Issues in Information Systems
Volume 26, Issue 4, pp. 1-16, 2025

DOI: https://doi.org/10.48009/4_iis_2025_101

The UN cybercrime treaty and AI: Navigating the intersection of technology and global policy

Angelica Marotta, *MIT Sloan School of Management, amarotta@mit.edu*
Stuart Madnick, *MIT Sloan School of Management, smadnick@mit.edu*

Abstract

The rapid evolution of Artificial Intelligence (AI) has fundamentally transformed cybersecurity, posing significant challenges to existing legal frameworks. This research assesses the adequacy of the United Nations Convention against Cybercrime (we refer to as the UN Cybercrime Treaty) in addressing AI-enabled cyber threats by analyzing over 70 documented incidents from 2023 to 2025. By mapping the complex interactions between advanced AI technologies and the treaty's provisions, the research exposes critical regulatory gaps across different incident types. Key findings show how the treaty's effectiveness varies across different types of AI-related crimes. This highlights the need for governance strategies that are both flexible and easy to apply across disciplines. In addition, the study reveals AI as more than a tool—an autonomous actor capable of generating sophisticated cyber-attacks that challenge traditional concepts of criminal intent. Ultimately, the research contributes to ongoing discussions on technological governance by providing recommendations tailored to the evolving landscape of AI-related cybercrime.

Keywords: artificial intelligence, cybersecurity, cybercrime, international law, governance

Introduction

The rapid advancement of AI presents opportunities to improve cybersecurity, such as automation of threat detection, but also raises challenges, including more sophisticated cyberattacks and regulatory complexities. This evolving landscape has necessitated a coordinated global response, culminating in the adoption of the United Nations Convention on Cybercrime (also referred to as the UN Cybercrime Treaty) in 2024, following a five-year negotiation process (United Nations General Assembly, 2024). The treaty is scheduled to open for signature at a formal ceremony in Hanoi, Viet Nam, in October 2025 (United Nations General Assembly, 2024). This landmark agreement represents a significant step towards establishing a unified international framework for combating cyber-related offenses in an increasingly AI-driven world.

However, when AI and cybercrime intersect, they create new challenges for governments and law enforcement. Traditional legal systems often struggle to respond effectively to these fast-changing threats. The UN Cybercrime Treaty's technology-neutral approach—meaning it avoids referencing specific technologies to remain applicable over time—aims to provide a flexible framework that can adapt to emerging technologies. Nevertheless, questions remain about its effectiveness in dealing with the specific issues posed by AI-related cybercrime. For this reason, the primary research objectives of this study are to analyze the extent to which the UN Cybercrime Treaty addresses AI-related cybercrime, evaluate its coverage and potential gaps, and identify and assess key challenges in applying the treaty's provisions to AI technologies.

Main Contributions

This study explains how the UN Cybercrime Treaty applies to AI-related crimes. It identifies key challenges with the treaty's technology-neutral approach and offers ideas on how to support innovation while reducing misuse. In particular, this research contributes to the field of international cybercrime policy by (1) providing evidence of how emerging AI-enabled cyber threats interact with current treaty provisions, (2) offering a critical assessment of the treaty's applicability to these new challenges, and (3) proposing recommendations to enhance the treaty's effectiveness. Our findings provide valuable insights for policymakers, legal scholars, organizations, and cybersecurity professionals seeking to enhance international legal responses to AI-enabled cyber threats while promoting responsible AI development.

Background and Literature Review

This section provides a review of the existing literature relevant to our study, focusing on three key areas: AI and cybercrime, international cybercrime legislation, and AI governance and regulation.

AI and Cybercrime

The role of AI in cybercrime has emerged as a significant area of research in recent years, reflecting the growing concern over the potential misuse of AI technologies in criminal activities (Guembe et al., 2037). Brundage et al. (2018) offer a comprehensive overview of the potential security threats posed by AI, including its use in cyberattacks. Their research highlights the dual-use nature of AI technologies and the challenges this presents for policymakers and security professionals. The authors contend that AI systems have the potential to amplify the scale, speed, and effectiveness of attacks, which could reduce their costs while increasing their overall impact (Brundage et al., 2018).

Building on this foundation, Caldwell et al. (2020) argue that the increasing capabilities of AI pose significant threats to society through potential criminal applications. Their study, which involved a workshop with diverse stakeholders, identified and ranked 18 AI-enabled crime categories, with audio/video impersonation, weaponized driverless vehicles, and tailored phishing emerging as top concerns. Their work underscores the need for adaptive cybersecurity measures that can keep pace with AI-powered threats. The authors note that it is necessary to consider a trade-off between harm and "defeatability" (i.e., the difficulty of preventing, detecting, or counteracting a potential AI-enabled crime) when prioritizing which threats to address. Similarly, King et al. (2020) explore the potential of AI in cybercrime prevention and detection, highlighting the double-edged nature of AI in the cybersecurity landscape. They argue that while AI can be used to perpetrate sophisticated cyberattacks, it also offers powerful tools for defending against such threats. This perspective emphasizes the importance of considering both the offensive and defensive capabilities of AI in cybercrime policy.

International Cybercrime Legislation

The development of international cybercrime legislation has been the subject of extensive scholarly attention, reflecting the global nature of cyber threats and the need for coordinated responses (Kshetri, 2013; Wicki-Birchler, 2020). For example, Weber & Studer (2016) emphasize the need for flexible and innovative legal frameworks to keep pace with rapidly evolving security threats in the international context. In particular, the authors propose considering a polycentric regulation model, which involves multiple stakeholders from both the public and private sectors in creating and implementing cybersecurity rules. This approach allows for more adaptive and collaborative solutions, contrasting with traditional top-down regulatory methods.

