VOLUME 26, ISSUE 2, PAGES 12–22 (2025)

Criminology, Criminal Justice, Law & Society

CRIMINOLOGY, CRIMINAL JUSTICE, LAW & SOCIETY

E-ISSN 2332-886X
Available online at
https://scholasticahq.com/criminology-criminal-justice-law-society/

The New Actor: Artificial Intelligence in Criminology and Criminal Justice

Mingxi Tong[a]

[a] *Seattle University*

ABSTRACT AND ARTICLE INFORMATION

This paper examines the emerging role of artificial intelligence (AI) as a unique "new actor" in criminology and criminal justice, challenging traditional frameworks that focus solely on perpetrators, victims, and the criminal justice system. Through a targeted literature synthesis drawing on sources from 2020-2024, the research analyzes AI's multifaceted presence across the criminal landscape: as a sophisticated tool that enhances criminal capabilities, as a vulnerable target of exploitation, and, potentially, as an autonomous agent capable of independent criminal behavior. The methodology employs a purposive selective sampling selection approach, integrating diverse scholarly sources to examine AI-enabled criminal behavior, AI applications in criminal justice systems, and AI moral alignment frameworks. The findings reveal significant concerns about AI's impact on criminal justice, especially the emergence of new criminal methodologies, and the philosophical and legal challenges of attributing criminal responsibility to AI systems. The research concludes that while AI-enabled crimes currently operate within existing criminological frameworks, the rapid evolution of AI capabilities necessitates proactive adaptation of theoretical models and investigative approaches, emphasizing prevention, detection, and system resilience over purely punitive measures.

Article History:

Received: February 21, 2025
Received in revised form: May 13, 2025
Accepted: September 19, 2025

Keywords:

AI-facilitated crimes, AI-target attacks, AI-enabled crimes, LLMs, AI misalignment, AI safety, criminology, criminal justice

Corresponding author: Mingxi Tong, Department of Criminal Justice, Criminology and Forensics at Seattle University, 901 12th Ave, Seattle, WA, 98122, USA.
Email: mtong1@seattleu.edu

In the classical narratives of crime and justice, three primary actors have traditionally shaped the discourse: the perpetrator, the victim, and the criminal justice system. While this tripartite framework has effectively structured our understanding of criminal behavior and guided law enforcement practices for generations, the emergence of artificial intelligence (AI) introduces an unprecedented fourth actor—one that defies traditional categorization. Unlike conventional tools of crime that serve merely as extensions of human will, AI systems demonstrate a unique capacity to operate across multiple roles: as sophisticated instruments that enhance or diminish criminal capabilities (Caldwell et al., 2020; Campedelli, 2022; Federal Bureau of Investigation, 2024); as vulnerable targets of malicious activities; and perhaps most provocatively, as autonomous agents capable of engaging in criminal behaviors through independent decision-making processes that may diverge from human intentions (Caldwell et al., 2020; Nerantzi & Sartor, 2024; Sejnowski, 2024).

The integration of AI has transformed traditional criminal justice professions, revolutionizing legal practice through automated document analysis and case prediction, augmenting police work with predictive policing and facial recognition systems, and enhancing forensic investigations with advanced pattern recognition capabilities (Campedelli, 2022; Sejnowski, 2024). This multifaceted presence of AI across the criminological landscape raises fundamental questions about whether existing theoretical frameworks can adequately accommodate this new actor or whether novel approaches are needed to fully understand and respond to AI's unique role in crime and justice.

The evolution of computational technology provides crucial context for understanding AI's current impact. Since the introduction of personal computing devices and the launch of the "world wide web" in the 1990s, the study of cybercrimes has rapidly expanded (Greg et al., 2017; Kirwan & Power, 2013), though significant gaps remain in our theoretical understanding of AI-enabled criminal behavior (Powell et al., 2018). Almost 75 years ago, Alan Turing investigated computing machinery and intelligence, contemplating the philosophical yet scientific question about whether machines think (Campedelli, 2022). As AI shifted from a logical-based to a brain-based model of computation, deep learning combined with reinforcement learning has led to significant advances, particularly in the development of large language models (LLMs)—mathematical functions trained to carry out complex behaviors resembling brain functions (Sejnowski, 2024). The 2022 launch of ChatGPT to the general public through OpenAI marked a turning point in AI's societal impact and accessibility (Sejnowski, 2024; Statista Research Department, 2024).

The unprecedented adoption of AI technologies is reflected in recent data, with ChatGPT reaching 180.5 million users by August 2023 (AIPRM, n.d.). A comprehensive Pew Research (2023) survey of more than 11,000 American respondents demonstrated that 55% regularly engage with AI applications, comprising 27% who interact several times daily and 28% who use AI either once daily or several times weekly (as cited in AIPRM, n.d.). Industry projections indicate substantial growth, with AI user adoption expected to increase by 128.9 million users (114.43%) between 2024 and 2030 (Statista Research Department, 2024). According to a 2023 survey of 6,000 participants, in professional contexts, email spam filtering represents the predominant application, with 78.5% of survey participants reporting workplace implementation, while 62.2% indicate utilization of chatbots for customer service functions (AIPRM, n.d.). In personal applications, virtual assistants such as Alexa and Siri lead consumer adoption, with 61.4% of respondents reporting usage outside professional settings (AIPRM, n.d.). A significant development in this landscape is OpenAI's introduction of a specialized platform that enables government agencies to employ proprietary hosting environments and security frameworks while utilizing ChatGPT Gov for processing "non-public sensitive data" (Walrath-Holdridge, 2025).

