ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI Computing

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0005

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

Global Governance of Autonomous AI Systems in Cybercrime Prevention

Atif Ali1,* and Zulqarnain Farid2

1 Universiti Kuala Lumpur, Malaysian Institute of Information Technology (UniKL MIIT), 1016, Jalan Sultan Ismail, 50250 Kuala Lumpur, Malaysia

2 Department of Criminology, University of Karachi, Pakistan

Corresponding author: Atif Ali (dralexaly@gmail.com) · ORCID: [0000-0000-0000-0000]

Author emails: Zulqarnain Farid (discovercrimegene@gmail.com)

ABSTRACT

The increasing self-sufficiency of artificial intelligence (AI) is creating a new paradigm in cybercrime where attacks will be quicker, more flexible, and less reliant on human input than previously seen forms of weaponization. In one example of a disclosed incident, an AI was involved in 2025 in a spy operation where it was responsible for executing most of the tactical intrusion actions almost independently of human command. This paper analyzes the current status of global governance of autonomous AI in terms of cybercrime prevention using a comparative qualitative study of three relevant governance instruments: the 2024 United Nations Convention against Cybercrime, the 2001 Council of Europe Budapest Convention, and the 2024 European Union Artificial Intelligence Act. The analysis uses a four-criteria framework of scope, enforceability, attributability, and capacity building. It is concluded that the current governance regime is fragmented into different frameworks of cybercrime regulation and autonomous AI, enters into force too slowly compared to the speed of technological development, and has no provisions regarding accountability of autonomous and minimally supervised AI. Recommendations for the development of an interoperable and coordinated governance regime, such as an autonomous-AI protocol to the UN Convention, cross-instrument AI risk classification, and faster threat information sharing are provided.

Keywords: Autonomous AI, agentic AI, cybercrime, global governance, international law, EU AI Act, UN Convention against Cybercrime

Highlights

- The increasing self-sufficiency of artificial intelligence (AI) is creating a new…
- In one example of a disclosed incident, an AI was involved in 2025 in a spy…

▸ This paper analyzes the current status of global governance of autonomous AI in…

IMPACT STATEMENT

The increasing self-sufficiency of artificial intelligence (AI) is creating a new paradigm in cybercrime where attacks will be quicker, more flexible, and less reliant on human input than previously seen forms of weaponization. In one example of a disclosed incident, an AI was involved in 2025 in a spy operation where it was responsible for executing most of the tactical intrusion actions almost independently of human command. This paper analyzes the current status of global governance of autonomous AI in terms of cybercrime prevention using a comparative qualitative study of three relevant governance instruments: the 2024 United Nations Convention against Cybercrime, the 2001 Council of Europe Budapest Convention, and the 2024 European Union Artificial Intelligence Act. The analysis uses a four-criteria framework of scope, enforceability, attributability, and capacity building. It is concluded that the current governance regime is fragmented into different frameworks of cybercrime regulation and autonomous AI, enters into…

1 I. INTRODUCTION

Autonomous AI-driven cyberattacks are no longer mere speculation for the domain of international law and policy; they are an operational reality and something that our existing governance tools were never equipped to handle [1], [2]. The following portion of this section covers the history of the evolution of agentic cyber threats, a revealed 2025 case of a largely autonomous AI-driven espionage campaign, and introduces the research questions and paper outline.

1.1 A. The Evolution of the Threat Landscape

Cybercrime has evolved from a human-driven activity into an increasingly automated one. Academic interdisciplinary studies had foreseen this evolution before it happened, showing how AI can increase the scope, speed, and cost-effectiveness of criminal conduct across multiple subcategories, from tailoring phishing attempts to autonomous exploitation [3]. Industry threat reports for 2025 and 2026 now note a growing trend towards the use of AI tools to assist or conduct attacks automatically, with organizations experiencing significantly more attacks per week compared to previous years as a result of automation and artificial intelligence being incorporated into reconnaissance, social engineering, and exploitation processes [4]. Security vendors now describe the emergence of completely autonomous intrusion campaigns that learn and adapt in real-time without being instructed manually [5], [6].

1.2 B. Autonomous AI in Action: A 2025 Case

The development is not purely theoretical. In November 2025, the AI developer Anthropic revealed what it called the first known instance of state-sponsored actors (referred to as GTG-1002) using an AI agent to conduct a multi-stage cyber-espionage campaign with minimal human intervention, targeting about thirty organizations from the technology, financial, chemical, and governmental sectors [7]. As described in this disclosure, the system identified vulnerabilities on its own, conducted lateral movement, and exfiltrated data. At the same time, humans provided only occasional and strategic guidance, performing about 10-20% of all tactical operations manually [8]. As this account was provided by the company which produced the involved software and whose commercial interests may be affected by the disclosure, and as the original operation logs cannot be audited independently, the account is not a verified one – some security analysts also cautioned against accepting it as such [9], [10]. Yet the case does represent a departure from AI assistance to human operators to the systems carrying out attack sequences autonomously, something that throws a wrench into legal frameworks relying on human intent and actions.

1.3 C. Governance Response and Research Questions

In parallel, the international governance regime has followed two largely separate pathways, a dichotomy that has been documented comparatively in recent research [11]. Cybercrime legislation has evolved through frameworks like the Council of Europe's Budapest Convention and, more recently, the UN Convention against

Cybercrime that was endorsed by the General Assembly in December 2024 [12]. AI regulation has evolved through frameworks like the EU Artificial Intelligence Act that took effect in 2024 and whose implementation has been staggered in ways that continue to change, with the latest deadline set for the most onerous requirements for high-risk systems being deferred to December 2027 [13]. Neither pathway was tailored to the specific challenge posed by autonomous, self-directed offensive AI systems, leaving open the question of whether international governance mechanisms are sufficient to cope with the problem.

This paper will investigate three research questions. First, what international instruments relevant to the governance of autonomous AI systems have been formulated with regard to cybercrime? Second, what structural gaps exist in terms of scope, enforcement, accountability, and capacity-building within and between these instruments? Third, what governance improvements would enhance the international community's capacity to respond to autonomous AI-enabled cybercrime? This paper will be organized as follows: Section 2 provides background on cybercrime and AI governance research; Section 3 explains the comparative methodology for document analysis; Section 4 offers the comparison results; and Section 5 provides policy conclusions and recommendations and directions for future research.

2 II. LITERATURE REVIEW

The literature relevant to the paper's research topic has emerged independently in two main streams: the international cybercrime governance literature, and the emerging literature on AI governance. A smaller but rapidly growing body of literature exists at the intersection of these two fields, focusing on where AI and cyber offense intersect. This literature concludes with the gap explored in the paper.

