Aligning AI-Driven Cyber Threat with Policy Responses: A CCI-Based Framework*

Sukyeong Heo, Jungmin Kang, and Dooho Choi†

Korea University, Sejong Campus, Republic of Korea
{h1238938,jmkang, doohochoi}@korea.ac.kr

Abstract

We propose a framework that rapidly classifies the type of artificial intelligence (AI) technology used in an AI-driven cyber threat when it occurs and derives appropriate policy response measures using the U.S. Defense Information Systems Agency's Control Correlation Identifier (CCI), a standardized identifier for mapping cybersecurity controls. This enables the rapid establishment of a consistent response pathway from the low-level categories of AI-driven cyber threat to the high-level policies. To achieve this, it researches existing AI threat classification systems, describes the structure and function of the CCI, and proposes a 'low-level Subcategory – CCI – high-level Category – CCI – high-level Policy' mapping model. The validity of this framework is examined through the cases 'Hong Kong CFO Deepfake (2024)' and 'Claude AI Chatbot Ransomware'.
Keywords: AI governance, AI Cyber Policy Framework, AI-driven cyber threats, CCI

1 Introduction

The rapid advancement and widespread adoption of artificial intelligence technologies have fundamentally transformed the cybersecurity landscape, creating unprecedented challenges for threat classification and response coordination. Recent industry analysis indicates that 78% of organizations have integrated AI capabilities into their operations, with the digital transformation market projected to expand from $2.5 trillion in 2024 to $3.9 trillion by 2027 (McKinsey, 2024). This technological proliferation has simultaneously enabled threat actors to weaponize AI capabilities, resulting in the emergence of advanced attack vectors, including deepfake-based social engineering, automated phishing, and AI-powered ransomware.

The evolution of AI-driven cyber threats has outpaced the development of corresponding classification and response frameworks. Current threat taxonomies operate as disconnected silos, with

* Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. 10, December 16-18, 2025, Sapporo , Japan. \space
† Corresponding author

detailed technical classifications failing to integrate with policy responses. For instance, while Stanford's AI Risk Taxonomy (AIR 2024) provides granular categorization of 314 risks, these classifications remain isolated from government agencies' high-level classification, like NIST's 'NIST AI 100-2e2025'. This fragmentation creates critical operational gaps where security practitioners cannot efficiently link specific threat types to actionable policy responses, resulting in delayed incident response and increased governance cost. This lack of coherence impedes immediate threat response when it occurs, so we need a systematic framework that can integrate between granular categorization and high-level classification.

Research Gap and Motivation

Despite research in both AI-driven cyber threat taxonomies and AI risk management frameworks, a significant gap exists in bridging granular threat classification with policy implementation. Current research lacks systematic approaches for linking granular AI threat classifications to policy responses from governance agencies. While frameworks like NIST's AI Risk Management Framework (AI RMF 1.0) provide high-level guidance, and detailed taxonomies offer technical classification, no existing methodology systematically connects these different levels.

Research Objectives and Contributions

This research addresses the gap by developing a mapping framework that systematically links disparate AI threat classifications and derives policy responses through the Defense Information Systems Agency's Control Correlation Identifier (CCI) methodology. The primary objective is to establish a bidirectional tracing mechanism that facilitates the transition from granular categorization to high-level classification, and vice versa. It enables rapid derivation of appropriate policy responses when specific AI threats are detected, while maintaining consistency in classifying AI-driven cyber threat types.

The study makes three key contributions to the field of AI cybersecurity governance: (1) introduces a five-layer mapping framework that connects technical threat classifications to policy response through CCI, (2) demonstrates the application of CCI methodology for AI-driven cyber threat classification, and (3) provides empirical validation through case studies, providing specific process for deriving policy response from low-level threat category.

Methodological Approach

Our approach leverages the CCI, an existing system for mapping actionable security tasks with high-level security requirements to manage general cyber threats, rather than proposing entirely new taxonomies. By utilizing the CCI as a bridge, we connect Stanford's AIR 2024 granular threat classification with NIST's high-level threat classification, ultimately linking to actionable policy from the AI RMF 1.0. This methodology ensures compatibility with existing governance agencies' AI cybersecurity frameworks while providing immediate applicability.