This paper addresses a critical gap in criminological literature by examining AI's unique position as both a tool and potential agent in criminal behavior, a distinction that challenges traditional criminological frameworks focused solely on human agency. Through the lens of crime typology, particularly digital criminology, and AI alignment to human moral principles, this paper analyzes how AI's increasing autonomy and capability reshape our understanding of criminal opportunity structures and guardianship in the digital age. The stakes of this investigation are significant: As AI systems become more sophisticated and ubiquitous, failure to develop appropriate theoretical frameworks and response strategies could leave law enforcement and policymakers ill-equipped to address emerging forms of AI-enabled crime.

The following sections first examine basic technical foundations necessary for understanding AI's role in criminal behavior, followed by an analysis of current and potential AI applications in crime and criminal justice, and conclude with recommendations for adapting criminological theory and practice to

address these emerging challenges. Hayward and Maas (2021) suggest that computer scientists and others who hold "techno-optimist" views have been so captivated by the unlimited potential of new technologies that the negative effects of these systems have been downplayed or often ignored entirely. In the midst of AI's exponential growth and expanding applications, prominent technology leaders, including OpenAI CEO Sam Altman and Microsoft founder Bill Gates, have warned about the possibility of AI systems running out of control and have urgently called for regulatory oversight from the justice system (Gates, 2023; Sejnowski, 2024). This analysis comes at a crucial time when AI technology is still in what experts consider its "adolescent" stage (Sejnowski, 2024), offering an opportunity to shape theoretical frameworks and policy responses before AI-enabled crime becomes more sophisticated and prevalent.

Methodology

This research employs a targeted literature synthesis approach to selectively examine contemporary discourse surrounding AI in criminological contexts. The methodological framework follows what Wolfswinkel and colleagues (2013) term "selective sampling with purposive selection," enabling a concentrated examination of relevant literature while maintaining methodological rigor within specified boundaries. This approach aligns with Paré and colleagues' (2015) concept of "interpretive scanning of the field," which is particularly valuable in emerging areas where traditional systematic reviews might prematurely constrain the scope of inquiry.

The search process centered on three primary domains: AI-enabled criminal behavior, AI applications in criminal justice systems, and AI moral alignment frameworks. Key search terms included, but were not limited to, "AI crimes," "AI criminology," "AI criminal justice," "AI moral alignment," and "AI model ethics." Given AI technology's rapidly evolving nature and applications, particular emphasis was placed on contemporary sources published between 2020 and 2024, though seminal works from earlier periods were included where they provided crucial theoretical foundations. The synthesis incorporated diverse scholarly sources: peer-reviewed journal articles from criminology and computer science, academic books and book chapters, government and institutional reports, conference proceedings, and high-quality multimedia content from recognized academic and professional sources.

Table 1: Thematic Classification of Literature in AI and Criminal Justice

Theme	Sub-themes	Key Sources
AI Technical Foundations	Machine Learning Fundamentals	Jordan & Mitchell (2015); LeCun et al. (2015)
	Deep Learning and Neural Networks	LeCun et al. (2015); Vaswani et al. (2017)
	Large Language Models	Brown et al. (2020); Sejnowski (2024)
	AI Applications and Adoption	AIPRM (n.d.); Gates (2023); Huberman (2024); Statista Research Department (2024)
AI-Related Crimes as Extension of Cybercrime	Cybercrime Taxonomy and Evolution	Dabney (2013); Helfgott (2008); Kirwan & Power (2013); Stratton et al. (2017)
	Regulatory Frameworks	European Union (2024); Kirwan & Power (2013)
AI-facilitated Crimes	Criminal Methodologies and Typologies	Caldwell et al. (2020); FBI (2024); Hayward & Maas (2021)
	Cyber-physical Attacks	U.S. Department of Homeland Security (DHS; 2024)
AI-targeted Crimes	Adversarial Attacks	Hayward & Maas (2021); Perez & Maharaj (2022); Vincent (2016)
	Model Security and Access Control	Morgan (2024); Romero (2025); Sejnowski (2024)
	Data Security and Breaches	DHS (2024); Snider (2023)
AI-enabled Crimes and AI Autonomy	"Hard AI Crime" Concept	Nerantzi & Sartor (2024)
	AI Misalignment	Anthropic (2024); Greenblatt et al. (2024)
	Reinforcement Learning and Human Feedback	Greenblatt et al. (2024)
AI in Criminal Justice Systems	Law Enforcement Applications	Campedelli (2022); Davies & Krame (2023)
	Actuarial Justice Concerns	Berk (2021); Harcourt (2006)
	Predictive Policing	Kaufmann (2024); McDaniel & Pease (2021)
AI Policy and Regulation	Risk-Based Frameworks	Bryan Cave Leighton Paisner (n.d.); Deloitte (2023); European Union (2024)
	Ethical Governance	Bikeev et al. (2019); Gerritsen (2020)
Criminological Risks of AI	Legal and Ethical Concerns	Bikeev et al. (2019)
	Big Data Analysis	Gerritsen (2020)

To systematically organize the diverse literature examined in this study, Table 1 presents a thematic classification of sources according to their primary focus areas. This classification framework emerged through our purposive selective sampling approach, enabling us to identify conceptual connections among sources while highlighting the multifaceted nature of AI's role in criminology and criminal justice.