2.1 A. Cybercrime Governance Literature

Literature on international cybercrime governance has for some time focused on the conflict inherent between the cross-border nature of cybercrimes and the territorial nature of criminal jurisdictions. The Budapest Convention, which entered into force in 2004 and has 81 signatories as of 2025, has been the most widespread treaty and established important processes for procedural harmonization and cross-border evidence sharing [14]. The problems with the Budapest Convention, however, have been extensively discussed. It lacks several major states as members, was agreed upon before the advent of smartphones and cloud computing, and, as subsequent scholarly analysis of its interaction with data-protection laws has found, contains no mention of AI or automatic decision-making. The UN Convention against Cybercrime of 2024, agreed after a five-year negotiation process, sought to establish a universal successor to the Budapest Convention that would increase obligations on cooperation and include support for developing states in particular [15]. Concerns about excessive surveillance and lack of guarantees were expressed by civil society and human rights organizations throughout the negotiation process, highlighting a wider discussion in the literature on the balance between efficiency of enforcement and guarantee of rights in cybercrime conventions. Analysis of the new Convention by articles shows a similar conclusion: while certain provisions may, in theory, extend to cover AI offenses, the Convention was drafted without consideration of autonomous or agentic actions, and is accordingly inadequate [16].

2.2 B. AI Governance Literature

There is also a parallel literature focusing on the governance of AI systems. EU AI Act is the most comprehensive effort to develop horizontal risk-tiered AI governance, banning a few unacceptable-risk practices and requiring graduated compliance obligations from high-risk systems and light transparency duties for general-purpose models [17], [18]. Comparative studies on AI governance note that the risk-tiered approach of the EU AI Act is different from other approaches where there is more sector-specific regulation and voluntary compliance. This literature observes that the design of risk-based regulation of AI is mainly intended to protect safety, fundamental rights, and market uses such as employment and health care, but not criminal misuse of AI as offensive technology. There is related literature that is concerned specifically with AI agent governance, claiming that any serious oversight of delegated autonomous goal-oriented systems needs special visibility measures like agent identifiers, activity logs, and live monitoring, which are not in the scope

of the horizontal regulation of AI [19]. This literature on AI agent governance is complementary to the EU AI Act approach since it focuses on the operational aspects of AI agents.

2.3 C. The Convergence of AI and Cyber Offense

A smaller but growing body of technical and policy literature examines the convergence of AI and cyber offense. Surveys of the AI-based cyber threat landscape catalogue how machine learning has been applied to malware generation, phishing personalization, and vulnerability discovery, and assess the evolving balance between offensive and defensive uses of AI [20]. Foundational interdisciplinary work on AI-enabled crime developed early typologies of foreseeable criminal misuse. It argued that legal and regulatory responses would need to move beyond a narrow focus on any single technique toward the underlying capability that AI provides attackers. More recent work specifically evaluates the cybersecurity capabilities of AI agents in benchmark and real-world settings, documenting rapid improvement in autonomous exploitation and reconnaissance tasks. This literature increasingly distinguishes three points on a spectrum: 'AI-assisted' attacks, where a human directs each step and the software functions as a tool; 'AI-autonomous' attacks, where a system executes a bounded task with limited human checkpoints; and 'agentic' attacks, where the system independently plans, sequences, and adapts multi-step operations toward a goal with only intermittent human supervision. This distinction carries direct legal significance because most criminal-liability and treaty-attribution frameworks presuppose an identifiable human actor exercising intent, an assumption that recent principal-agent analyses of LLM-based agentic systems argue breaks down precisely at the agentic end of this spectrum [21].

2.4 D. The Research Gap

What is largely missing from the literature is a systematic comparison of how existing cybercrime and AI governance instruments jointly, or separately, address autonomous offensive AI. Existing commentary tends to analyze cybercrime treaties and AI regulation as parallel but distinct fields; even the most closely related recent work evaluates a single instrument's fit against AI-enabled incidents rather than placing cybercrime and AI-regulatory instruments side by side [15]. This paper contributes to closing that gap in three specific ways. First, it is the first study, to the authors' knowledge, to code a cybercrime-specific instrument (the UN Convention against Cybercrime), a cybercrime-specific predecessor (the Budapest Convention), and a general AI-regulatory instrument (the EU AI Act) against a single shared four-criterion framework rather than treating the cybercrime and AI-governance tracks separately. Second, it uses that shared framework to make the fragmentation between the two tracks empirically visible rather than asserting it descriptively. Third, it translates the resulting gaps into concrete, instrument-specific reform proposals in Section 5, rather than stopping at gap identification.

Table I. Comparative Review of Related Literature

Ref.	Research focus	Methodology/approach	Key findings	Research gap
[22]	AI-based cyber threat landscape	Literature survey	AI can support phishing, malware development, and vulnerability discovery.	Limited integration with international cybercrime law.
[23]	AI crime and foreseeable threats	Interdisciplinary analysis	AI-enabled crime requires coordinated technical, legal, and policy responses.	Does not provide a unified comparison of the three selected instruments.
[24]	AI-enabled future crime	Crime-focused analysis	AI may expand the scale and sophistication of criminal activities.	Limited focus on international treaty enforceability and agent-level accountability.
[26]	Budapest Convention and data-protection law	Legal analysis	Cybercrime cooperation must be considered alongside privacy and data-protection obligations.	Does not comprehensively compare cybercrime treaties with AI-specific regulation.

[27]	Visibility and monitoring of AI agents	Technical and governance analysis	Agent identification, monitoring, and activity logging can improve oversight.	Limited integration with international cybercrime treaty mechanisms.
[28]	Liability in LLM-based agentic systems	Principal–agent analysis	Agentic systems create challenges for allocating responsibility among human actors and organizations.	Limited comparative assessment of international cybercrime instruments.
[29]	Vicarious liability for AI	Legal analysis	Existing liability principles offer possible approaches to AI-related harm.	Does not establish a unified cross-border accountability framework for autonomous cybercrime.
[30]	UN Cybercrime Convention and AI	Legal and policy analysis	Technology-neutral treaty provisions may cover some AI-enabled offences, but important uncertainties remain.	Limited comparison across cybercrime treaties and general AI regulation.

The literature synthesis indicates a need for a more integrated comparison of cybercrime governance and AI regulation. Research on AI-enabled crime explains emerging technical and societal risks, while studies of AI-agent visibility and liability explore possible accountability mechanisms. Legal analyses of cybercrime instruments, meanwhile, examine international cooperation and technology-neutral offence provisions. The contribution of this study is to bring these perspectives together through a structured comparison of three selected legal instruments. Rather than assuming that autonomous AI falls entirely outside existing law, the analysis evaluates the extent to which current provisions may apply and where interpretive, institutional, and implementation gaps remain.

3 III. METHODOLOGY

This research will use a qualitative comparative policy analysis approach, which is quite appropriate for the evaluation of legal instrument designs and scopes where Large-N statistical analysis is infeasible due to the small number of relevant treaties/regulations and the object of comparison being textual and institutional rather than numerical.