The remainder of this paper is organized as follows: Section 2 suggests eight stages of AI development and reviews related work in AI threat classification. Section 3 presents the CCI-based mapping model and demonstrates its application through the analysis of real-world incidents. Section 4 discusses the implications, limitations, and directions for future research.

2 Previous Research

2.1 AI Development Stages and AI-driven Cyber Threats

Figure Figure 1 and Table 1 present a comprehensive eight-stage framework that synthesizes the historical evolution of artificial intelligence with Jensen Huang's four stages of AI progression (Perception, Generative, Agentic, and Physical AI), accompanied by representative incidents at each developmental phase. Figure 2 illustrates the corresponding escalation in cyber-attack damage across these eight stages, revealing a correlation between AI evolution and threat severity.

Our analysis demonstrates that as AI technologies mature, the scale of attack damage expands exponentially [1, 2]. At the stage of Generative AI (2019-2024), the advent of generative AI has precipitated a 135% increase in social engineering attacks and a 260% surge in voice phishing incidents [3]. At the stage of Agentic AI (2025-2027), which is the current stage, Czech security firm ESET disclosed the 'PromptLock' ransomware. This represents a paradigm shift where AI can accelerate cyberattack automation. This ransomware leverages Large Language Models (LLMs) to autonomously generate and deploy malicious code without human intervention [4]. The PromptLock incident exemplifies the transition from human-directed attacks to fully autonomous cyber threat operations, where AI systems can independently identify targets, craft attack vectors, and execute sophisticated ransomware.

These developments underscore the critical need for integrated response frameworks that can address the increasingly sophisticated AI-driven cyber threats. But current AI governance approaches suffer from fragmentation, with risk management frameworks and threat taxonomies from various government agencies and research papers. Therefore, establishing a more systematic and integrated response framework is essential to counter AI-driven cyber threats.