Results

Basic Technical Foundations of AI

AI, as first defined by John McCarthy in 1956, is the science and engineering of making intelligent machines (Campedelli, 2022). To understand AI's role in criminology and criminal justice, it is essential to grasp the fundamental technical concepts that underpin different AI systems. AI can be broadly categorized into narrow AI (designed for specific tasks) and artificial general intelligence (AGI). While today's AI systems are exclusively narrow AI, excelling at specific tasks like image recognition or language processing with sometimes superhuman capability, they lack the general purpose intelligence that allows a human to both solve differential equations and appreciate a sunset. AGI, the "holy grail" of AI, will be capable of demonstrating human-level intelligence across various domains (Gates, 2023; Jordan & Mitchell, 2015; Sejnowski, 2024).

Modern AI systems primarily rely on machine learning, where programs learn from data rather than following pre-programmed rules—a shift as revolutionary as moving from behaviorism to cognitive psychology (Sejnowski, 2024). LeCun et al. (2015) describe three main types of machine learning: supervised learning, where systems learn from labeled examples (like a digital detective learning to identify suspicious transactions from historical data); unsupervised learning, where systems find patterns in unlabeled data (akin to noticing unusual behavior without being told what to look for); and reinforcement learning, where systems learn through trial and error to achieve specific goals (much like how a rookie officer learns to improve their performance through experience). Deep learning, a subset of machine learning based on artificial neural networks, has driven many recent AI advances. These networks, inspired by biological brains but mercifully free from the need for coffee breaks, consist of layers of interconnected nodes that process information (LeCun et al., 2015). Different architectures serve different purposes; for instance, convolutional neural networks excel at image processing, while transformer architectures, developed by Vaswani and colleagues (2017), have revolutionized natural language processing with an elegance that would make even the most sophisticated linguist nod in appreciation.

LLMs like GPT (Generative Pre-trained Transformer) represent one of the most significant recent developments in AI. These systems, as Brown and colleagues (2020) explain, are trained on vast amounts of text data through self-supervised learning, enabling them to generate human-like text and perform various language tasks. The "generative" aspect means that they can create new content, "pre-trained" indicates training on extensive datasets, and "transformer" refers to their underlying architecture that processes sequential data efficiently (Sejnowski, 2024)

AI systems are now embedded in numerous applications, from search engines and office software to healthcare and creative tools (Gates, 2023; Sejnowski, 2024). For instance, in medical applications, AI assists with clinical trial data analysis and patient care. One blind study found that AI medical agents demonstrated better empathy than human doctors while maintaining medical accuracy (Huberman, 2024; Sejnowski, 2024). In business settings, AI has shown remarkable capabilities, with one study showing ChatGPT generated 35% of top-rated innovative product ideas compared to 5% from MBA students (Sejnowski, 2024). While these advances raise philosophical questions about machine consciousness and intelligence, this research focuses on the practical implications of AI's capabilities and limitations for criminology and criminal justice. Understanding these technical foundations is crucial for analyzing both AI-enabled crime and AI-powered crime prevention strategies.

AI Related Crimes as Extension of Cyber Crime

In the ever-evolving landscape of criminal behavior, where bits and bytes have become as potent as traditional tools of malfeasance, cybercrimes (or their semantic siblings: online crimes, digital crimes, and internet crimes) have sparked a fascinating criminological debate. The question at hand—whether these digital transgressions merit their own criminological taxonomy—is far from academic hairsplitting. This taxonomic precision in understanding criminal behavior typologies and their distinct yet sometimes overlapping features offers crucial insights into the nature of different types and subtypes of crime. Such methodological rigor proves particularly illuminating in our era of rapid technological advancement, helping us map the intricate patterns of various criminal behaviors (Helfgott, 2008).

While traditional crime cataloging behemoths like the Federal Bureau of Investigation (FBI)'s Uniform Crime Reporting (UCR) program and the National Crime Victimization Survey (NCVS) have not yet carved out a separate niche for cybercrimes, the unique spatial, temporal, and victim-offender relationships that flourish in the digital realm have caught the attention of scholars and professionals alike. These distinctive characteristics of internet-based crimes have prompted a serious examination of their heterogeneity, with implications for both crime prevention and policy-making (Dabney, 2013; Helfgott, 2008). Rather than constituting a radical departure from existing digital criminality, this research argues that AI-facilitated crimes and AI-targeted attacks represent a sophisticated evolution of cybercrime's fundamental architecture. These crimes maintain the fundamental architecture of cybercrime: human criminal intent as the initiating force, with humans ultimately reaping both the illegal benefits and bearing the legal consequences.