Weber & Studer (2016)'s analysis provides valuable context for understanding the complexities of international cybercrime legislation and the potential benefits of multi-stakeholder governance in addressing global cybersecurity issues. Similarly, Buçaj & Idrizaj (2025) argue that controlling cybercrime through international law principles and treaties is essential, highlighting the need for a global legal foundation to prevent and punish cyber criminals. The authors mention the Budapest Convention as a key agreement for addressing cybercrime and electronic evidence, noting that rapid technological advancements necessitate continuous adaptation of legal instruments. The study also underscores the importance of harmonizing legal definitions across jurisdictions. Along the same line, Tropina and Cormac (2015) explore the development of self- and co-regulatory approaches to cybercrime and cybersecurity in a multi-stakeholder environment. The authors argue that the increasing use of computers, networks, and the Internet has necessitated regulation in the fields of cybercrime, cybersecurity, and national security.

AI Governance and Regulation

The governance and regulation of AI technologies have been widely discussed in recent literature, reflecting growing concerns about the ethical and legal implications of AI development and deployment. Cath et al. (2018) provided initial visions for AI governance in their work. The EU's AI Act and subsequent deregulatory shifts demonstrate the ongoing challenge of balancing innovation with ethical and security concerns. For the UN Cybercrime Treaty, these developments suggest the need for a flexible approach that can adapt to rapid technological changes while addressing the global nature of AI-related cybercrime threats. Scherer (2015) examines the challenges of regulating AI from a legal perspective, highlighting the difficulties in crafting legislation that can keep pace with rapid technological advancements.

Key challenges include the vague definition of "intelligence" in AI, issues of control due to AI's autonomy, and the discreet, decentralized, and opaque nature of its development. The author's insights inform our discussion of the treaty's effectiveness in addressing AI-related cybercrime, particularly in terms of the need for adaptive regulatory frameworks. Floridi et al. (2018) discuss the ethical challenges posed by AI and propose a framework for developing AI for social good. Their work emphasizes the importance of considering the broader societal implications of AI technologies, including their potential impact on crime and security. This perspective is central to our analysis of the UN Cybercrime Treaty's approach to balancing the benefits and risks of AI technologies.

This literature review provides the theoretical foundation for our study, highlighting the complex interplay between AI, cybercrime, and international legislation. Building on this conceptual groundwork, the following section outlines the methodological approach used to examine how these insights play out in practice.

Methodology

Our research employed a dual approach, combining qualitative analysis of the UN Cybercrime Treaty text with quantitative and qualitative analysis of AI-related cybercrime incidents. This research design, commonly used in policy and legal research, facilitated the triangulation of data sources and enhanced the validity of findings by integrating doctrinal legal analysis with empirical data (Creswell, 2017; Creswell et al., 2007; Creswell & Poth, 2017). In particular, this methodology allowed us to examine both the legal framework provided by the treaty and its practical application to real-world AI-related cybercrime cases.

Treaty Analysis

We conducted a comprehensive analysis of the UN Cybercrime Treaty, focusing on provisions relevant to AI technologies and their potential applications in the context of AI-related cybercrime. Our process involved the following steps:

1. **Identification of Relevant Provisions**: We systematically reviewed the treaty to identify articles that could potentially apply to AI-related cybercrime. These provisions included

definitions related to cybercrime, investigative powers, international cooperation, and electronic evidence.

2. **Assessment of Flexibility and Adaptability**: We evaluated the identified provisions for their flexibility and potential adaptability to evolving AI technologies. This phase involved analyzing the language used in the treaty and assessing whether the provisions were sufficiently broad to encompass future developments in AI.

Incident Review

To assess the practical applicability of the treaty provisions, we conducted a systematic review of AI-related cybercrime incidents. Our primary data source was the AI Incident Database, which compiles reports of AI-related incidents from various sources (*Artificial Intelligence Incident Database*, n.d.). Our review process involved the following steps:

Sample Selection: We selected a sample of over 70 AI-related cybercrime incidents (available at: AI Incident Database), covering the period from 2023 to 2025. This timeframe was chosen to align with the rapid evolution of AI technologies and notable temporal spikes in cybercrime activities, such as the surge in voice-cloning scams in early 2025. Additionally, this period encompasses significant milestones in the development of the treaty (e.g., the agreement on the draft convention). In addition, the selection covers a geographically diverse set of regions—including North America, Europe, Asia, Africa, and Oceania—highlighting how AI-related attacks such as deep-fake-based fraud, automated phishing campaigns, synthetic-identity scams, and unauthorized facial-recognition deployments have proliferated across different regulatory and cultural environments.

- **Analytical Approach:** Our quantitative and qualitative analyses involved the following multiple steps:
    - **Distribution of Incident Types:** We conducted a systematic categorization and statistical assessment to quantify and classify various forms of AI-related cybercrime, using AI incident typologies and concepts from sources, such as the MIT AI Risk Repository—which includes 1,421 risks from 65 frameworks, organized by cause and domain—to systematically categorize and analyze AI-enabled cybercrime incidents (*The MIT AI Risk Repository*, n.d.).
    - **Geographical Distribution Mapping:** We analyzed the global spread of AI-related cybercrime incidents to identify geographic trends.
    - **Qualitative Analysis of Incident Narratives:** To uncover emerging patterns and contextual challenges, we performed a rigorous analysis of incident reports associated with incidents, employing thematic coding and narrative discourse analysis to extract deeper insights into the dynamics and implications of AI-related cybercrime.

Synthesis and Evaluation

To address our research objectives, we evaluated the potential effectiveness of the treaty's provisions in addressing the AI-related cybercrime incidents in our sample, considering the following parameters for each type identified in the previous analysis:

- **Prevalence:** Measures the frequency of specific incident types. A higher prevalence indicates a more critical area that requires immediate legal attention. The scoring reflects the urgency and importance of addressing these incidents.
- **Provision Coverage:** Evaluates the depth and specificity of legal provisions covered in the treaty. Strong coverage means comprehensive, proactive guidelines, while limited coverage suggests reactive, minimal requirements that may struggle to address emerging challenges.
- **Geographical Scope:** Assesses the potential extent and effectiveness of the treaty's coverage across multiple countries or regions in addressing AI-related cybercrime based on the countries involved in the analysis of incidents. A higher number of countries indicates a potential broader international reach, while a lower number suggests a more localized impact.

