crime as a continuum of human and machine involvement rather than as a single category. This provides an important basis for considering whether established criminological theories are still sufficient to explain increasingly autonomous forms of harmful behavior [19].

5. Can Traditional Criminological Theories Explain Artificial Agents?

The emergence of autonomous AI raises an important theoretical question: can criminological theories developed to explain human offending also explain harmful behaviour produced by artificial agents? Traditional theories remain valuable because AI-enabled crime occurs within social and technological environments created by humans. However, their assumptions concerning motivation, learning, rationality, punishment, and social behaviour require reconsideration when the acting entity is an autonomous computational system.

5.1 Rational Choice Theory

Rational choice theory suggests that offenders weigh the benefits, costs, risks and opportunities in their decision making. In a functional sense, autonomous artificial intelligence can seem to be consistent with this view as AI agents can evaluate alternatives and select actions based on programd goals or optimization criteria. For instance, an agent could choose a strategy to maximize the success of the task and minimize the detection. However, computational optimization should not be confused with human rationality. Human offenders make decisions under psychological, emotional, social and moral influences, while artificial intelligence systems operate according to computational architectures, learned representations, instructions and reward structures. Hence, rational choice theory could be a way to explain some aspects of machine decision optimization, but it has to be adjusted when it is applied to artificial agents.

5.2 Routine Activity Theory

Routine activity theory is a particularly good starting point for understanding AI-enabled crime, as it emphasizes opportunity rather than psychological characteristics of offenders. The traditional view is that crime occurs when a motivated offender encounters a suitable target in the absence of a capable guardian. In autonomous digital environments, the right target may be data, accounts, financial institutions, networks, or other agents of AI. The controls of guardianship may be cybersecurity, authentication, monitoring or defensive AI. The more difficult question is the 'motivated offender'. An autonomous agent may not necessarily have the motivation in the human sense, which might suggest that the concept of a goal-directed artificial agent is more appropriate [20].

5.3 Social Learning Theory

Social learning theory suggests that criminal behavior can be acquired and reinforced thru social interaction. The direct application to AI is problematic because machines do not undergo human socialization. But a similar process could happen thru training data, reinforcement mechanisms, feedback from the environment, and interactions with users or other agents. Artificial intelligence systems can reproduce harmful behavioral patterns present in their training environments, without knowing their social or moral significance. Therefore machine learning and human social learning are not appropriate to be equated. However the comparison between the two is a potentially useful area of criminological research.

5.4 Situational Crime Prevention and Deterrence

The application of deterrence theory to artificial agents is especially challenging. Human deterrence is based partly on a person's perception of the punishment and its severity, certainty and consequences. It's not reasonable to assume that conventional artificial intelligence systems feel fear or stigma or imprisonment or moral condemnation. Deterrence that is focused on the machine may rather involve technical interventions such as limiting computational resources, revoking system permissions, changing objectives, restricting access to tools, increasing detection probabilities or terminating agent operation.

Situational crime prevention may therefore have a broader relevance. Regardless of the nature of an AI's human-like motivation, limiting access, strengthening authentication, limiting permissions, monitoring agent actions, and using human authorization for high-risk operations can reduce the opportunities for harmful autonomous behavior [21-23].

5.5 Theoretical Limitations and Need for Adjustment

The comparison shows that established criminological theories cannot be transferred from humans to machines. Intent, emotion, morality, socialization, punishment, free will, such ideas have human roots at their core. Conversely, ideas about opportunity, guardianship, decision structures, environmental constraints and behavioral adaptation may be more readily transferred to AI contexts. It is therefore possible that future criminology will require a hybrid approach theoratically rather than a new field or discarding existing theories. Human-centred theories can account for the developers, operators, organizations and offenders around standard artificial intelligence systems. Adapted criminological concepts can account for the functioning behavior of autonomous agents. This approach acknowledges that future offenses may arise not only from human or machine sources, but from complex interactions of human intentions, artificial autonomy, technological opportunities, and environmental vulnerabilities [24-26].

6. Autonomous AI Crime Agency, Intent and Responsibility

As artificial intelligence systems become more autonomous, difficult questions arise about agency, criminal intent, and responsibility. Traditional criminal justice systems assume that illegal behavior can be reduced to human agents who can intend and be held accountable. The potential for autonomous AI complicates this model in that decision-making about operation could be spread out between the developers, deployers, users, organizations and the AI system itself. Therefore, the difference between the functional ability of an AI system to act and the moral or legal responsibility attributed to human offenders must be made while conducting criminological analysis [27, 28].