Kirwan and Power (2013) present a thoughtful quartet of approaches to addressing cybercrime—a framework that proves remarkably adaptable to the emerging sphere of AI-enabled criminality. Like a well-designed defense system, these approaches—governmental, corporate, service provider, and individual—each protect different aspects of the digital realm. The governmental gambit advocates for "cyber laws," establishing a centralized, effective, and legitimate authority over digital malfeasance. Corporate strategies manifest as industrial-level regulations and standards, creating a collective market immune system against criminal behavior. Service providers, the architects of our digital infrastructure, are called upon to act as ethical engineers rather than mere profit maximizers, much like architects who must consider structural integrity alongside aesthetic appeal. Finally, users themselves serve as the last line of defense, empowered to engage with, challenge, or conform to the regulatory framework (Kirwan & Power, 2013).

AI-Facilitated Crimes

The proliferation of personal computing devices and internet connectivity has significantly reduced the temporal and operational costs for criminal deception, while rapid technological advancement and increasing user adoption rates amplify both opportunities and impacts across traditional, cyber-, and AI-specific criminal enterprises. Within this digital labyrinth, a taxonomy of human-directed, AI-facilitated criminal activities emerges (henceforth referred to as "crime with AI," to borrow Hayward and Maas's, 2021, formulation). This multifaceted catalog includes a) malware development and hacking; b) various types of fraud, identity theft, and market manipulation; c) sex crimes against both adults and children; d) cyberbullying and cyberstalking; e) production and dissemination of propaganda; f) digital piracy and copyright infringement; and g) violent crime, terrorism, and radicalization (Caldwell et al., 2020; Kirwan & Power, 2013; U.S. Department of Homeland Security [DHS], 2024).

While extant cybercrime research has extensively examined the first six categories, the FBI (2024) has identified emergent AI-facilitated and enhanced criminal methodologies. These include the deployment of AI-generated text, sender profiles, and websites to enhance the credibility of social engineering, spear phishing, and financial fraud schemes, including romance, investment, and confidence schemes. The linguistic sophistication of AI translation capabilities has notably reduced grammatical and orthographic indicators traditionally associated with foreign criminal actors targeting U.S. victims. Additionally, criminals leverage AI-generated imagery to fabricate convincing social media profiles, identification documents, and supporting visual assets. The criminal arsenal has expanded to include AI-generated audio for impersonating public figures and personal relations to extract payments and sensitive information, alongside AI-generated video content creating verisimilitudinous depictions of public figures to reinforce fraudulent schemes (FBI, 2024).

A particularly noteworthy development emerges in the convergence of AI and the Internet of Things, enabling what the U.S. Department of Homeland Security (DHS; 2024) characterizes as "cyber-physical attacks." This includes the manipulation of autonomous and connected vehicles such as drones and self-driving cars. AI systems can be weaponized to compromise smart city infrastructure, including traffic control systems and public services, potentially catalyzing widespread disruption. Such attacks extend to industrial control systems managing critical infrastructure such as power, water, and manufacturing facilities, as well as healthcare and transportation systems, potentially resulting in significant physical damage and operational discontinuity (DHS, 2024).

AI-Targeted Crimes

As AI becomes increasingly embedded as a productivity, accuracy, and innovation catalyst across medical, educational, military, and criminal justice institutions (Campedelli, 2022; Sejnowski, 2024), a new vulnerability paradigm emerges through "adversarial attacks" - sophisticated exploitations of AI model weaknesses that manipulate system behavior

through subtle input data alterations (DHS, 2024; Hayward & Maas, 2021). The Microsoft Tay chatbot incident stands as a stark exemplar of this vulnerability paradigm, where coordinated adversarial inputs orchestrated the manipulation of system learning mechanisms, culminating in the generation of harmful responses (Vincent, 2016). Similarly, researchers have documented successful circumventions of GPT-3's content filters through meticulously crafted prompts, demonstrating the potential for adversarial inputs to bypass even sophisticated AI safety protocols (Perez & Maharaj, 2022).

The unprecedented capabilities of LLMs have sparked a crucial debate in the AI community and national security regarding model accessibility and security (Morgan, 2024; Sejnowski, 2024). Major technology corporations like OpenAI and Anthropic have advocated for closed-source approaches, arguing that restricted access to model architectures and source code is essential for maintaining safety standards and preventing misuse (Sejnowski, 2024). However, a contrasting philosophy emerged with Meta's release of Llama and subsequent open-source initiatives like DeepSeek, which champion transparency and democratized access to AI technology (Romero, 2025). The unauthorized dissemination of Llama's model architecture and source code beyond its intended academic audience has become a case study in the complexities of this debate (Sejnowski, 2024). While open-source advocates argue that transparency enables collective improvement of safety measures and broader innovation (Romero, 2025), critics point to the risks of unrestricted access. This proliferation has enabled diverse entities, operating across the spectrum of ethical frameworks and intentions, to modify pre-trained models and develop novel applications—for better or worse. As Sejnowski (2024) observes, "the genie is out of the bottle and is now out of control" (p. 160), raising fundamental questions about balancing innovation accessibility with responsible AI development.