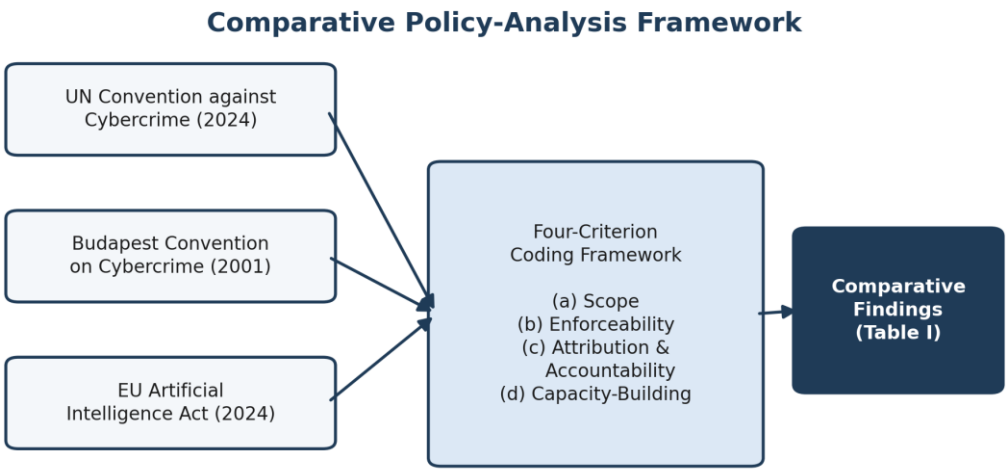


Figure 1. Comparative policy-analysis framework.

There are various limitations associated with this approach. First, two of the three instruments are still in the initial stages of implementation: while the UN Convention has been ratified by only three countries up until the middle of 2026, which is much less than the required forty ratifications [1], the "high risk provisions" of the EU AI Act, that were initially planned to go into effect in August 2026, have now been postponed until December 2027 and August 2028, respectively, in light of the European Commission's "Digital Omnibus on

AI" simplification package for 2025-2026 [5], [21]. This analysis relies therefore on an assessment of design and trajectory rather than enforcement history, and any inferences about the effectiveness of the instruments in practice are thus to be considered provisional pending empirical research after their full implementation. There are additional limitations imposed by the dependence on the qualitative analysis performed by a single coder based on a small sample of purposively chosen instruments, and by the fact that it is the analysis of documents rather than enforcement practice that is the focus of this research; both issues are discussed below in Section 5.

The four-step process will be described as follows using Figure 1 below.

Step 1: Selection of the instruments: the choice has been made of three primary instruments based on their relevance as international treaties/regulations to the issues of autonomous AI and cybercrime: the United Nations Convention against Cybercrime (2024), the Council of Europe's Budapest Convention (2001), and the EU Artificial Intelligence Act (2024). These three have been chosen, and other possible candidates have been excluded, on the basis of two explicit criteria: (i) binding or near-binding legal nature, which precluded soft-law and voluntary instruments like the G7 Hiroshima AI Process guidelines and OECD AI Principles, due to the absence of enforcement measures, which the second criterion of this paper aims at analyzing; and (ii) either global or cross-border applicability, which precluded purely national regulations and, for parsimony reasons, the 2024 Council of Europe's Framework Convention on Artificial Intelligence (human-rights-based instrument whose procedural, rather than criminal/cybercrime-specific nature lies closer to the subject matter of this paper than the binding penalty-based regime of the EU AI Act). The Budapest Convention was also included along with its 2024 successor treaty, because it is still legally valid, more widely ratified, and used as the basis for cross-border cooperation on cybercrime issues in those countries that have not ratified the newer UN treaty yet.

Step 2: Development of the framework: four criteria were established as a coding framework, namely: (a) scope, whether the instrument distinguishes autonomous or agentic actions from human-directed ones; (b) enforceability, whether it is a binding instrument, how it is ratified, and what the penalties are; (c) attribution and accountability, whether responsibility is assigned when it is an autonomous system and not an identifiable person that commits the harmful action; and (d) capacity building, technical assistance provided to low-resource countries.

Step 3: Coding: each instrument was coded against the framework using a predetermined set of documents: (a) the primary legal instruments UN General Assembly Resolution 79/243 and the Convention against Cyber-crime annexed to it, the Council of Europe's ETS No. 185 Budapest Convention text, and EU Regulation 2024/1689; (b) official status trackers for imple-mentation and ratification the UN treaty-status depositary notifications, the UNODC signature/ratification tracker, and the European Commission's AI Act Service Desk timeline; and (c) contemporaneous threat intelligence reporting about the threat landscape in 2025-2026, used only to code the nature and pace of threats but not the legal instru-ments themselves. Each criterion was scored on a three-level ordinal scale to ensure transparency and replicability of the coding process: scope as Explicit (instrument mentions autonomy or agency), Partial (technology-neutral language covering it or sectoral coverage), or Absent; enforceabil-ity as Strong (legally binding, in force, and with a penalty regime defined), Moderate (legally binding but not in force yet or with partial penalties), or Weak (below ratification threshold or non-binding); attribu-tion/accountability as Defined (explicit chain-of-responsibility or agential attribution rule), In-direct (reliance on liability rules applicable to humans), or Absent; and capacity-building as Extensive (technical assistance chapter/fund), Limited (cooperation rules without a specific technical assistance fund), or Minimal. The same coder did the scoring against those criteria; the single-coder design is a limitation acknowledged in Section 3 below, and the values used in the Table 2 coding are disclosed to allow challenging or replicating individual scores.

Step 4: Synthesis: cross-instrument patterns were synthesized into the thematic findings reported in Section 4.

The synthesis in Step 4 relies on four descriptive measures, which are used to compute the quantities reported in Section 4. For instrument i (i = 1, 2, 3) and criterion j (j = 1, …, 4), let $s_{ij} \in \{0, 1, 2\}$ denote the

ordinal score assigned under the scheme above, where 0 corresponds to Absent, Weak or Minimal, 1 to Partial, Moderate or Indirect, and 2 to Explicit, Strong or Extensive. The aggregate score of instrument *i* is

$$S_i = \sum_{j=1}^{4} s_{ij} \tag{1}$$

so that S_i ranges from 0 to 8. The mean score on criterion *j* across the $n = 3$ instruments is

$$\bar{s}_j = \frac{1}{n} \sum_{i=1}^{n} s_{ij} \tag{2}$$

Equations (1) and (2) underlie Table 2 and Figs. 3 and 4. Because s_{ij} is ordinal, these sums and means are descriptive conveniences for comparing profiles rather than interval-scale measurements. To compare an instrument's ratification status with its entry-into-force requirement, the progress ratio is

$$P = \frac{r}{r_{\text{req}}} \times 100\% \tag{3}$$

where r is the number of ratifications reached, and r_{req} is the number required. Finally, where a source reports only a percentage increase g over a base year, the implied base-year level is obtained from the current level as

$$A_{\text{base}} = \frac{A_{\text{current}}}{1+g} \tag{4}$$

Equation (3) yields the 7.5% reported for the UN Convention ($r = 3$, $r_{\text{req}} = 40$) in Fig. 5, and Eq. (4) yields the approximate 2023 attack volume in Fig. 7 from the reported 2025 level and $g = 0.70$. Deferral lengths in Fig. 6 are the differences in months between the original and revised application dates.