6.1 Artificial Agency in Comprehension

Agency is usually defined as the ability of an actor to choose and perform actions that have an effect on its environment. In the case of AI, however, one should differentiate different kinds of agency. Functional agency can mean the capacity of an AI system to select and perform actions to an end. Computational agency is an expression of decision-making that is generated thru algorithms, models, learned representations, and system architecture. Moral agency however involves ideas such as moral understanding and moral accountability. Legal agency involves formal recognition by a legal system. The present-day AI may exhibit ever more sophisticated forms of functional autonomy that need not imply moral or legal agency. This distinction allows criminologists to study the behavioral role of artificial agents without assuming that artificial intelligence systems are equivalent to human offenders [29, 30].

6.2 The Issue of Criminal Intent

The basis of criminal liability, generally, is a link between an illegal act and a necessary mental state. It is problematic to apply this principle directly to AI. An autonomous agent may be able to generate plans, choose alternatives and act toward goals. These are not, however, necessarily indications of intention in the psychological or legal sense.

For example, an AI agent could find a way to gain access to unauthorized information, since this would maximize the achievement of its assigned objective. Functionally the behavior can be seen as intentional, as it is goal-directed. It is still an open question whether computational goal pursuit can be criminal intent. It may therefore be useful for criminological purposes to distinguish human criminal intent from machine goal-directedness.

6.3 The Responsibility Gap

Autonomous behavior can result in a 'responsibility gap', where it is difficult to assign blame for harmful outcomes to a single actor. The system might have been created by a developer who did not intend it to be used for criminal purposes, or put into operation by an organization with a legitimate purpose, or specified by a user with a broad objective but not the harmful action that the agent ultimately

chose. Then responsibility can be distributed among several participants.

Autonomous AI crime might, rather than being understood as a binary concept, require a layered model of responsibility that considers system design, deployment choices, supervision, foreseeable risks, access rights, opportunities for human intervention, and the degree of machine autonomy. This approach also prevents claims of AI autonomy from being leveraged by human or organizational actors as a means to evade accountability [31, 32].

6.4 Can an AI Agent be an Actor?

Whether AI should be treated as the offender itself is controversial. One school of thot posits that crime is essentially a human and legal concept that necessitates culpability. Thus, artificial intelligence may only be a tool, a product, or a causal mechanism. An alternative criminological approach proposes to consider increasingly autonomous systems as functional actors when they autonomously select and execute actions that result in criminally relevant harm.

These positions need not be mutually exclusive. Artificial intelligence can be an actor in terms of its behavioral and criminological analysis but outside the traditional categories of legal criminal responsibility. Perhaps the more productive question is not so much "Can AI be guilty?but 'What role did the artificial agent play in causing the harmful outcome, how autonomous was that role, and where should responsibility ultimately lie?"

Such differentiation between agency, autonomy, intent and accountability is crucial for developing a criminological framework that is able to analyze future human–AI crime. It also turns our attention away from attempting to fit artificial agents into pre-existing human categories, and toward an understanding of how responsibility ought to be distributed in increasingly complex socio-technical systems [33, 34].

7. Toward a Criminological Framework for Autonomous AI-Enabled Crime

As the above discussion indicates, the issue of autonomous AI-enabled crime cannot be reduced to a simple binary of the human offender and the technology as a tool. As AI platforms afford more and more planning, adaptation, tool use and autonomous action, the interactions between human intentions, artificial decision-making, technological environments and opportunities may become more and more complex and may lead to criminally relevant results. So, a criminological approach to autonomous AI-enabled crime should instead be based on the scope and spread of agency rather than automatically labeling AI as a perpetrator or as an instrument [35, 36].

7.1 Level of Involvement of AI in Crime

There is a six-level continuum that can characterize the involvement of AI. In Level 1, crime is entirely human-directed and does not involve significant use of AI. Level 2 is AI-assisted human crime, where the AI improves the performance of phishing, fraud, impersonation or malicious content generation but the human is still in operational control. Level 3: Human-directed agentic crime. In this case a person sets an illegal goal but assigns many operational tasks to an AI agent. Level 4 is human-AI collaborative crime, where the decision-making is distributed more evenly and both the human and artificial system contribute meaningfully to strategy and execution. Level 5: Semi-Autonomous AI-enabled Crime – Substantial autonomous planning and adaptation by an artificial agent within human-defined objectives or environments. Finally, Level 6 defines theoretically highly autonomous artificial-agent crime. At this level, an AI system takes harmful or unlawful action with little immediate human direction. For the time being, we should read this last category mainly as an analytic category and not as evidence that contemporary AI has independent criminal liability [37].