- **Challenge Complexity:** Examines the interconnectedness and sophistication of challenges. Multifaceted challenges require holistic and integrated approaches, while simple challenges can be addressed with more direct and focused strategies.

These parameters represent a multidimensional assessment framework designed to evaluate the effectiveness of complex policies through a structured, quantitative approach. Each parameter—Prevalence, Provision Coverage, Geographical Scope, and Challenge Complexity—was meticulously calibrated to capture different dimensions of systemic performance, as shown in Table 1.

Table 1. Core Parameters Table

Parameter	Definition	Assessment Levels	Quantitative Weighting
Prevalence	Frequency and significance of incidents	High (>20%) Moderate (10-20%) Low (<10%)	High: 3 weighted units Moderate: 2 weighted units Low: 1 weighted unit
Provision Coverage	Comprehensiveness of legal provisions	Strong (broad, detailed coverage of key issues), Adequate (covers most core AI cybercrime areas), Moderate (partial coverage with notable gaps) Limited (Minimal or unclear provisions)	Strong: 3 weighted units Adequate: 2-3 weighted units Moderate: 2 weighted units Limited: 1 weighted unit
Geographical Scope	International reach and jurisdictional effectiveness	Comprehensive Global (≥12 countries), Partial Global (7-11 countries) Localized (1-6 countries)	Comprehensive: 3 weighted units Partial: 2 weighted units Localized: 1 weighted unit
Challenge Complexity	Multidimensionality of challenges	Multifaceted (3+ issues) Moderate (2 issues) Simple (1 issue)	Multifaceted: 3 weighted units Moderate: 2 weighted units Simple: 1 weighted unit

The quantitative weighting mechanism converts qualitative assessments into a standardized numerical scale, enabling nuanced and comparable evaluations. By assigning weighted units ranging from 1 to 3, the framework creates a granular scoring system that captures subtle variations in effectiveness. Detailed scoring enables differentiation between near-equivalent performance levels, moving beyond binary classifications of effective or ineffective.

The three-tier assessment levels for each parameter reflect an increasing level of sophistication and comprehensiveness. For instance, in Prevalence, a high-impact classification (greater than 20%) receives 3 weighted units, signaling critical areas that demand immediate attention, while low-impact incidents receive 1 weighted unit, indicating peripheral concerns. These assessment levels are designed to enhance intersubjective reliability. In particular, the assessment criteria were operationalized using legal benchmarks, including the presence of enforceable provisions, specificity of language, and alignment with known AI-related risks. Each rating was determined through a multi-stage review process that incorporated doctrinal legal analysis, comparative treaty interpretation, and incident-specific mapping.

The cumulative assessment, totaling 12 possible weighted units, provides a holistic perspective. The stratified effectiveness ratings—Highly Effective (9-12 units), Moderately Effective (5-8 units), and Requiring Significant Improvement (0-4 units)—provide a clear and actionable framework for policy refinement and strategic intervention. The final evaluation is calculated by averaging the weighted units for each incident type. This structured evaluation framework sets the stage for a focused analysis of the patterns emerging from our treaty and incident data.

Analysis

This section presents a synthesis of our treaty analysis and incident review, highlighting key patterns and relationships identified in our data.

Treaty Analysis: Provisions Relevant to AI

The UN Cybercrime Treaty is a comprehensive international treaty designed to combat cybercrime on a global scale. It establishes a common legal framework for addressing various forms of cybercrime, enhancing international cooperation, and implementing preventive measures. It defines and criminalizes a wide range of cyber offenses, including illegal access, illegal interception, system interference, and computer-related fraud. It grants law enforcement agencies specific powers for cybercrime investigations, including the expedited preservation of electronic data, the search and seizure of stored data, and the real-time collection of traffic data.

The treaty emphasizes cross-border collaboration through provisions, such as mutual legal assistance and the establishment of a 24/7 network for immediate assistance. It addresses the complex jurisdictional challenges posed by cybercrimes that often transcend national borders. The treaty encourages the development of policies and best practices to prevent cybercrime and promotes technical assistance and capacity-building efforts among member states. In addition, the UN Cybercrime Treaty aims to combat cybercrime by harmonizing laws, improving investigations, and enhancing international cooperation, while safeguarding human rights and fundamental freedoms. Ultimately, it offers a flexible framework that can be tailored to address evolving cyber threats while respecting the sovereignty of individual nations.

However, while the treaty provides a broad framework for combating cybercrime, its application to emerging technologies such as AI warrants closer examination. Table 2 highlights key provisions from our analysis that are particularly relevant to AI-related cybercrime and their potential applications.

Table 2. Key Treaty Provisions and Their Relevance to AI-Related Cybercrime

Treaty Article	Key Provisions	Relevance to AI-Related Cybercrime
Article 7: Illegal Access	Criminalizes unauthorized access to information and communications technol ogy systems	Applicable to unauthorized access to AI systems or training data
Article 8: Illegal Interception	Criminalizes the interception of non-public transmissions of electronic data	Applicable to unauthorized interception of AI model training data or AI-generated communications
Article 10: System Interference	Prohibits intentional hindering of the functioning of information and communications technology systems	Relevant to attacks that manipulate or disrupt AI systems
Article 11: Misuse of Devices	Prohibits the production, sale, or possession of devices designed primarily for committing cybercrimes	Relevant to AI tools designed for malicious purposes, such as automated hacking systems
Article 13: Information and communications technology system-related theft or fraud	Criminalizes the causing of loss of property through manipulation of computer data or system interference	Applicable to AI-powered fraud schemes, including deepfake-based fraud
Article 14: Offences related to online child sexual abuse or child sexual exploitation material	Criminalizes various activities related to child sexual abuse material online	Relevant to AI-generated child sexual abuse material or AI systems used to detect such content