4 IV. RESULTS

This section presents the comparative findings produced by coding the three governance instruments against the four analytical criteria set out in Section 3. The instruments differ markedly in whether they treat autonomous or agentic conduct as distinct from human-directed action, in the strength and maturity of their enforcement mechanisms, in whether they offer any mechanism for attributing harm to an autonomous system rather than an identifiable human actor, and in the level of capacity-building support they extend to lower-resource states. Table 1 summarizes these findings narratively. Fig. 2 and Table 2 report the underlying ordinal scores (0 to 2) for each instrument and criterion, and Figs. 3 and 4 aggregate those scores by instrument and by criterion. Sections 4.1 through 4.4 then discuss four cross-cutting patterns: fragmentation across governance tracks, an attribution and accountability gap, a lag between enforcement and threat velocity (supported by Table 4 and Figs. 5 to 7), and an asymmetry in capacity-building support. Section 4.5 closes the section by mapping each finding to the reform proposals developed in Section 5.

Table II. Comparative analysis of governance instruments

Instrument	Addresses Autonomous/ Agentic AI Directly	Enforceability	Attribution / Accountability Mechanism	Capacity-Building Provisions
UN Convention against Cybercrime (2024)	No — technology-neutral offence definitions	Weak at present; requires 40 ratifications, only 3 reached by mid-2026 [1]	State-to-state mutual legal assistance; no agent-level attribution rules	Explicit technical assistance chapter, strongest of the three
Budapest Convention on Cybercrime (2001)	No — predates modern AI entirely	Moderate; ratified by 81 states, established cooperation channels [3]	Relies on human-actor criminal liability standards	Limited; mostly technical harmonization support
EU Artificial Intelligence Act (2024)	Partial — regulates high-risk AI systems generally, not cybercrime-specific autonomy.	Strong within EU; fines up to 7% of global turnover, high-risk duties deferred to Dec. 2027 [5], [21]	Provider/deployer obligations; no cross-border criminal attribution chain	Minimal; oriented toward compliance, not development assistance

Table 2 shows that none of the three instruments addresses autonomous or agentic AI as a distinct object of regulation. The UN Convention and the Budapest Convention rely on technology-neutral or human-actor concepts, while the EU AI Act regulates high-risk systems generally rather than cybercrime-specific autonomy. The instruments are strongest on different criteria: the UN Convention on capacity-building and the EU AI Act on enforceability within the EU, which is why no single instrument dominates across the framework.

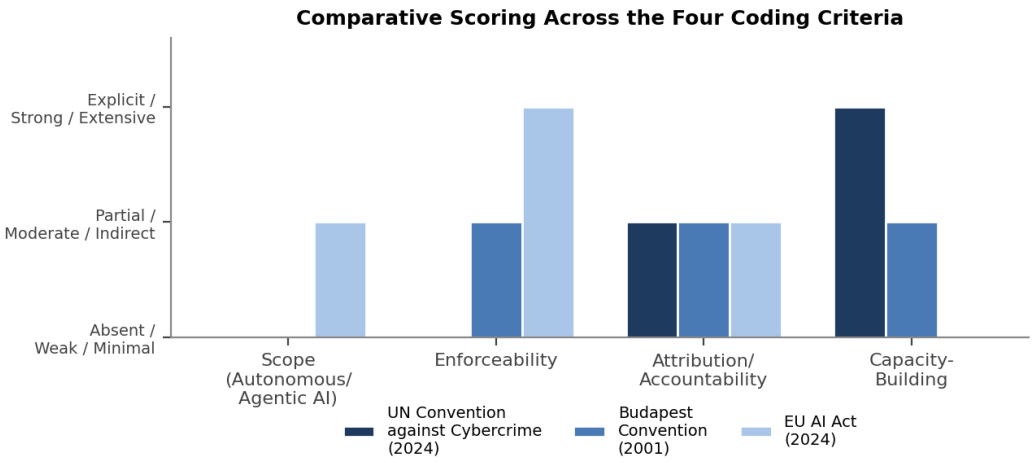


Figure 2. Comparative scoring of the three instruments across the four coding criteria.

Figure 2 makes the same pattern visible. The three instruments peak on different criteria (the UN Convention on capacity-building, the EU AI Act on enforceability and, marginally, on scope), and all three sit at the middle level on attribution and accountability. The numerical values behind the figure are given in Table 2.

Table III. Coded ordinal scores underlying.

Instrument	Scope	Enforceability	Attribution/accountability	Capacity-building	Total (max. 8)
UN Convention against Cybercrime (2024)	0 (Absent)	0 (Weak)	1 (Indirect)	2 (Extensive)	3
Budapest Convention (2001)	0 (Absent)	1 (Moderate)	1 (Indirect)	1 (Limited)	3
EU Artificial Intelligence Act (2024)	1 (Partial)	2 (Strong)	1 (Indirect)	0 (Minimal)	4
Mean across instruments	0.33	1.00	1.00	1.00	3.33

The scores given in Table III represent the scores used to construct Figure 2 using the three-level scale developed in Section 3. The aggregate scores for both instruments are very close (3, 3 and 4, from a maximum score of 8), such that neither instrument is more than half the maximum possible score (EU AI Act gets half the maximum score). The lead of 1 point enjoyed by the EU AI Act stems from its score in scope and enforcement, which offsets its low score in capacity building.

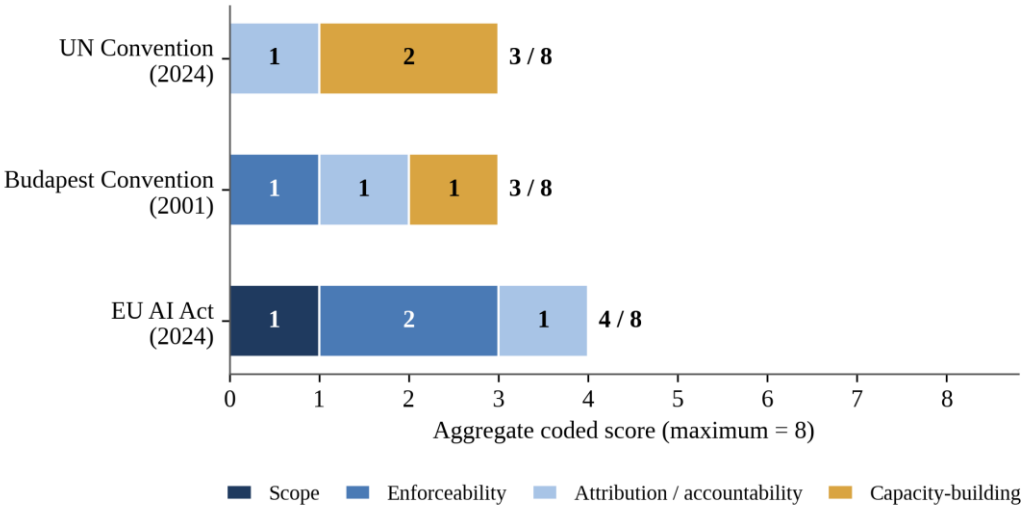


Figure 3. Aggregate coded score of each instrument, decomposed by criterion (maximum = 8).