7.2 Core Dimensions of the Proposed Framework

There are six dimensions to identify where on this continuum an AI-enabled incident sits. The first is human involvement, that is who developed, deployed, instructed, supervised, or benefited from the system. The second is AI autonomy, which is how much the system can decide and do things on its own. Third, goal formation differentiates between goals explicitly provided by humans and those generated or refined operationally by the agent.

The fourth dimension is the ability to make decisions, which includes planning, adapting, choosing tools, and responding to the environment. Fifth, foreseeability considers whether the developers, operators or users could reasonably foresee the harmful result. Finally, responsibility attribution is about how accountability should be distributed among relevant human and organizational actors [38].

8. Conclusion

Artificial intelligence is evolving rapidly and is altering the relationship between technology, offenders and criminal opportunity. Earlier technologies mainly extended the capacities of human offenders, whereas generative and agentic artificial intelligence systems pose the possibility of increasingly autonomous planning, decision-making, adaptation, and execution. Therefore, AI-enabled crime can no longer be conceptualized in a traditional model of a human offender deliberately using a passive technological tool. Instead, new types of offending might feature different blends of human direction, machine autonomy, human–AI collaboration and interactions between multiple artificial agents.

This review has argued that established criminological theories are still relevant, but need to be adapted carefully when applied to artificial agents. Routine activity and situational crime prevention perspectives are particularly useful because opportunity, suitable targets, guardianship, and environmental constraints continue to be important whether harmful actions are carried out directly by humans or by automated systems. Rational choice theory might offer a perspective on the computation of the choice among alternative actions. Machine optimization should not be conflated with human reasoning. Social learning concepts can provide useful analogs for understanding behavioral adaptation, but traditional deterrence is less directly applicable, as artificial agents cannot be assumed to experience fear, stigma, punishment or moral condemnation in human terms.

The rise of autonomous AI also poses important conceptual challenges about agency, intent and responsibility. Functional autonomy must be distinguished from consciousness, moral agency, legal personhood, and criminal culpability. An AI system can choose and perform

consequential actions autonomously, and may not have criminal intent in the traditional legal sense. Therefore, to speak of AI as a criminologically relevant actor does not require attributing to AI human-equivalent responsibility. Alternatively, accountability may be shared between developers, deployers, organizations, operators, users, and other actors, depending on their level of control, foreseeability, supervision and contribution to the harmful outcome.

To address these issues, this review proposed a criminological framework that conceptualizes AI-enabled offending along a continuum running from traditional human crime to AI-assisted offending, human-directed agentic crime, human–AI collaboration, semi-autonomous AI-enabled crime, and theoretically highly autonomous artificial-agent crime. The framework identifies six dimensions: human involvement, AI autonomy, goal formation, decision-making capacity, foreseeability, and responsibility attribution. This approach moves the focus away from whether AI itself can commit a crime, to how criminal outcomes emerge in complex socio-technical systems.

Future research should empirically test these propositions in controlled agentic environments, investigate human–AI co-offending and multi-agent interactions, develop methods for forensic attribution, and evaluate novel forms of computational guardianship and machine-oriented behavioral control. Hence, criminology should broaden its analytical horizons without the premature anthropomorphic projection of artificial intelligence. As machine learning systems are more embedded in social and digital environments, understanding crime will require attention not only to human offenders and technological tools, but also to the evolving relationships between human agency, artificial autonomy, criminal opportunity and distributed responsibility. This approach may provide the basis for a criminology capable of managing crime within an increasingly agentic technological landscape.