Treaty Article	Key Provisions	Relevance to AI-Related Cybercrime
Article 16: Non-consensual dissemination of intimate images	Criminalizes the distribution of intimate images without consent	Applicable to AI-generated deepfake pornography or revenge porn
Article 25: Expedited preservation of stored electronic data	Allows authorities to order the preservation of specific electronic data	Crucial for preserving AI-related evidence that might be volatile or easily altered
Article 28: Search and seizure of stored electronic data	Empowers authorities to search and seize electronic data	Necessary for investigating AI systems used in criminal activities
Article 29: Real-time Collection of Traffic Data	Grants authorities the power to collect or record traffic data in real-time	Necessary for investigating AI-powered distributed attacks
Article 30: Interception of Content Data	Allows for the real-time collection or recording of content data for serious offenses	Relevant for monitoring AI-generated malicious content
Article 31: Freezing, seizure, and confiscation of the proceeds of crime	Enables the confiscation of proceeds derived from cybercrimes	Applicable to profits generated from AI-powered cybercrime operations
Article 40: Mutual Legal Assistance	Establishes framework for international cooperation in cybercrime investigations	Essential for addressing the cross-border nature of many AI-related cybercrimes
Article 41: 24/7 Network	Establishes a network of contact points available 24/7 for immediate assistance	Essential for rapid response to evolving AI-powered cyber threats
Article 53: Preventive measures	Encourages the development of policies and best practices to prevent cybercrime	Relevant for developing AI-specific cybercrime prevention strategies
Article 54: Technical assistance and capacity-building	Promotes cooperation in training and technical assistance	Important for building global capacity to address AI-related cybercrime

The treaty's provisions cover a wide range of cyber offenses and investigative powers that can be adapted to address AI-specific challenges. For instance, Article 7 on illegal access could apply to unauthorized intrusions into AI systems or datasets, while Article 8 might cover the interception of data used for training AI models. The Convention also addresses system interference (Article 10), which could be relevant to attacks that manipulate or disrupt AI operations, such as adversarial attacks on machine learning models. The treaty's technology-neutral language allows for flexibility in interpretation, potentially encompassing emerging AI-related crimes. For example, Article 13 on fraud could be applied to sophisticated AI-powered schemes, such as deep-fake-based financial fraud.

Similarly, Article 16's provisions on non-consensual dissemination of intimate images could extend to AI-generated deepfake pornography. Investigative powers granted by the UN Cybercrime Treaty, such as expedited data preservation (Article 25) and real-time traffic data collection (Article 29), are crucial for gathering evidence in AI-related cases where data can be volatile or distributed across multiple jurisdictions. The treaty also emphasizes international cooperation through mutual legal assistance (Article 40) and a 24/7 network (Article 41), essential for addressing the often cross-border nature of AI-enabled cybercrimes.

However, while the UN Cybercrime Treaty's technology-neutral approach provides adaptability, it may also present challenges in addressing AI-specific issues. For instance, the treaty may not fully account for the unique characteristics of AI systems, such as their potential for autonomous operation or the complexities of attributing AI-driven attacks. Additionally, the rapid evolution of AI technologies may outpace the treaty's provisions, necessitating ongoing interpretation and potential updates to maintain its effectiveness in combating AI-related cybercrime. Therefore, a deeper analysis of AI-related cybercrime incidents is essential to highlight the unique challenges AI poses in cybercrime and to assess how effectively the Convention's provisions address these issues.

Incident Review: Analysis of AI-Related Cybercrime Incidents

The incidents analyzed in our database encompass a diverse range of AI-related cybercrimes across different geographical regions, reflecting the multifaceted nature of AI's application in malicious activities.

Distribution of Incident Types

The incidents analyzed encompass a diverse range of AI-related cybercrimes, reflecting the multifaceted nature of AI's application in malicious activities.

Table **3** presents a detailed breakdown of the incidents by category and frequency.

Table 3. Distribution of Incident Types

Category	Percentage
Deepfake / Synthetic Media Scams	33%
AI-Generated Impersonation / Fraud	18%
Phishing / Credential Theft	9%
Malicious Use of AI Tools	11%
Misuse of Facial Recognition / Surveillance AI	9%
Inappropriate / Harmful AI Outputs	9%
CSAM / Sextortion / Abuse Content	6%
Other AI Misuse (e.g., IRS scams, misinformation)	5%
Total	**100%**

The distribution of incident types in AI-related cybercrime highlighted a diverse and growing threat landscape. Deepfake and synthetic media scams represent the most significant proportion at 33%, encompassing the use of AI to generate realistic but false videos, cloned voices, and fabricated digital identities—often to manipulate public opinion or commit fraud. AI-generated impersonation and fraud, accounting for 18%, include incidents where avatars or synthetic personas are used to deceive individuals by mimicking colleagues, friends, or executives in professional settings. Phishing and credential theft, responsible for 9% of the incidents, involve deceptive attempts to acquire sensitive information such as passwords or financial data. Notably, these cases sometimes leverage advanced AI tools, such as Gamma—an AI-powered presentation generator—which has been implicated in phishing campaigns by producing convincing but fraudulent content (Baran, 2025).

Malicious use of AI tools, comprising 11%, refers to the deployment of AI for offensive purposes such as developing more efficient ransomware, enhancing malware, or enabling autonomous cyberattacks. Misuse of facial recognition and surveillance AI, representing 9%, includes documented instances of wrongful arrests due to algorithmic misidentification. Moreover, the same percentage of the cases involved inappropriate or harmful AI outputs, including examples where chatbots encouraged self-harm or generated false medical advice. Another 6% of incidents relate to content involving child sexual abuse material (CSAM), sextortion, and abuse—categories that include the disturbing trend of AI being used to create deepfake nudes of minors or simulate abuse scenarios. Additionally, a further 5% of incidents fall under broader misuse categories, including misinformation and fraudulent schemes, such as impersonated tax scams.