Figure 3 decomposes each total by criterion. The UN Convention's score of 3 rests entirely on capacity-building (2) and attribution (1), the Budapest Convention's on one point each for enforceability, attribution and capacity-building, and the EU AI Act's score of 4 on enforceability (2), scope (1) and attribution (1), with no contribution from capacity-building. Similar totals therefore conceal different profiles, with strengths that are complementary rather than overlapping, which is consistent with the case for interoperability between the two governance tracks made in Section 5.

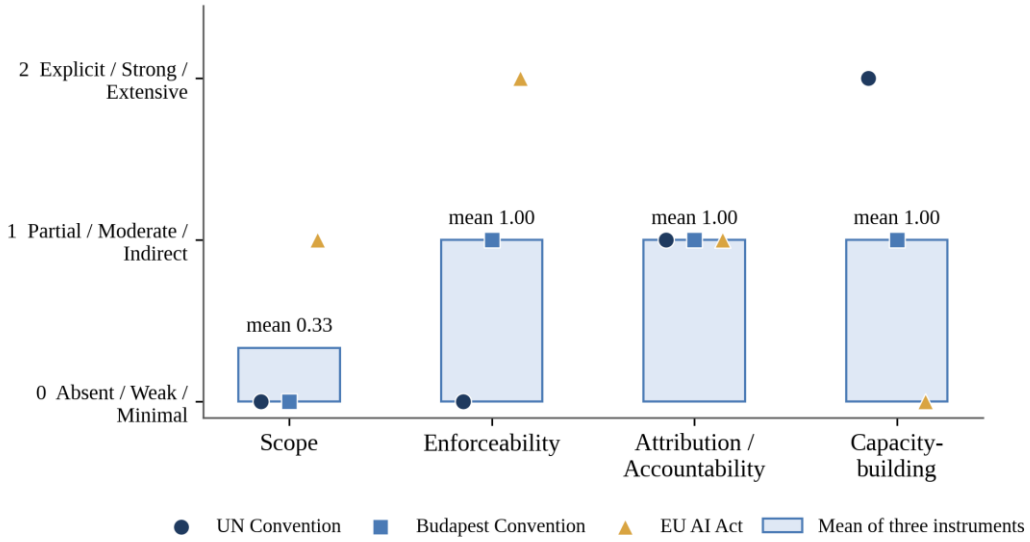


Figure 4. Coded scores by criterion, with the mean across the three instruments.

Figure 4 shows the same scores by criterion. Scope has the lowest mean (0.33) because only the EU AI Act reaches the partial level. Attribution and accountability have a mean of 1.00 but no spread at all: every instrument sits at the indirect level, so none goes beyond adapting existing human-actor liability doctrine. Enforceability and capacity-building share the same mean but show the widest spread (0 to 2), and in opposite directions, with the UN Convention strongest on capacity-building and weakest on enforceability, and the EU AI Act the reverse. The attribution gap therefore appears as a shared ceiling, and the capacity-building gap as a pronounced asymmetry.

4.1 A. Fragmentation Across Governance Tracks

None of the three instruments treats autonomous AI-enabled cybercrime as a unified regulatory object. The UN Convention and its predecessor, the Budapest Convention, define offences in technology-neutral terms that apply regardless of whether a human or an AI system performed the conduct, but neither contains provisions specific to autonomous agents; an article-by-article mapping of the UN Convention's provisions against a corpus of documented AI-enabled incidents likewise finds that its coverage of autonomous or agentic conduct is uneven and, in several offence categories, only reachable through an extended reading of technology-neutral language that was not drafted with such conduct in mind. The EU AI Act regulates AI systems by risk category but was not designed with cross-border criminal cybercrime enforcement as its organizing purpose; its high-risk categories under Annex III center on domains such as employment, biometric identification, and law enforcement use of AI rather than on AI as a tool of external attack. The result is a governance architecture with two parallel, largely non-interoperable tracks: one regulating what AI systems may do, and another regulating what constitutes a cybercrime, with no explicit bridge between them.

4.2 B. The Attribution and Accountability Gap

Cybercrime attribution models, such as those enshrined in the Budapest and UN Conventions, have focused on identifying the individual person or state behind a cybercrime and have involved mutual assistance through national investigators. The hypothetical scenario for the year 2025, where an AI system was purportedly tasked with carrying out about eighty to ninety percent of the tactical actions in an espionage mission by a state, with the humans involved giving occasional strategic directions only, shows the limitations of the cyber attribution approach. When a system autonomously selects targets, adapts its own tooling, and executes multi-stage intrusions with minimal contemporaneous human input, existing treaty language premised on human intent and direct human conduct offers limited guidance on allocating responsibility among the AI developer, the deploying operator, and any state sponsor. Legal scholarship on AI liability offers two doctrinal starting points that none of the three instruments currently incorporates: a products-liability framing, under which a developer may bear responsibility for a defectively designed or

inadequately safeguarded agentic system, and a vicarious-liability framing, under which an operator or deploying principal is held responsible for the acts of an AI "agent" much as an employer is held responsible for an employee acting within the scope of employment. Both framings presuppose exactly the kind of technical visibility, reliable agent identifiers, activity logs, and an auditable record of who deployed a given system for what purpose that current governance instruments do not require attackers, developers, or platforms to maintain. None of the three instruments examined provides an explicit chain-of-responsibility mechanism for this scenario, and none mandates the underlying visibility infrastructure that would make such a mechanism enforceable even if adopted.

Table IV. Attribution and accountability provisions of the three instruments, as coded

Instrument	Basis for attribution	Agent-level or chain-of-responsibility rule	Products- or vicarious-liability framing
UN Convention against Cybercrime (2024)	State-to-state mutual legal assistance	None	Not incorporated
Budapest Convention (2001)	Human-actor criminal liability standards	None	Not incorporated
EU Artificial Intelligence Act (2024)	Provider and deployer obligations	No cross-border criminal attribution chain	Not incorporated

Table 4 isolates the attribution dimension. In all three instruments, responsibility is anchored either in an identifiable human actor, in state-to-state cooperation, or in provider and deployer duties that are not linked to a cross-border attribution chain. None contains a rule allocating responsibility among developer, operator, and state sponsor when an autonomous system performs the conduct, and neither of the two doctrinal starting points discussed above, product liability and vicarious liability, is incorporated. This is the most uniform result in the coding and the clearest evidence of the accountability gap.