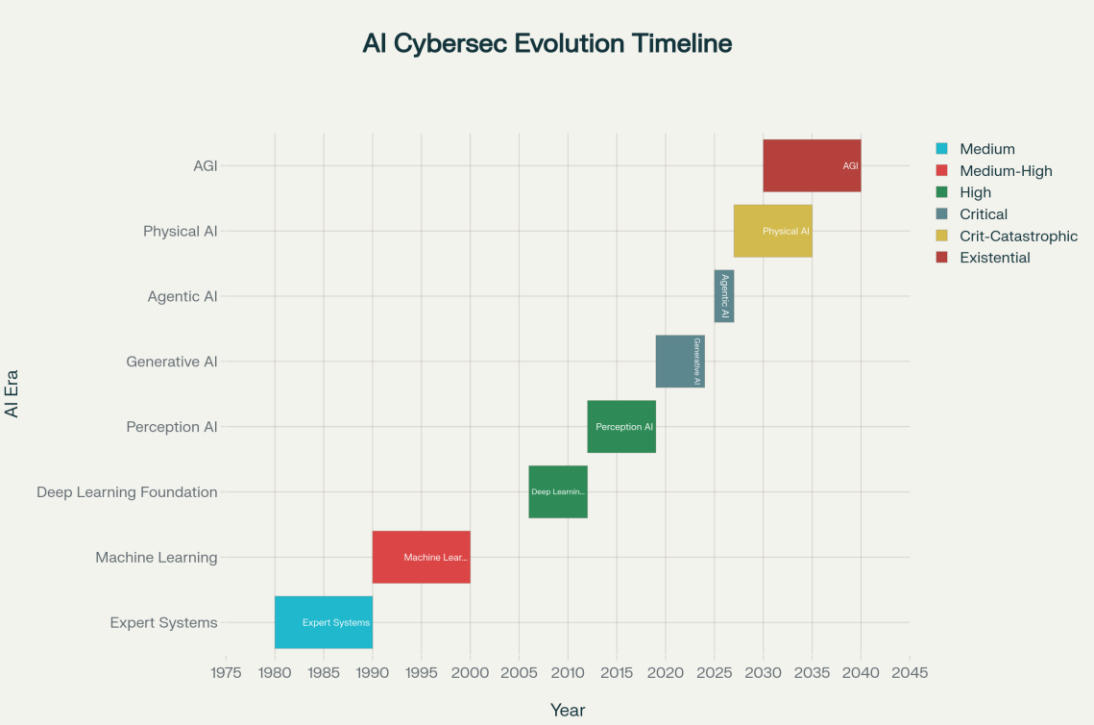

Figure 1: AI Evolution Timeline

AI Era	Attack Type	key Incident
Expert Systems (1980-1990)	Knowledge Base Manipulation	※ **No documented AI-based attacks** Era dominated by traditional malware (Brain virus 1986, Morris Worm 1988)
Machine Learning (1990-2000)	Statistical Model Evasion	• **Melissa Worm (1999)** Bypassed early Bayesian spam filters via feature manipulation, infected 1M computers, caused $80M+ damage
Deep Learning Foundation (2006-2012)	Gradient-based Adversarial Examples	※ **No real-world AI-driven attacks** Foundational research only; first systematic adversarial examples by Goodfellow et al. (2014)
Perception AI (2012-2019)	Adversarial Example Attack	• **Tesla Autopilot stop sign manipulation (2017-2019)** Adversarial patches caused vehicles to ignore stop signs, multiple safety incidents
Generative AI (2019-2024)	AI-Generated Social Engineering	• **Hong Kong deepfake CFO scam (2024)** AI-generated video call resulted in $25M financial loss
Agentic AI (2025-2027)	Autonomous Multi-step Attack with Local LLM	• **PromptLock ransomware (Aug 2025)** First fully autonomous ransomware using local LLMs for target identification and encryption without human intervention
Physical AI (2027+)	Cyber-Physical System Hijacking (predicted)	※ **Predicted** Autonomous vehicle fleet manipulation causing mass transportation disruption and casualties
AGI (2030-2040)	Self-Modifying Autonomous Attack (predicted)	※ **Predicted** AGI could autonomously attack and manipulate critical infrastructure, causing civilization-scale catastrophic damage