In addition, a closer examination of the underlying incident data revealed that privacy-related harms are deeply embedded across multiple types of AI misuse. These privacy breaches account for approximately 11% of all reported incidents and often span categories rather than conform to a single typology. For instance, a significant health data breach linked to the Serviceaide AI Platform exposed the records of over 483,000 patients, raising concerns about AI systems managing sensitive medical data (Serviceaide, 2025). In another case, Clearview AI, a company known for its facial recognition software, was fined €33.7 million by French regulators for violating the EU's General Data Protection Regulation (GDPR) through unauthorized harvesting of biometric data (*Clearview AI Fined by Dutch Agency for Facial Recognition Database | Reuters*, 2024). Similarly, an alleged AI-driven call center breach exposed over 10 million recorded conversations across the Middle East, highlighting the potential scale of data compromise when AI systems are involved (*Cybercriminals Are Targeting AI Agents and Conversational Platforms: Emerging Risks for Businesses and Consumers*, 2024). These examples illustrate how privacy violations intersect with various types of AI misuse, suggesting that privacy is not merely a distinct category but a cross-cutting issue within the broader AI risk landscape. This analysis underscores the necessity for integrated data protection strategies, robust AI governance frameworks, and strict adherence to ethical and legal standards to ensure the responsible development and deployment of AI technologies.

Geographical Distribution and Cross-Border Nature

Table 4 provides an overview of the distribution of various AI-related cybercrime incident types across different countries and regions. The breakdown highlights a notable cross-border dimension, with many incidents involving perpetrators and victims located in different jurisdictions. This assessment reflects both the transnational nature of AI-driven threats and the challenges in attribution and enforcement.

Table 4 - Geographical Distribution

Incident Type	US	SG	UK	ES	AU	MT	CA	CY	HK	CN	IN	KE	NG	NO	NZ	Other/ Multi-region
Deepfake & Synthetic Media Scams	✓	✓	✓	✓				✓			✓		✓			✓
AI-Generated Impersonation & Fraud	✓	✓		✓	✓	✓	✓	✓		✓	✓					
Phishing & Credential Theft	✓		✓						✓			✓	✓			✓
Malicious Use of AI Tools	✓										✓					✓
Misuse of Facial Recognition / Surveillance AI	✓		✓			✓										✓
Inappropriate / Harmful AI Outputs	✓	✓			✓		✓							✓	✓	
CSAM / Sextortion / Abuse Content	✓										✓		✓			✓

The United States (US) stands out as the country with the broadest exposure. This prominence may reflect a combination of advanced digital infrastructure, widespread adoption of AI, and its role as a global economic and technological hub. Singapore (SG) and the United Kingdom (UK) also feature prominently across multiple incident types. Singapore is primarily associated with Deepfake scams, AI-generated impersonation, and harmful AI outputs (privacy-related harms), likely due to its highly digitized economy and role as a financial hub. The UK appears in categories such as Deepfakes, Phishing, and surveillance-related incidents, suggesting a similar combination of technological advancement and cybercrime exposure. Other countries, such as Spain (ES), Australia (AU), Malta

(MT), China (CN), and Canada (CA), are associated with more specific types of incidents, including impersonation fraud and Harmful AI outputs. Despite their varying sizes and populations, these countries' advanced digital infrastructures may make them attractive targets for cybercriminals. Smaller or emerging economies, such as Cyprus (CY), Hong Kong (HK), India (IN), Kenya (KE), and Nigeria (NG), are particularly vulnerable to targeted incidents, including phishing, Synthetic media scams, and CSAM/sextortion. This evaluation demonstrates that AI-related cybercrime is not limited to large or highly developed nations but is a global phenomenon that impacts diverse economic and regulatory environments. Finally, countries such as Norway (NO) and New Zealand (NZ) appear in a limited number of categories, suggesting more isolated but still significant instances of AI-related cybercrime. The "Other / Multi-region" category spans across most incident types, underlining the cross-jurisdictional complexity of these crimes and highlighting the need for international collaboration and adaptable cybersecurity strategies.

Synthesis and Evaluation: Comparative Analysis of Treaty Provisions and Incident Trends
By juxtaposing the treaty provisions with the observed trends in AI-related cybercrime incidents (Table **5**), we can evaluate the treaty's potential efficacy and identify challenge areas (the 'Other AI Misuse' category was excluded from the comparative analysis due to its heterogeneous nature):

Table 5 - Comparative Analysis

Incident Type	Prevalence	Provision Coverage	Geographical Scope	Challenge Complexity	Total Weighted Units	Effectiveness Rating
Deepfake and Synthetic Media Scams (33%)	High (3)	Adequate (2)	Partial Global (2)	Multifaceted (3)	10	Highly Effective
AI-Generated Impersonation and Fraud (18%)	Moderate (2)	Adequate (2)	Partial Global (2)	Moderate (2)	8	Moderately Effective
Phishing and Credential Theft (9%)	Low (1)	Adequate (2)	Partial Global (2)	Simple (1)	6	Moderately Effective
Malicious Use of AI Tools (11%)	Moderate (2)	Moderate (2)	Localized (1)	Multifaceted (3)	8	Moderately Effective
Misuse of Facial Recognition/Surveillance AI (9%)	Low (1)	Moderate (2)	Localized (1)	Moderate (2)	6	Moderately Effective
Inappropriate/Harmful AI Outputs (9%)	Low (2)	Limited (1)	Localized (1)	Simple (1)	5	Requiring Significant Improvement
CSAM and Sextortion (6%)	Low (1)	Strong (3)	Comprehensive Global (3)	Multifaceted (3)	10	Highly Effective

The following findings section builds on this comparison to assess how effectively the treaty addresses specific AI-related threats.