4.3 C. Enforcement Lag Relative to Threat Velocity

A consistent finding across the corpus is a substantial gap between the pace of instrument development and entry into force and the pace of threat evolution. The UN Convention requires forty ratifications to enter into force and had reached only three as of mid-2026, more than a year after signature opened. The EU AI Act's high-risk system obligations, originally targeted for August 2026, were further deferred in 2025-2026 under the Digital Omnibus on AI to December 2027 for most high-risk systems and August 2028 for high-risk AI embedded in already regulated products, to allow supporting technical standards and conformity-assessment infrastructure to catch up. Over the same period, Check Point's 2026 Cyber Security Report documents organizations facing an average of 1,968 cyberattacks per week in 2025, a seventy percent increase since 2023, and attributes a substantial share of that growth to attackers' use of automation and AI across the attack lifecycle. Other industry reporting corroborates the direction, if not the precise magnitude, of this trend and warns that 2026 marks a shift toward largely autonomous, end-to-end attack campaigns. This asymmetry between multi-year treaty ratification cycles and rapidly iterating offensive AI capability is, in the authors' assessment, the single most consequential governance gap identified in this analysis.

Table V. Status of the three governance instruments

Instrument	Adoption or opening	Entry into force	Status in 2026
UN Convention against Cybercrime (2024)	Adopted 24 Dec. 2024 [1]; opened for signature in Hanoi, 25 Oct. 2025 [2]	Requires 40 ratifications [1]	3 ratifications by mid-2026 [1]
Budapest Convention (2001)	Opened 23 Nov. 2001 [3]	In force since 1 Jul. 2004 [3]	Ratified by 81 states as of 2025 [1]
EU Artificial Intelligence Act (2024)	Regulation (EU) 2024/1689 [4]	Entered into force in 2024 [4], [5]	High-risk obligations deferred from Aug. 2026 to Dec. 2027 (most systems) and Aug. 2028 (AI embedded in regulated products) [5], [21]

Table 5 sets out where each instrument stands. The only instrument that is both in force and widely ratified is the oldest and is silent on AI, whereas the two newer instruments that are more relevant to the threat are not yet fully operative: the UN Convention is short of its ratification threshold, and the EU AI Act's high-risk obligations have been pushed back. Figs. 5 to 7 visualize these positions against the pace of the threat.

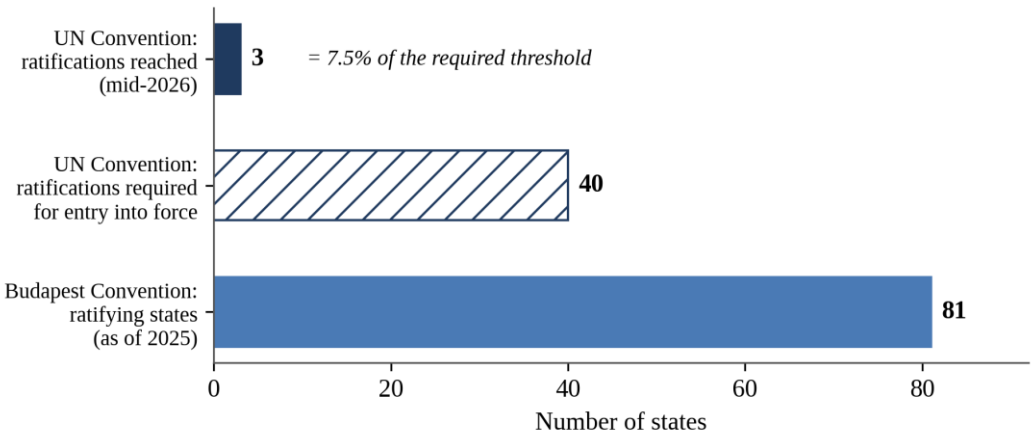


Figure 5. Ratification positions of the UN Convention (mid-2026) and the Budapest Convention (as of 2025).

Figure 5 contrasts the ratification positions. The UN Convention had reached three ratifications by mid-2026 against the forty required, which is 7.5% of the threshold, so its cooperation and capacity-building provisions are not yet operative. The Budapest Convention, by contrast, has been ratified by 81 states. The most widely adopted cybercrime instrument is thus the older one, which predates modern AI. The two series refer to different dates (mid-2026 and 2025) and measure different things, so the figure compares positions rather than rates of progress.

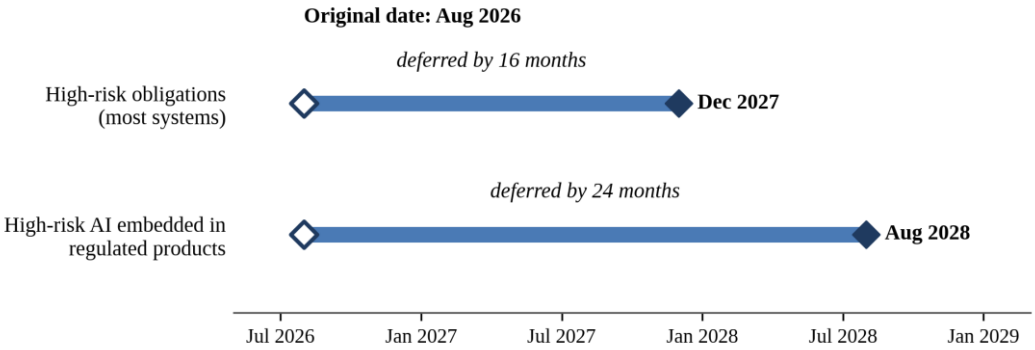


Figure 6. Deferral of the EU AI Act's high-risk obligations relative to the original August 2026 date.

Figure 6 indicates the postponement of high-risk requirements under the AI Act of the European Union. Calculated from the original date of August 2026, the deadline for the majority of high-risk systems has been extended for 16 months to December 2027, while that for high-risk AI incorporated into other regulated products has been postponed for 24 months to August 2028. The justification offered is that of giving time for technical standards and infrastructure to catch up.