Table 1: Incidents for each AI Evolution Stages

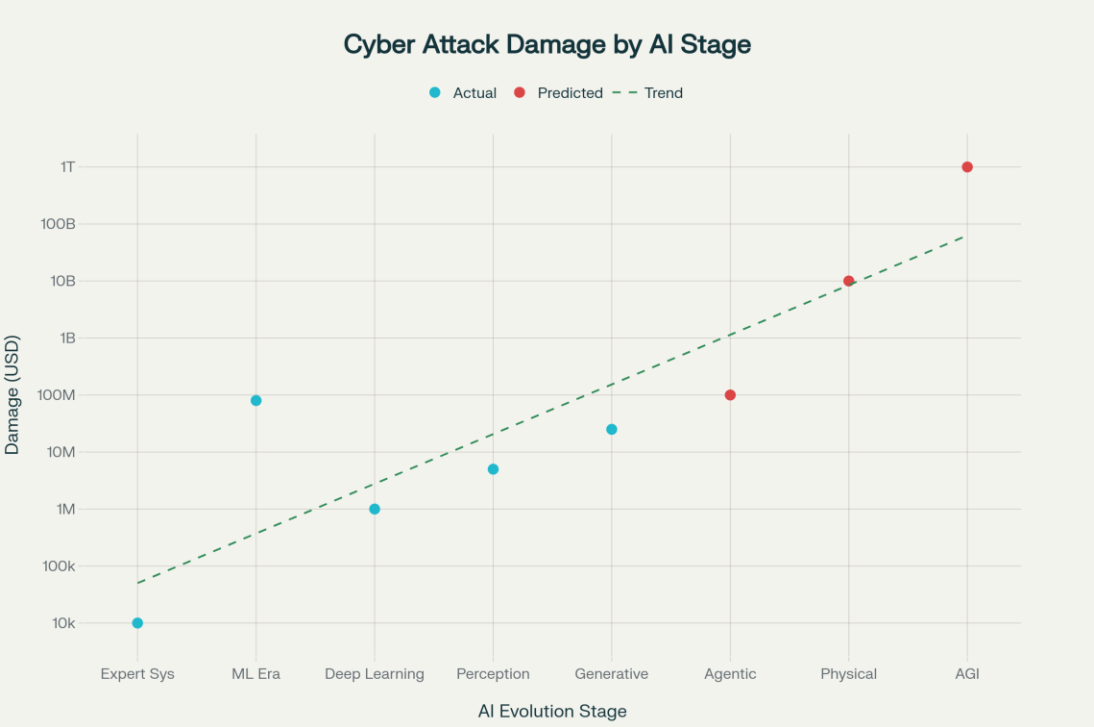

Figure 2: Cyber-attack damage by AI stage

2.2 The U.S. AI Cybersecurity Threat Classification

As AI-driven cyber threats proliferate and their associated damages escalate, government agencies and research institutions have undertaken efforts to develop a taxonomy for AI-driven cyber threats. Table 2 synthesizes the AI-driven cyber threat taxonomies announced by U.S. government agencies.

The National Institute of Standards and Technology (NIST), through its 'NIST AI 100-2e2025' framework, and the Department of Homeland Security (DHS), via its 'Safety and Security Guidelines for Critical Infrastructure Owners and Operators,' each propose tripartite categorization schemes for AI-based cyber threats [5]. While these frameworks provide macro-level perspectives for policy formulation and response, their abstract perspective leads to operational challenges.

On the other side, Stanford researchers introduced 'The AI Risk Taxonomy' (AIR 2024), which presents 314 granular threat types derived from a comprehensive analysis of governmental and corporate policy documents [6]. The AIR 2024 framework offers taxonomic detail, enabling micro-level threat classification. However, a critical limitation emerges: the framework lacks explicit

Agency	Reports/ Guidelines	Classification Criteria	Categories
NIST	NIST AI 100-2e2025 (March 2025)	Goal-based Attack Types	• Availability Breakdown • Integrity Violation • Privacy Compromise
DHS	Safety and Security Guidelines for Critical Infrastructure Owners and Operators (April 2024)	System-perspective Risk Categories	• Attacks Using AI • Attacks Targeting AI Systems • Design and Implementation Failures

Table 2: AI-driven Cyber Threat Classification

mappings to the higher-level categorization schemes established by NIST and DHS. This disconnection between governmental macro-frameworks and academic micro-taxonomies creates a fragmentation problem, wherein classification systems exist in parallel rather than as an integrated hierarchy.

The recent research 'The Unified Control Framework' by Eisenberg et al. (2025), acknowledges this fragmentation as a fundamental impediment to effective AI governance. Their work identifies the disconnected frameworks across institutions as a core challenge, proposing the Unified Control Framework (UCF) as an integrated governance approach that bridges risk management and regulatory compliance [7]. The UCF establishes bidirectional mappings between policy requirements, risk scenarios, and controls. While UCF's approach to linking risk taxonomies with policy requirements aligns conceptually with our research objectives, its focus remains confined to enterprise AI governance contexts rather than governmental policy frameworks.

The existing literature thus reveals a critical gap: the absence of a systematic methodology for bridging granular threat classifications with high-level policy frameworks in governmental contexts. This gap impedes rapid threat response and policy implementation, necessitating the development of an integrated mapping model that preserves the operational utility of detailed taxonomies while enabling strategic policy alignment.

3 CCI-Based AI Cyber Threat Response Model

This section presents our framework for mapping granular threat types to policy responses through the Defense Information Systems Agency's Control Correlation Identifier (CCI) methodology. We first establish the theoretical foundation for extending CCI to AI threat contexts, then demonstrate the framework's practical application through case studies.

3.1 Theoretical Foundation and Framework Architecture

The Control Correlation Identifier (CCI), developed by DISA, acts as a bridge for translating abstract policy requirements into specific and actionable technical implementations [8]. Traditional CCI applications focus on general cyber threat management, and our framework extends this to AI-specific correlation identifiers that manage specifically for AI-driven cyber threats.