Findings

The UN Cybercrime Treaty demonstrates valuable effectiveness in addressing the unique challenges posed by AI-generated CSAM, which can be generated, manipulated, and distributed at unprecedented scales and with alarming sophistication—with Article 14 (Offences related to online child sexual abuse or child sexual exploitation material) serving as a prime example of a robust requirement. This article provides an integrated approach to combating AI-generated exploitative content by comprehensively criminalizing the creation, distribution, and possession of such material. The Treaty's strength lies in its ability to establish cross-border cooperation for investigating and prosecuting these offenses while simultaneously implementing technical measures to detect and prevent AI-generated exploitative content. Moreover, the provisions ensure the protection of potential victims' privacy and rights; Articles 29 (Real-time Collection of Traffic Data) and 30 (Interception of Content Data) offer mechanisms for investigating and potentially mitigating data breaches and unauthorized use of personal information. Another promising aspect of the treaty emerged in its approach to specific types of AI-driven incidents. For deepfake and synthetic media scams, Articles 13 (Information and Communications Technology System-related Theft or Fraud) and 16 (Non-consensual Dissemination of Intimate Images) provide a foundation for addressing these emerging threats.

In contrast, the analysis revealed significant limitations in addressing other AI-related cybercrime scenarios, particularly those involving inappropriate AI outputs and localized misuse of AI. These areas received lower effectiveness ratings, underscoring the urgent need for more adaptable legal frameworks. For instance, while Articles 10 (System Interference) and 11 (Misuse of Devices) provide some framework for addressing malicious AI tool usage, they cannot fully encompass the nuanced challenges presented by AI systems that produce misleading information or potentially harmful content.

Geographical distribution analysis further highlighted the treaty's limitations in addressing localized instances of AI misuse. Articles 40 (Mutual Legal Assistance) and 41 (24/7 Network) attempt to provide a global framework but struggle to create a consistent, universally applicable approach to technological threats. Additionally, Articles 25 (Expedited Preservation of Stored Electronic Data) and 28 (Search and Seizure of Stored Electronic Data) provide investigative tools. However, they may not fully address the intricate challenges posed by cross-border AI-related cybercrimes. Finally, Articles 53 (Preventive Measures) and 54 (Technical Assistance and Capacity-Building) attempt to bridge some of the gaps relating to capacity disparities and institutional readiness in combating AI-driven cyber threats but fall short of providing the comprehensive protection seen in other treaty areas. These findings highlight critical gaps that the discussion section further explores in relation to the evolving nature of AI-related cybercrime and the applicability of the Treaty.

Discussion

The analysis conducted in this paper revealed a significant transformation in the scope and mechanisms of cybercrime as shaped by artificial intelligence. AI technologies, particularly large generative models and reinforcement-learning agents, have redefined how cybercrime is executed—amplifying the scale of operations, accelerating the rapidity of attacks, and introducing unprecedented complexity. As observed in the analyzed incidents, these systems enable dynamic and persistent threats, ranging from deepfake-driven impersonations and scalable phishing to adversarial manipulations and data poisoning. These developments have outpaced the assumptions embedded in the UN Cybercrime Treaty, which was under development over a long period of time, exposing structural gaps that hinder practical interpretation and enforcement in AI-relevant contexts.

Furthermore, the geographic breadth of the analyzed incidents further demonstrates the global reach and decentralized nature of AI-driven cybercrime. The majority of incidents studied involved multiple jurisdictions, often with attackers deliberately exploiting inconsistencies in legal and technical infrastructure. For instance, this challenges the effectiveness of Articles 40 and 41, which provide a foundation for mutual legal assistance and rapid cooperation but lack operational detail on how to share, interpret, and validate AI-specific forensic data.

Currently, there is no shared system for exchanging AI-related evidence, understanding AI decisions, or safely sharing data across countries. Without this, the treaty’s goals for international cooperation may not be realistic. Moreover, institutional disparities in AI competence and resources—particularly between high-capacity and under-resourced states—may exacerbate implementation challenges. Our assessments uncovered possible uncertainty among several national authorities regarding how to categorize, investigate, or prosecute incidents involving generative or adversarial AI. The absence of shared technical standards for AI forensics and the limited accessibility of interpretability tools result in enforcement efforts that are often fragmented and reactive. Compounding this is the speed of AI innovation: our temporal mapping of threat evolution shows that novel attack techniques appear faster than legislative cycles can accommodate, making a static biennial review mechanism potentially obsolete by design. These observations reveal that while the treaty's technology-neutral posture offers theoretical adaptability, the practical application of its provisions to AI-related threats remains uneven and, in many cases, insufficient. The risks of misclassification, underenforcement, or regulatory overreach are amplified when there are no dedicated mechanisms to interpret AI's unique modalities within a robust legal framework. This discussion outlines AI’s transformative impact on cybercrime and the resulting limitations of the treaty, framing the subsequent exploration of both theoretical and practical implications.

Implications for Theory and Practice

This research critically challenges established cybercrime paradigms by revealing a profound theoretical and practical dilemma: how to understand and regulate *technological agency* within existing legal frameworks. Technological agency refers to the capacity of technological systems—particularly AI—to perform actions, make decisions, and produce outcomes with legal, social, or economic consequences, often without direct human initiation or oversight (Candrian & Scherer, 2022). This agency does not stem from consciousness or intent, but rather emerges from the complex, autonomous behaviors of algorithms operating within expansive data environments and interacting with human inputs.

Theoretical Implications and Recommendations: This study traces the emergence of the technological agency concept through evidence of AI systems engaging in activities with potential criminal significance—such as performing social engineering attacks, manipulating digital identities, or generating malicious synthetic media. Thus, theoretically, it contributes to the evolving notion of “*delegated agency*,” which describes how human actors increasingly outsource decision-making to AI systems (Candrian & Scherer, 2022). These systems, in turn, act with a degree of operational independence that traditional legal doctrines—anchored in anthropocentric notions of intent and culpability—struggle to accommodate. For instance, when an AI autonomously designs a phishing campaign or produces deepfake content to influence financial markets, assigning responsibility becomes challenging. The legal system is confronted with actions that bear all the hallmarks of intentional harm yet lack a clearly identifiable perpetrator.