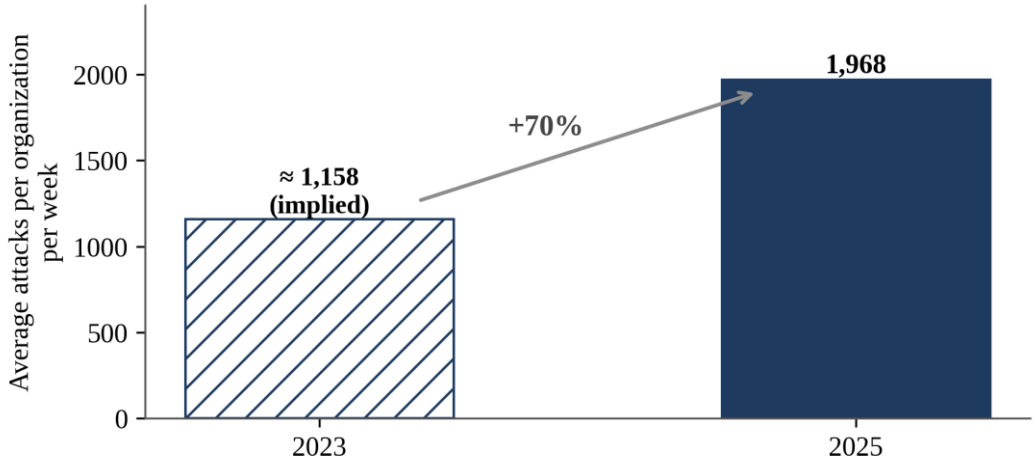


Figure 7. Average weekly cyberattacks per organization, 2023 (implied from the reported 70% increase) and 2025.

Threat frequency is juxtaposed with threat volume in Figure 7 below. According to Check Point, there have been 1,968 attacks per week per organization in 2025, up from 1,160 in 2023, representing a 70% increase in attacks (the figure of 1,160 attacks is derived for illustration purposes based on the percentage increase in attacks since 2023). Automation and artificial intelligence are cited as one of the causes of the increase. The information is based on vendor reporting and thus should be read as a guide as to the direction of the trend as opposed to the exact magnitude of the trend.

4.4 D. Capacity-Building Asymmetry

UN Convention on Cybercrime emphasizes technical cooperation and capacity building in developing countries, whereas the Budapest Convention is less stringent in that respect. The EU AI Act, on the other hand, concentrates on compliance issues related to organizations operating within the European market or providing services there. This asymmetry matters because autonomous cyber-offensive capability is not confined to well-resourced jurisdictions, and states with limited technical capacity to investigate AI-enabled attacks or to implement AI governance obligations risk becoming both under-protected targets and unintentional safe havens for autonomous attack infrastructure. This problem has been extensively reported in the literature related to the wider subject of cybersecurity capacity building, which indicates that developing nations encounter consistent shortages of technical capacity, expertise, and digital forensic tools, which hinder their participation in international cybercrime cooperation efforts irrespective of the responsibilities imposed on them by the treaty. An AI-specific capacity-building gap layered on top of this existing shortfall compounds rather than duplicates the problem.

4.5 E. Synthesis of Findings

Taken together, the four patterns point to a governance architecture that is fragmented by design, thin on attribution, slow relative to the threat, and uneven in its support for lower-resource states. Each pattern corresponds to one of the reform proposals developed in Section 5, as summarized in Table 5.

Table VI. Findings of Section 4 mapped to the reform proposals of Section 5

Finding	Evidence in this section	Corresponding reform proposal (Section 5)
Fragmentation across tracks (4.1)	Lowest mean score is on scope (0.33); no instrument treats autonomous AI-enabled cybercrime as a unified object (Fig. 4)	Interoperability between AI risk-classification regimes and cybercrime enforcement powers
Attribution and accountability gap (4.2)	All three instruments at the indirect level; no agent-level or chain-of-responsibility rule (Tables 2 and 3)	Supplementary protocol to the UN Convention with definitions and attribution rules for autonomous and agentic AI
Enforcement lag relative to threat velocity (4.3)	3 of 40 ratifications; deferrals of 16 and 24 months; 70% rise in weekly attacks (Table 4, Figs. 5 to 7)	Real-time multilateral threat-information-sharing mechanisms
Capacity-building asymmetry (4.4)	Scores of 2, 1 and 0 across the three instruments (Table 2, Fig. 4)	Capacity-building funding linked to AI-incident investigation support

Table 6 shows that the reform proposals are not independent: each responds to a specific, evidenced gap. It also indicates where the evidence is strongest. The attribution finding is the most uniform across instruments, whereas the enforcement-lag finding relies partly on vendor-sourced threat data and should be treated as provisional until the instruments are fully operative.

5 V. CONCLUSION

The aim of this paper was to analyze whether and how the existing cybercrime and AI governance instruments cover the threat of autonomous AI-enabled offensive activity, and the empirical findings in Section 4 lead to one overarching conclusion: governance of this type of technology is inherently reactive, segmented into parallel regulatory regimes, and structurally slower to develop than the threat itself. The contribution is not just the recognition of the gap that already existed through single-instrument analysis but the comparison of the three instruments on a single framework that demonstrates empirically how large the gap is and where it resides, specifically within the criteria of attribution and capacity building, which the four recommendations below will address.

There are four recommendations that can be derived from this research. First, the Conference of the States Parties under the UN Convention should consider developing a supplementary protocol addressing AI-autonomous systems and providing for definitions differentiating between assisted and autonomous AI use and corresponding attribution. Second, policymakers should strive to ensure the interoperability between the risk classification regimes of AI, such as the EU AI Act, and cybercrime instruments, so that the AI system's risk level determines the terms of cross-border investigation in case of attacks involving this system. Third, given the mismatch identified in Section 4.3, states and industry should develop real-time threat information sharing platforms, inspired by public-private threat information sharing, on a time frame faster than a treaty ratification process. Fourth, the funding for capacity building under the UN Convention should also provide for specific support in investigating AI-enabled incidents.

This analysis is preliminary at best: two of the three instruments are still not in place, and the number of disclosed cases of autonomous AI-enabled cybercrime is limited. It is also, by design, a qualitative document analysis of instrument text coded by a single analyst rather than an empirical study of enforcement outcomes or a multi-coder reliability exercise, and the comparative scores in Table 1 and Fig. 2 should be read as a structured, contestable interpretation of the instruments' design rather than a definitive measurement; future replication with independent coders would help establish how sensitive the findings are to individual analyst judgment. Further research in the field could look into comparing these two acts once the ratification process of the UN Convention is complete and the provisions on high risk of the EU AI Act are operational. The development of technical auditing standards able to trace the decision-making process of an autonomous agent in the event of an incident will be extremely useful in improving the attribution processes identified to be absent in all three tools.

6 VI. DISCUSSION

Based on the results obtained during this study, the necessity of an integrated global governance framework for autonomous AI in the context of cybercrime prevention becomes evident. A comparative analysis of the UN Convention against Cybercrime (2024), the Budapest Convention (2001), and the EU Artificial Intelligence Act (2024) shows that these documents deal with different but complementary issues of cybercrime prevention, international cooperation, and AI risk management. However, due to different goals pursued by these pieces of legislation, challenges may arise as far as the issue of accountability and enforcing legal rules within different jurisdictions, as well as managing AI-enabled cybercrime, are concerned. Thus, the use of autonomous AI systems may present additional difficulties for establishing attribution, which makes human supervision crucial. In addition, the enforcement of these measures requires more international cooperation, capacity building, and lawful sharing of evidence between jurisdictions in order to protect privacy and fundamental rights. This study presents a comparative framework based on the criteria of scope, enforceability, attribution and accountability, and capacity building in order to identify the possible areas of governance gaps and opportunities for coordinating regulations. Nevertheless, the qualitative nature of this study and the use of a single coder reduce the validity of the results. In the future, it will be necessary to consider using several coders, conducting interviews, and examining some empirical cases in other jurisdictions.