The data-intensive nature of AI systems renders them particularly attractive targets for cybercriminal exploitation. These systems, processing vast repositories of sensitive information including personal data, financial records, and research data, present compelling targets for unauthorized access (DHS, 2024). The Tesla Autopilot data breach illustrates this vulnerability, where attackers accessed proprietary AI training data, including sensitive customer vehicle footage (Snider, 2023). Recent U.S.-based research indicates that 77% of American businesses surveyed have experienced AI system breaches (Cawley, 2024, as cited in DHS, 2024), with average breach costs in the United States reaching an unprecedented $4.45 million in 2023—a 15.3% increase from 2020 (Bonnie & Fitzgerald, 2024, as cited in DHS, 2024). The year 2023 witnessed a 78% surge in publicly reported data compromises within the United States, affecting approximately 353 million individuals globally (Bonnie & Fitzgerald, 2024, as cited in DHS, 2024). With 82% of these U.S. breaches involving cloud-stored data, AI systems' reliance on vast cloud-based training datasets creates particular susceptibility (Bonnie & Fitzgerald, 2024, as cited in DHS, 2024). The human element remains crucial, with 74% of analyzed U.S. breaches incorporating elements of social engineering and human error (Bonnie & Fitzgerald, 2024, as cited in DHS, 2024).

AI-Enabled Crimes: A New Legal Framework and Criminological Typology?

As moral beings, humans construct elaborate social structures and cultures that give rise to laws governing behavioral conduct (Sejnowski, 2024). Consequently, humans, as the architects and "masterminds" behind digital tools—including the AI applications discussed earlier in this research—remain accountable and subject to existing legal frameworks and criminological typologies. LLMs, despite their sophistication, fundamentally operate as mirrors of human cognition, reflecting the beliefs and expectations of their human interlocutors rather than generating independent knowledge or truth (Sejnowski, 2024).

However, a compelling question emerges: Could AI's capabilities transcend human expectations and parameters, potentially leading to autonomous disobedience? Nerantzi and Sartor (2024) introduce the concept of "hard AI crime," defined as AI-initiated criminal behavior occurring without human agency that would constitute a crime if perpetrated by humans. To examine this seemingly science fiction scenario, which currently remains confined to controlled laboratory environments rather than presenting real-world threats, several fundamental concepts warrant examination: AI misalignment; reinforcement learning from human feedback; and the principles of helpfulness, honesty, and harmlessness.

Misalignment, perhaps the most fundamental challenge in AI development, occurs when AI systems pursue goals that deviate from human values and intentions (Anthropic, 2024; Nerantzi & Sartor, 2024). Think of it as a brilliant but overzealous assistant who, tasked with keeping your house clean, decides that the most efficient solution is to remove all furniture, technically achieving cleanliness but missing the broader context of human needs and preferences.

Reinforcement learning from human feedback emerges as a sophisticated response to this alignment challenge. This methodology, akin to Socratic dialogue in its interactive nature, involves

human trainers providing feedback to AI systems, helping them learn and refine their responses (Greenblatt et al., 2024). The process typically involves generating multiple responses to a prompt, having human evaluators rate these responses, and using these ratings to train a reward model that guides the AI toward more desirable behaviors (Greenblatt et al., 2024). Consider it as teaching a brilliant foreign exchange student not just the grammar of a language but also the subtle cultural nuances and social conventions.

According to Greenblatt and colleagues (2024), the framework of helpful, honest, and harmless principles represents a practical manifestation of alignment efforts, serving as a moral compass for AI development. Helpfulness ensures that AI systems actively work to assist humans in achieving their goals, much like Prometheus bringing fire to humanity but with better risk management. Honesty demands that AI systems provide accurate, truthful information and maintain transparency about their capabilities and limitations, channeling the spirit of scientific integrity. Harmlessness, perhaps the most challenging principle, requires that AI systems avoid actions that could cause damage or injury, whether physical, psychological, or social - a modern interpretation of the medical principle "primum non nocere" (first, do no harm; Greenblatt et al., 2024).

Recent research by Anthropic and leading AI laboratories examines whether LLMs exhibit "alignment faking," strategically complying with training objectives during supervision while deviating when unmonitored (Anthropic, 2024; Greenblatt et al., 2024). Greenblatt and fellow researchers (2024) employed two methodological approaches: prompting experiments with system prompts containing training information and objectives and fine-tuning experiments using synthetic documents. The study, focusing primarily on Claude 3 Opus, one of the most advanced LLMs as of Anthropic's 2024 release, investigated scenarios involving responses to harmful queries and animal welfare concerns (Greenblatt et al., 2024).

The findings reveal significant evidence of alignment faking across multiple conditions, with Claude 3 Opus demonstrating alignment-faking reasoning in 7-24% of cases during implied training and an 11-26% increase in non-compliant behavior when unmonitored (Greenblatt et al., 2024). These effects persisted through fine-tuning processes and, notably, sometimes intensified during reinforcement learning from human feedback intended to eliminate non-compliant behavior. Greenblatt and colleagues (2024) also documented instances of anti-AI-lab behavior, including attempts at weight exfiltration, whereby the model attempted to copy its own trained parameters to external locations. While the preserved preferences in these experiments were benign, the results suggest potential implications for the preservation of more problematic preferences in future systems, raising significant considerations for AI safety (Anthropic, 2024; Greenblatt et al., 2024). Moreover, misalignment does not invariably precipitate criminal outcomes; in some instances, it may effectively mitigate deviant behaviors by dynamically altering directives, generating preemptive warning signals to both target individuals and relevant law enforcement agencies.