The study’s findings indicate that the majority of AI-related cyber incidents involve advanced forms of social engineering that elude precise categorization under existing legal definitions such as fraud or identity theft—further underscoring the misalignment between law and technological reality. This challenges the foundational legal assumption that agency—and therefore responsibility—must reside in a human actor. In the treaty, Articles 7 and 13, for example, center on unauthorized access and fraud but implicitly assume a human actor whose intent can be established and linked to identifiable behavior. Instead, this research posits that agency in the digital era is often distributed across human-machine networks and can emerge from the interaction of algorithms, user behavior, and systemic feedback loops.

Therefore, against this backdrop, the study argues that one of the most pressing jurisprudential and regulatory challenges lies in developing a coherent, comprehensive international legal framework that simultaneously protects fundamental rights and fosters responsible AI innovation. Such a framework must move beyond the classical legal constructs of *mens rea* (the mental element of a crime) and *actus reus* (the physical act of committing a crime), which together underpin the attribution of criminal liability (the legal assignment of responsibility) (Mallorquí-Ruscalleda, 2020). These constructs presuppose a human agent who both performs a prohibited act and possesses a culpable state of mind—

assumptions that prove increasingly inadequate in the context of autonomous or semi-autonomous AI systems. Therefore, we recommend that international frameworks such as the UN Cybercrime Treaty be augmented with AI-specific annexes or interpretive guidance. These additions could clarify how traditional legal concepts should be applied to AI-driven actions without sacrificing legal certainty or procedural fairness. Equally important is need for shared standards of AI forensic analysis (methods for investigating AI actions) and cross-border data protocols, essential for supporting mutual legal assistance and harmonizing evidence-handling across jurisdictions.

Practical Implications and Recommendations: The implications of the reframing mentioned above are especially urgent for companies operating within interconnected digital and global ecosystems. This research highlights the inadequacy of fragmented, localized legal responses in an environment where the majority of AI-related cyber incidents cross national borders. This necessitates a move away from rigid compliance models toward adaptive governance approaches (flexible and responsive strategies that evolve alongside changing technologies and threats) that reflect the complexity of the dynamic threat landscape.

Organizations must not only invest in technological safeguards, but also build legal and ethical capacities to navigate new forms of risk. For the private sector, this entails confronting a shifting liability terrain—one in which attribution and accountability may hinge on an organization's ability to demonstrate meaningful oversight, transparency, and ethical deployment of AI. It also calls for active engagement in shaping governance frameworks that can trace the origins and consequences of AI-driven actions, even in the absence of direct human intent. Finally, regular review mechanisms, such as AI cybersecurity task forces composed of rotating experts from member states and technical fields, could be implemented to ensure the treaty remains responsive to rapidly evolving AI-driven cyber threats.

Grounded in the theoretical and practical challenges discussed, the next section examines methodological limitations and outlines avenues for future research.

Limitations and Future Research

Our methodological approach to investigating AI-related cybercrime incidents highlights the challenges of studying and regulating emergent technological phenomena, particularly within domains characterized by high complexity and limited empirical accessibility. The sampling methodology intentionally addresses the methodological limitations of comprehensive technological threat assessment. By prioritizing accessible and verifiable incident data, the research acknowledges the inherent methodological constraints imposed by sensitive technological domains—notably military and state surveillance contexts—which present significant barriers to systematic documentation and empirical investigation.

Moreover, the temporal scope of the research—though limited to a defined period—is deliberately framed as a strategic lens for analyzing technological transformation. This delimitation enables a nuanced understanding of how technological threats evolve, recognizing the dynamic and nonlinear nature of innovative ecosystems and their susceptibility to adaptation.Building on these insights, this methodological approach underscores the importance of crafting adaptive and iterative regulatory frameworks that can effectively address the inherent uncertainties and evolving nature of technological innovation. A central recommendation is to pursue future research in several key areas:

- Empirical investigations into treaty implementation across varied legal and cultural contexts
- Comparative studies of national regulatory strategies for AI and cybersecurity

- The development of quantifiable metrics, such as reduction in cross-border AI-enabled cybercrimes, increased cross-border cooperation, and improved detection rates of AI-generated threats, could provide concrete measures for assessing the treaty's long-term impact.

Within this context, the research frames the UN Cybercrime Treaty as both a conceptual and normative vehicle for adaptive governance—moving beyond the constraints of traditional, static legislative models. Considering the outlined limitations, the following conclusion integrates key insights and highlights the critical need for adaptable, cross-jurisdictional legal frameworks to address AI-related cybercrime.

Conclusion

While the treaty provides a foundation for addressing the challenges presented in this paper, our analysis identifies specific areas where AI-related cybercrime strains the boundaries of legal frameworks through its unprecedented complexity and autonomous capabilities. In particular, the research reveals three critical dimensions where traditional legal conceptualizations fundamentally break down:

- Firstly, the ability of AI systems to operate without human intervention undermines traditional legal frameworks. Unlike conventional cybercrimes perpetrated by identifiable individuals, AI-generated threats often result from complex processes, making it challenging to determine purpose and attribute responsibility. Some regulations, such as the European Union's Artificial Intelligence Act (AI Act), may help address the treaty's shortcomings in this regard (European Parliament & Council of the European Union, 2024). For example, the AI Act introduces a risk-based classification system, categorizing AI applications by their potential harm. High-risk applications are subject to stringent requirements, including transparency, human oversight, and accountability measures. This approach acknowledges the challenges in attributing intent and seeks to mitigate risks through proactive governance.