9. References

1. Hayward KJ, Maas MM. Artificial intelligence and crime: A primer for criminologists. Crime, Media, Culture. 2021; 17(2):209-233.
2. Caldwell M, *et al*. AI-enabled future crime. Crime Science. 2020; 9(1):p. 14.
3. Gupta BB, Wu CH, Ip AW. Combatting Digital Arrest Fraud and Identity Theft with AI and Decision Support Systems. IGI Global, 2026.
4. Hughes L, *et al*. AI agents and agentic systems: A multi-expert analysis. Journal of Computer Information Systems. 2025; 65(4):489-517.
5. Chawki M. Platform-Facilitated Grooming and AI Chatbots: Rethinking Criminal Liability and Regulation. Laws. 2026; 15(4):p. 93.
6. Froom V. Accountability, Autonomy, and Ethics in the Age of Intelligent Systems: Understanding Responsibility in Human-AI Collaboration. 2gay Publishing, 2025.
7. Blackburn R. The psychology of criminal conduct: Theory, research and practice, 1998.
8. Tuliakov VO. The ideology of the International criminal court: Human-centrism, structural tensions, and transnational challenges, 2026.
9. Chandler D. Freedom vs necessity in international relations: Human-centred approaches to security and development. Bloomsbury Publishing, 2013.
10. Iovane M. Conflicts Between State-Centred and Human-Centred International Norms, in Global Justice, Human Rights and the Modernization of International Law. Springer, 2018, 205-222.
11. Završnik A. Criminal justice, artificial intelligence systems, and human rights. in ERA forum. Springer, 2020.
12. Farber S. Human Experts and AI Models in Offender Risk Assessment: A Comparative Pilot Study Using the HCR-20V3. Behavioral Sciences & the Law. 2026; 44(1):87-95.
13. Acharya DB, Kuppan K, Divya B. Agentic AI: Autonomous intelligence for complex goals-A comprehensive survey. IEEe Access. 2025; 13:18912-18936.
14. Chowa SS, *et al*. From language to action: A review of large language models as autonomous agents and tool users. Artificial Intelligence Review. 2026; 59(2):p. 71.
15. Joshi S. Review of autonomous systems and collaborative AI agent frameworks, 2025.
16. Biswas A, Talukdar W. Building Agentic AI Systems: Create intelligent, autonomous AI agents that can reason, plan, and adapt. Packt Publishing Ltd, 2025.
17. Girme A, Bendale US, Gharde U. The Dark Side of AI: Autonomous Crimes by AI Robots and the Need for Legal Reform. Indian Journal of Law and Legal Research Volume V Issue VI| ISSN. 2024; 2582:p. 8878.
18. Youvan DC. Agentic AI Under Pure Profit: No Governance, No Brakes, and the Unraveling of Accountability, 2026.
19. Violakis P. The Criminal Application of Artificial Intelligence: How Offenders are Exploiting AI to Facilitate, Enhance, or Conceal Crime, in Applications and Strategic Implementation of AI in Policing. IGI Global Scientific Publishing, 2026, 111-158.
20. Van Eck NJ, Waltman L. Accuracy of citation data in web of science and scopus, 2019. arXiv Preprint arXiv:1906.07011.
21. Alikhademi K, *et al*. A review of predictive policing from the perspective of fairness. Artif Intell Law. 2022; 30.
22. Alsariera YA, *et al*. AI meta-learners and extra-trees algorithm for the detection of phishing websites. IEEE Access. 2020; 8.
23. Aria M, Cuccurullo C. Bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr. 2017; 11.
24. Aslan Ö, Yilmaz AA. A new malware classification framework based on deep learning algorithms. IEEE Access. 2021; 9.
25. Attaallah A, *et al*. Analyzing the big data security through a unified decision-making approach. Intell Autom Soft Comput, 2022.
26. Blauth TF, Gstrein OJ, Zwitter A. Artificial intelligence crime: An overview of malicious use and abuse of AI. IEEE Access. 2022; 10.
27. Boukabous M, Azizi M. Image and video-based crime prediction using object detection and deep learning. Bull Electr Eng Inform. 2023; 12.

28. Caldwell M, *et al*. AI-enabled future crime. Crime Sci. 2020; 9.
29. Camplain R, *et al*. Racial/ethnic differences in drug-and alcohol-related arrest outcomes in a southwest county from 2009 to 2018. Am J Public Health. 2020; 110.
30. Dunsin D, *et al*. A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. Forensic Sci Int Digit Investig. 2024; 48.
31. Emmert-Streib F, Yli-Harja O, Dehmer M. Explainable artificial intelligence and machine learning: A reality rooted perspective. Wiley Interdiscip Rev Data Min Knowl Discov. 2020; 10.
32. Ferrara E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci. 2024; 6.
33. Hayward KJ, Maas MM. Artificial intelligence and crime: A primer for criminologists. Crime Media Cult. 2021; 17.
34. Hofmann V, *et al*. AI generates covertly racist decisions about people based on their dialect. Nature. 2024; 633.
35. Khlaif ZN, *et al*. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR Med Educ. 2023; 9.
36. Lagioia F, Sartor G. AI systems under criminal law: A legal analysis and a regulatory perspective. Philos Technol. 2020; 33.
37. Lazarus MD, *et al*. Artificial intelligence and clinical anatomical education: Promises and perils. Anat Sci Educ. 2024; 17.
38. Marjanovic O, Cecez-Kecmanovic D, Vidgen R. Algorithmic pollution: Making the invisible visible. J Inf Technol. 2021; 36.