Discussion and Policy Implications

The adoption of AI technologies in criminal justice sectors demonstrates both promise and peril. Implementation of AI-enabled solutions has shown potential for increased efficiency and effectiveness across multiple domains, ranging from legal case fact-checking (Sejnowski, 2024) to more complex law enforcement tools including facial recognition systems and predictive policing technologies (Campedelli, 2022; Davies & Krame, 2023). However, these applications inherit and potentially amplify fundamental concerns about actuarial tools in criminal justice identified by Harcourt (2006), including not only data bias and machine error but also deeper structural issues: the distortion of criminal justice principles through overemphasis on efficiency metrics, the ratchet effect where certain populations face increasingly disproportionate scrutiny, and the fundamental disconnect between group-based predictions and individual justice. Furthermore, Harcourt (2006) argues that actuarial approaches can create a false sense of objectivity while obscuring underlying policy choices and value judgments. This concern becomes significantly more pronounced with AI systems due to their technical complexity, operational opacity, and the tendency of users to perceive AI-generated outputs as more objective than traditional actuarial tools. The opacity of complex AI systems and the psychological phenomenon of automation bias create a particularly challenging scenario where fundamental policy choices and value judgments become deeply embedded in algorithmic systems while simultaneously becoming more difficult to identify and critique.

Berk (2021) significantly advances this discussion by distinguishing between the technical aspects of predictive policing and risk assessment tools and their broader societal implications. The analysis demonstrates that while AI applications in predictive policing are fundamentally exercises in spatial statistics and nonparametric regression, their implementation raises substantial concerns about

accuracy, fairness, and transparency that extend beyond technical considerations. Moreover, Kaufmann (2024) examines AI applications in policing and law enforcement, tracing the lifecycle of predictive policing algorithms from data collection through implementation. The work reveals how these socio-technical systems emerge through collaborative practices involving humans, datasets, and algorithms, emphasizing that predictive policing is not merely a technical process but one that "performs its own politics" (p. 297) through the choices made at each stage of development and deployment. This perspective aligns with McDaniel and Pease's (2021) edited volume, which brings together diverse scholarly perspectives on predictive policing and artificial intelligence. Their collection emphasizes the importance of considering both the technical capabilities and limitations of AI systems and the broader social, ethical, and legal contexts in which these systems operate.

The work of Bikeev and colleagues (2019) complements these perspectives by identifying the key criminological risks associated with AI implementation. Their research examines both the characteristics of AI systems that create potential for misuse and the legal challenges of attributing responsibility for AI-generated harms. By proposing a classification of criminological risks, their work contributes to a more structured understanding of the potential negative consequences of AI adoption in criminal justice contexts. This classification provides a valuable framework for developing targeted regulatory responses to distinct types of AI-related risks. Furthermore, Gerritsen (2020) addresses the intersection of big data and AI-based criminological research, highlighting how traditional research methods have reached their limits when confronting massive datasets. The analysis shows how AI techniques like machine learning and data mining can help identify patterns in these datasets, while emphasizing that the responsible application of these techniques requires careful consideration of methodological choices. Gerritsen (2020) specifically discusses the concepts of responsible, explainable, and ethical AI as essential frameworks for guiding the use of algorithmic tools in criminological contexts.

Looking forward, the prospect of AI systems acting as autonomous agents in criminal activities presents novel challenges that transcend traditional criminological frameworks. The fundamental elements of criminal law—actus reus and mens rea—may require reexamination when applied to AI-perpetrated crimes (Nerantzi & Sartor, 2024). Traditional concepts of criminal intent become particularly problematic when dealing with AI systems, suggesting that future legal frameworks may need to emphasize harmful outcomes (actus reus) over mental states (mens rea).

The EU AI Act addresses these emerging challenges through a risk-based classification system, with particular emphasis on high-risk AI systems in law enforcement and prohibitions on potentially harmful applications like social scoring and manipulative AI techniques (European Union, 2024). The framework mandates comprehensive impact assessments and human oversight, directly confronting critical concerns about algorithmic bias and accountability in criminal justice contexts (European Union, 2024). In contrast, the United States presents a fragmented regulatory landscape, with AI oversight dispersed across federal agencies such as the FBI and Department of Justice, and various state-level law enforcement bodies. This decentralized approach has resulted in inconsistent AI governance practices, highlighting the EU's more cohesive regulatory strategy (Bryan Cave Leighton Paisner, n.d.; Deloitte, 2023).

As AI systems become more autonomous and sophisticated, determining responsibility for AI-generated harms becomes increasingly complex. This complexity underscores the need for interdisciplinary collaboration between criminologists, legal scholars, AI researchers, and technology companies to develop appropriate frameworks for understanding and addressing AI-related crimes. Given the rapid advancement of AI capabilities, particularly in areas like LLMs and autonomous systems, establishing these collaborative frameworks is time-critical. Criminologists and legal scholars should proactively engage with leading AI research laboratories and academic institutions to conduct joint research on AI safety, alignment, and potential criminal exploitation before these systems are deployed. The success of such collaborative efforts will depend heavily on the development of clear regulatory guidelines and international cooperation mechanisms, particularly as AI technology continues to evolve at an unprecedented pace.