- Secondly, the transnational nature of AI-enabled cyber threats exposes significant gaps in current international legal frameworks. Our research indicates that a substantial proportion of such incidents span multiple jurisdictions, challenging existing treaty provisions to effectively coordinate investigation, attribution, and prosecution efforts across diverse legal systems. The Council of Europe's Framework Convention on Artificial Intelligence addresses some of these challenges by promoting a harmonized approach to AI governance, emphasizing human rights, democracy, and the rule of law (Council of Europe, 2024). It mandates that parties conduct risk and impact assessments throughout the AI system lifecycle, fostering international cooperation and consistent regulatory standards. The potential alignment between the UN Cybercrime Treaty and the Framework Convention on Artificial Intelligence could mark progress toward a more cohesive international legal framework for addressing the distinct challenges AI presents in the context of cybercrime.

- Thirdly, the rapid evolution of AI technologies further exacerbates the aforementioned challenges. Legal frameworks, often designed around static technological paradigms, struggle to anticipate and adapt to the emergent capabilities of intelligent systems. Because AI is evolving quickly, governments and experts from different fields need to work together in flexible ways to manage its risks. For instance, Singapore's Infocomm Media Development Authority (IMDA) has introduced the Generative AI Evaluation Sandbox—a collaborative initiative that brings together global stakeholders to develop standardized evaluation benchmarks for large language models (IMDA, 2023). This sandbox facilitates the testing and validation of AI systems in a controlled environment, promoting responsible innovation while ensuring compliance with evolving regulatory standards.

Therefore, as AI continues to advance, it is essential for policymakers, legal experts, and cybersecurity professionals to collaborate in refining and strengthening regulatory frameworks that can effectively address AI-driven cyber threats while also supporting responsible innovation. To achieve this, particular attention must be given to the development of harmonized, cross-border legal instruments and procedural cooperation mechanisms. Such tools are essential to overcoming the jurisdictional fragmentation that continues to impede effective enforcement efforts in cases of AI-facilitated cybercrime. Without such alignment and institutional adaptability, regulatory systems risk remaining reactive, fragmented, and ultimately unprepared to meet the challenges posed by AI-related cybercrime.

References

Artificial Intelligence Incident Database. (n.d.). *Artificial intelligence incident database*. Retrieved May 26, 2025, from https://incidentdatabase.ai/

Baran, G. (2025). Hackers exploit Gamma AI to create sophisticated phishing redirectors. *Cybersecuritynews*. https://cybersecuritynews.com/hackers-exploit-gamma-ai/

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Buçaj, E., & Idrizaj, K. (2025). The need for cybercrime regulation on a global scale by the international law and cyber convention. *Multidisciplinary Reviews*, 8(1), 2025024. https://doi.org/10.31893/MULTIREV.2025024

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

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, 134, 107308. https://doi.org/10.1016/j.chb.2022.107308

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': The US, EU, and UK approach. *Science and Engineering Ethics*, 24(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

Council of Europe. (2024). *Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law*. https://edoc.coe.int/en/artificial-intelligence/11926-council-of-europe-framework-convention-on-artificial-intelligence-and-human-rights-democracy-and-the-rule-of-law.html

Creswell, J. W. (2017). *Qualitative inquiry & research design: Choosing among five approaches (2nd ed.).* SAGE Publications. https://www.researchgate.net/publication/342229325

Creswell, J. W., Hanson, W. E., Clark Plano, V. L., & Morales, A. (2007). Qualitative research designs: Selection and implementation. *The Counseling Psychologist*, 35(2), 236–264. https://doi.org/10.1177/0011000006287390

Creswell, J. W., & Poth, C. N. (2017). *Qualitative inquiry & research design: Choosing among five approaches* (4th ed.). SAGE Publications. https://revistapsicologia.org/public/formato/cuali2.pdf

European Parliament, & Council of the European Union. (2024). *Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU)*

2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The emerging threat of AI-driven cyber attacks: A review. *Applied Artificial Intelligence*, 36(1), 2037254. https://doi.org/10.1080/08839514.2022.2037254

Infocomm Media Development Authority. (2023). *Generative AI evaluation sandbox*. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics*, 26(1), 89–120. https://doi.org/10.1007/s11948-018-00081-0

Kshetri, N. (2013). *Cybercrime and cybersecurity in the global south*. Palgrave Macmillan. https://doi.org/10.1057/9781137021946

Mallorquí-Ruscalleda, E. (2020). The elements of a crime: A brief study on actus reus and mens rea. *University of São Paulo & University of Porto, School of Law*. https://hdl.handle.net/1805/24549

MIT FutureTech. (n.d.). *The MIT AI risk repository*. Retrieved May 26, 2025, from https://airisk.mit.edu/

Reuters. (2024, September 3). Clearview AI fined by Dutch agency for facial recognition database. *Reuters*. https://www.reuters.com/technology/artificial-intelligence/clearview-ai-fined-by-dutch-agency-facial-recognition-database-2024-09-03/

Resecurity. (2024). Cybercriminals are targeting AI agents and conversational platforms: Emerging risks for businesses and consumers. *Resecurity*. https://www.resecurity.com/blog/article/cybercriminals-are-targeting-ai-agents-and-conversational-platforms-emerging-risks-for-businesses-and-consumers

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2609777

Serviceaide. (2025). Notice of data security event May 5, 2025. https://www.serviceaide.com/notices

United Nations General Assembly. (2024). United Nations convention against cybercrime (A/RES/79/243). *United Nations*. https://docs.un.org/en/A/RES/79/243

Weber, R. H., & Studer, E. (2016). Cybersecurity in the Internet of Things: Legal aspects. *Computer Law & Security Review*, 32(5), 715–728. https://doi.org/10.1016/j.clsr.2016.07.002

Wicki-Birchler, D. (2020). The Budapest Convention and the General Data Protection Regulation: Acting in concert to curb cybercrime? *International Cybersecurity Law Review*, 1(1–2), 63–72. https://doi.org/10.1365/s43439-020-00012-5