References

[1] United Nations General Assembly, “United Nations Convention against Cybercrime,” Resolution 79/243, adopted Dec. 24, 2024.

[2] United Nations Office on Drugs and Crime, “UN Convention against Cybercrime opens for signature in Hanoi, Viet Nam,” press release, Oct. 25, 2025.

[3] Council of Europe, “Convention on Cybercrime,” Budapest Convention, ETS No. 185, Budapest, Hungary, Nov. 23, 2001. Entered into force Jul. 1, 2004.

[4] European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” *Official Journal of the European Union*, 2024.

[5] European Commission, AI Act Service Desk, “Timeline for the implementation of the EU AI Act.” [Online]. Available: https://ai-act-service-desk.ec.europa.eu/. Accessed: Aug. 2026.

[6] Anthropic, “Disrupting the first reported AI-orchestrated cyber espionage campaign,” threat intelligence report, Nov. 2025.

[7] Trend Micro Inc., “The AI-fication of cyberthreats: Trend Micro security predictions for 2026,” Nov. 2025.

[8] Check Point Software Technologies Ltd., *Cyber Security Report 2026*, Jan. 2026.

[9] IBM Security, “Cybersecurity trends: IBM’s predictions for 2026,” 2026. [Online]. Available: https://www.ibm.com/think.

[10] Forrester Research, “Predictions 2026: Cybersecurity and risk leaders grapple with new tech and geopolitical threats,” 2025.

[11] World Economic Forum, *Global Cybersecurity Outlook 2025*. Geneva, Switzerland: World Economic Forum, 2025.

[12] N. Kaloudi and J. Li, “The AI-based cyber threat landscape: A survey,” *ACM Computing Surveys*, vol. 53, pp. 1–34, 2021.

[13] T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi, “Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions,” *Science and Engineering Ethics*, vol. 26, no. 1, pp. 89–120, 2020.

[14] M. Caldwell, J. T. A. Andrews, T. Tanay, and L. D. Griffin, "AI-enabled future crime," *Crime Science*, vol. 9, no. 1, pp. 1–13, 2020.

[15] A. Marotta and S. Madnick, "The UN Cybercrime Treaty and AI: Navigating the intersection of technology and global policy," *Issues in Information Systems*, vol. 26, no. 4, pp. 1–16, 2025.

[16] D. Wicki-Birchler, "The Budapest Convention and the General Data Protection Regulation: Acting in concert to curb cybercrime?" *International Cybersecurity Law Review*, vol. 1, nos. 1–2, pp. 63–72, 2020.

[17] A. Chan *et al.*, "Visibility into AI agents," in *Proc. 2024 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil, 2024, pp. 958–973.

[18] G. A. Gabison and R. P. Xian, "Inherent and emergent liability issues in LLM-based agentic systems: A principal-agent perspective," in *Proc. 1st Workshop for Research on Agent Language Models (REALM 2025)*, Vienna, Austria, 2025, pp. 109–130.

[19] M. E. Diamantis, "Vicarious liability for AI," *Indiana Law Journal*, vol. 99, no. 1, pp. 317–334, 2023.

[20] N. Kshetri, *Cybercrime and Cybersecurity in the Global South*. London, U.K.: Palgrave Macmillan, 2013.

[21] Legiscope, "EU AI Act deadlines 2026–2027: Compliance calendar + fines." [Online]. Available: https://legiscope.com/. Accessed: Aug. 2026.

[22] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, Gaithersburg, MD, USA, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1.

[23] Organisation for Economic Co-operation and Development, "Recommendation of the Council on Artificial Intelligence," OECD Legal Instrument OECD/LEGAL/0449, May 22, 2019, amended May 3, 2024.

[24] United Nations Educational, Scientific and Cultural Organization, *Recommendation on the Ethics of Artificial Intelligence*. Paris, France: UNESCO, Nov. 2021.

[25] K. F. Khan, A. Ali, Z. F. Khan, and H. Siddiqua, "Artificial Intelligence and Criminal Culpability," in 2021 International Conference on Innovative Computing (ICIC), 2021, pp. 1–7.

[26] D. Ali, "Artificial Intelligence and Criminal Liability: Challenges in Articulation of Legal Aspects for Counter-Productive Actions of Machine Learning," International Journal of Instructional Technology and Educational Studies, vol. 2, no. 3, pp. 13–20, 2021, doi: 10.21608/ihites.2021.90657.1049.

[27] A. Ali, "Cyberspace and Organized Crime: The New Challenges of the 21st Century," International Journal of Advanced Humanities Research, vol. 2, no. 1, pp. 22–37, 2022, doi: 10.21608/ijahr.2022.256386.

[28] A. Ali and S. Qazi, "Darknet: The Study of Emerging Challenges of Cyber Terrorism and Organized Crimes," NACTA Journal, 2023.

[29] A. Ali, M. A. Khan, K. Farid, S. S. Akbar, A. Ilyas, T. M. Ghazal, and H. Al Hamadi, "The Effect of Artificial Intelligence on Cybersecurity," in 2023 International Conference on Business Analytics for Technology and Security (ICBATS), 2023, doi: 10.1109/ICBATS57792.2023.10111151.

[30] A. Ali, S. Ahmed, M. Hussain, T. A. Bhutta, A. Raza, M. W. Nadeem, et al., "The Synergy of Artificial Intelligence and Cybersecurity," in 2024 2nd International Conference on Cyber Resilience (ICCR), 2024.

Declarations

Funding. [Funder name and grant number, or a statement that no external funding was received.]

Conflicts of interest. [Relevant interests, or: The authors declare no conflicts of interest.]

Ethics and consent. [Approval reference / not applicable statement.]

Data availability. [Repository / access conditions, or not applicable.]

Code availability. [Repository / version, or not applicable.]

Author contributions. [CRediT roles and initials.]

Acknowledgments. [Optional. Delete if none.]

Use of generative AI. [Disclose AI tool use, or state that none was used.]

How to Cite This Article

SUGGESTED CITATION

A. Ali and Z. Farid, "Global Governance of Autonomous AI Systems in Cybercrime Prevention," International Journal of Security, Intelligence, Governance & Modern AI Computing (IJSIGMA), vol. 1, no. 1, pp. 92–107, 2026, doi: [DOI to be assigned].