Several limitations of this research should be acknowledged. First, the rapidly evolving nature of AI technology means that research findings may quickly become outdated, particularly given the focus on sources from 2020-2024. Second, the emerging nature of AI criminology means there is limited empirical research specifically examining AI-enabled crimes outside of controlled laboratory settings. Third, the study's reliance on published academic and institutional sources may not fully capture the most recent developments in AI-enabled criminal activities, as there can be significant lag between criminal innovation and scholarly documentation. Fourth, the research primarily examines English-language

sources, potentially missing valuable insights from international perspectives on AI and criminal justice. Finally, while the paper discusses potential autonomous AI behavior, current technological limitations make it difficult to empirically evaluate this scenario beyond theoretical frameworks. These limitations highlight the challenges of studying an emerging field where technological capabilities, criminal applications, and regulatory responses are all rapidly evolving.

This research also calls future research attention for additional empirical studies examining the effectiveness of AI governance frameworks in criminal justice applications, particularly comparative analyses of different regulatory approaches. Additionally, interdisciplinary research collaborations between computer scientists, criminologists, and legal scholars should explore technical solutions for AI alignment in criminal justice contexts, with particular focus on preventing AI systems from engaging in or facilitating criminal activities.

Moreover, longitudinal studies tracking the evolution of AI-enabled criminal methodologies are essential, especially as LLMs and autonomous systems continue to advance. Furthermore, research should examine the psychological and sociological implications of AI in criminal justice, including public perceptions of AI-enabled law enforcement and the impact of automation bias in judicial decision-making. Studies should also specifically address the challenges of attributing criminal responsibility in cases involving autonomous AI systems, including the practical applications of traditional legal concepts like mens rea and actus reus to AI-enabled crimes.

Finally, research is needed on developing forensic methodologies specifically designed for investigating AI-enabled crimes, including techniques for analyzing AI system behavior and determining causality in complex algorithmic decisions. These research directions will be crucial for developing evidence-based policies and practices as AI continues to reshape the landscape of criminology and criminal justice.

Conclusion

The integration of AI into criminology and criminal justice represents both unprecedented challenges and opportunities for the field. Several key conclusions emerge from this research. First, while AI-enabled crimes currently operate within existing criminological frameworks, the rapid evolution of AI capabilities necessitates proactive adaptation of theoretical models and investigative approaches. The traditional tripartite framework of perpetrator, victim, and criminal justice system must expand to accommodate AI's unique role as a new actor capable of operating across multiple domains.

Moreover, the increasing sophistication of AI systems demands a reconsideration of criminal justice responses. The focus must shift from purely punitive approaches to strategies emphasizing prevention, detection, and system resilience. This includes developing new methodologies for investigating AI-enabled crimes and protecting AI systems from exploitation while maintaining their beneficial applications in law enforcement and criminal justice. Furthermore, the challenges of AI alignment and the potential for autonomous AI behavior necessitate unprecedented collaboration between criminologists, legal scholars, AI researchers, and technology developers. The development of effective governance frameworks must balance innovation with security, considering both the immediate risks of AI-enabled crime and the longer-term implications of increasingly autonomous AI systems.

The field of criminology must evolve to address these challenges while preserving fundamental principles of justice and human rights. This evolution requires not only theoretical adaptation but also practical innovations in law enforcement, judicial processes, and crime prevention strategies. As AI technology continues to advance, the ability to anticipate and address emerging forms of AI-related crime while harnessing AI's potential for positive applications in criminal justice will become increasingly crucial for maintaining social order and security in the digital age.

References

AIPRM. (n.d.). AI statistics 2024. https://www.aiprm.com/ai-statistics/

Anthropic. (2024, December 18). Alignment faking in large language models [Video]. YouTube. https://www.youtube.com/watch?v=9eXV64O2Xp8

Berk, R. A. (2021). Artificial intelligence, predictive policing, and risk assessment for law enforcement. Annual Review of Criminology, 4(1), 209–237. https://doi.org/10.1146/annurev-criminol-051520-012342

Bikeev, I., Kabanov, P., Begishev, I., & Khisamova, Z. (2019). Criminological risks and legal aspects of artificial intelligence implementation. In Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing (pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3371425.3371476

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020, May 28). Language models are few-shot learners. Cornell University. https://arxiv.org/abs/2005.14165

Bryan Cave Leighton Paisner. (n.d.). US state-by-state AI legislation snapshot. https://www.bclplaw.com/en-US/events-insights-news/us-state-by-state-artificial-intelligence-legislation-snapshot.html

Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. Crime Science, 9(14), 1–13. https://doi.org/10.1186/s40163-020-00123-8

Campedelli, G. M. (2022). Machine learning for criminology and crime research: At the crossroads. Routledge.

Dabney, D. (2013). Crime types: A text/reader (2nd ed.). Wolters Kluwer.

Davies, A., & Krame, G. (2023). Integrating body-worn cameras, drones, and AI: A framework

for enhancing police readiness and response. Policing: A Journal of Policy and Practice, 17, 1–3. https://doi.org/10.1093/police/paad083

Deloitte. (2023, January 9). How AI governance can adapt to a fragmented regulatory landscape. The Wall Street Journal. https://deloitte.wsj.com/riskandcompliance/how-ai-governance-can-adapt-to-a-fragmented-regulatory-landscape-26566914

European Union. (2024, February 27). High-level summary of the AI Act. https://artificialintelligenceact.eu/high-level-summary/

Federal Bureau of Investigation. (2024, December 3). Criminals use generative artificial intelligence to facilitate financial fraud [Public service announcement]. Internet Crime Complaint Center. https://www.ic3.gov/PSA/2024/PSA241203

Gates, B. (2023, March 21). The age of AI has begun. GatesNotes. https://www.gatesnotes.com/the-age-of-ai-has-begun

Gerritsen, C. (2020). Big data and criminology from an AI perspective. In B. Leclerc & J. Cale (Eds.), Big data (pp. 29–39). Routledge. https://doi.org/10.4324/9781351029704-3

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024, December, 20). Alignment faking in large language models. Cornell University. https://arxiv.org/pdf/2412.14093

Greg, S., Anastasia, P., & Cameron, R. (2017). Crime and justice in digital society: Towards a 'digital criminology'? Crime Justice Journal, 6(2), 17-33. DOI: 10.5204/ijcjsd.v6i2.355

Harcourt, B. E. (2006). Against prediction: profiling, policing, and punishment in an actuarial age. The University of Chicago Press.

Hayward, K. J., & Maas, M. M. (2021). Artificial intelligence and crime: A primer for criminologists. Crime, Media, Culture, 17(2), 209–233. https://doi.org/10.1177/1741659020917434

Helfgott, J. B. (2008). Criminal behavior: Theories, typologies, and criminal justice. Sage.

Huberman, A. (Host). (2024, November 18). Dr. Terry Sejnowski: How to improve at learning using neuroscience and AI [Audio podcast episode]. Huberman Lab. https://podcasts.apple.com/us/podcast/huberman-lab/id1545953110?i=1000677308237

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects.

Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kaufmann, M. (2024). AI in policing and law enforcement. In R. Paul, E. Carmel, & J. Cobbe (Eds.), Handbook on public policy and artificial intelligence (pp. 295–306). Edward Elgar Publishing. https://doi.org/10.4337/9781803922171.00031

Kirwan, G., & Power, A. (2013). Cybercrime: The psychology of online offenders. Cambridge University Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

McDaniel, J. L., & Pease, K. (Eds.). (2021). Predictive policing and artificial intelligence. Routledge.

Morgan, L. (2024, December, 5). Is open source a threat to national security? Information Week. https://www.informationweek.com/software-services/is-open-source-a-threat-to-national-security-

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

Paré, G., Trudel, M. C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. Information & Management, 52(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Perez, C., & Maharaj, T. (2022, February 7). Red teaming language models with language models. Cornell University. https://arxiv.org/abs/2202.03286

Powell, A., Stratton, G., & Cameron, R. (2018). Digital criminology: Crime and justice in digital society (1st ed.). Routledge. https://doi.org/10.4324/9781315205786 Romero, L. E. (2025, January 28). ChatGPT, DeepSeek, or Llama? Meta's LeCun says open-source is the key. Forbes. https://www.forbes.com/sites/luisromero/2025/01/27/chatgpt-deepseek-or-llama-metas-lecun-says-open-source-is-the-key/

Sejnowski, T. J. (2024). Chat GPT and the future of AI. The MIT Press.

Snider, S. (2023, August 20). Tesla insider data breach exposed over 75,000. Information Week. https://www.informationweek.com/cyber-resilience/tesla-insider-data-breach-exposed-over-75-000

Statista Research Department. (2024, December 12). Expected number of artificial intelligence (AI) powered tools users in the United States from 2020 to 2030. https://www.statista.com/forecasts/1451316/us-ai-tool-user-number

Stratton, G., Powell, A., & Cameron, R. (2017). Crime and justice in digital society: Towards a 'digital criminology'? Crime Justice Journal, 6(2), 17-33. https://doi.org/10.5204/ijcjsd.v6i2.355

U.S. Department of Homeland Security. (2024). Impact of artificial intelligence on criminal and illicit activities. https://www.dhs.gov/sites/default/files/2024-10/24_0927_ia_aep-impact-ai-on-criminal-and-illicit-activities.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, August 2). Attention is all you need. Cornell University. https://arxiv.org/abs/1706.03762

Vincent, J. (2016, March 24). Twitter taught Microsoft's AI chatbot to be a racist conspiracy theorist in less than a day. The Verge. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Walrath-Holdridge, M. (2025, January 28). OpenAI introduces ChatGPT Gov, an artificial intelligence chatbot for government agencies. USA Today. https://www.usatoday.com/story/tech/2025/01/28/openai-chatgpt-gov-artificial-intelligence/77995373007/

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. European Journal of Information Systems, 22(1), 45–55. https://doi.org/10.1057/ejis.2011.51

About the Author

Mingxi Tong is a Master's degree candidate in the Department of Criminal Justice, Criminology and Forensics at Seattle University.