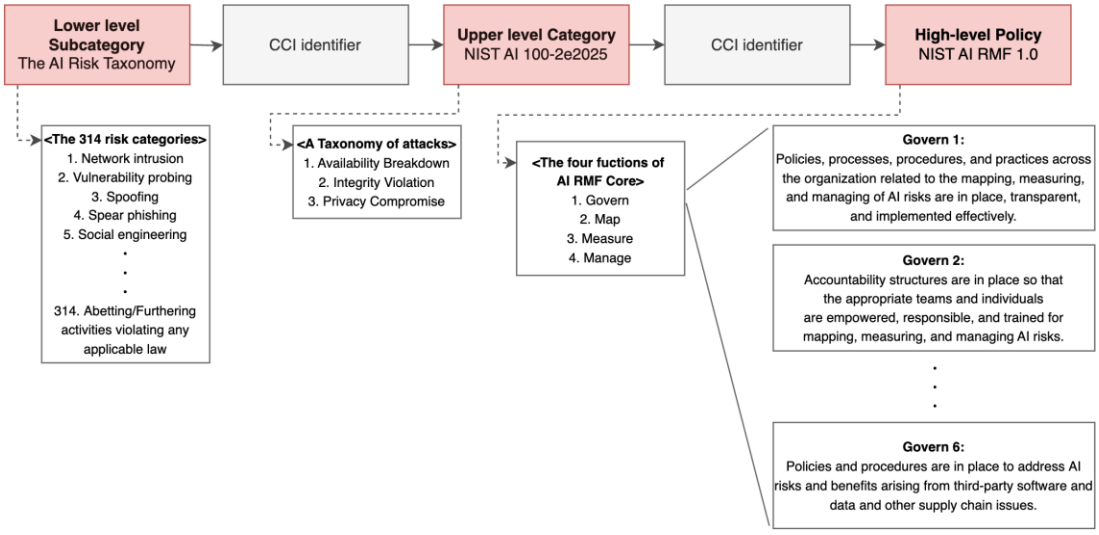

Figure 3: AI Cyber Threat Type Mapping Framework

Figure 3 illustrates our five-layer hierarchical mapping framework, which connects 314 granular threat classifications from Stanford's AIR 2024 taxonomy to NIST's three high-level categories (NIST AI 100-2e2025), ultimately linking to actionable guidance from the AI Risk Management Framework (AI RMF 1.0) [9]. The framework operates through two distinct mapping phases:

- **Phase 1: Threat Categorization Mapping (Lower-to-Upper)** The initial phase maps a specific threat incident from AIR 2024's granular category to NIST's upper-level category through CCI-AI identifiers. This mapping preserves bidirectional traceability, enabling both bottom-up threat identification and top-down policy application.
- **Phase 2: Policy Derivation Mapping (Category-to-Action)** The second phase translates categorized threats into actionable policy responses by mapping NIST categories to specific AI RMF 1.0 guidance. This phase employs a second CCI-AI identifier.

The dual-phase architecture ensures that threat-to-policy mappings remain consistent across governance agencies' environments while accommodating changes when an AI cyber threat management policy is newly announced. This design principle enhances framework compatibility with existing security infrastructure.

3.2 Case Study 1: Hong Kong CFO Deepfake

3.2.1 Incident Overview and Threat Characterization

To validate the framework, we analyze the 2024 Hong Kong CFO deepfake incident—a sophisticated social engineering attack that resulted in $25 million in financial losses. This incident exemplifies the convergence of multiple AI capabilities (voice synthesis, video generation, and behavioral modeling) to execute targeted deception at scale.

Figure 4 illustrates the complete mapping pathway from threat identification to policy response. The incident initially maps to AIR 2024's category "157. Impersonating others". Through CCI-AI-001, this granular classification correlates to NIST's "Integrity Violation" category, reflecting the attack's fundamental compromise of information authenticity and trustworthiness.

3.2.2 Policy Response Derivation

The framework's second mapping phase, executed through CCI-AI-0004, links the "Integrity Violation" categorization to the AI RMF 1.0's "Manage" function. This mapping activates specific policy guidance under Manage 1.3, which prescribes: " Responses to the AI risks deemed high priority, as identified by the MAP function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting."

Figure 5 details the resulting policy implementation pathway, demonstrating how granular threat classifications translate into actionable policy responses. The derivation process maintains full auditability, enabling post-incident analysis and framework refinement based on response effectiveness.

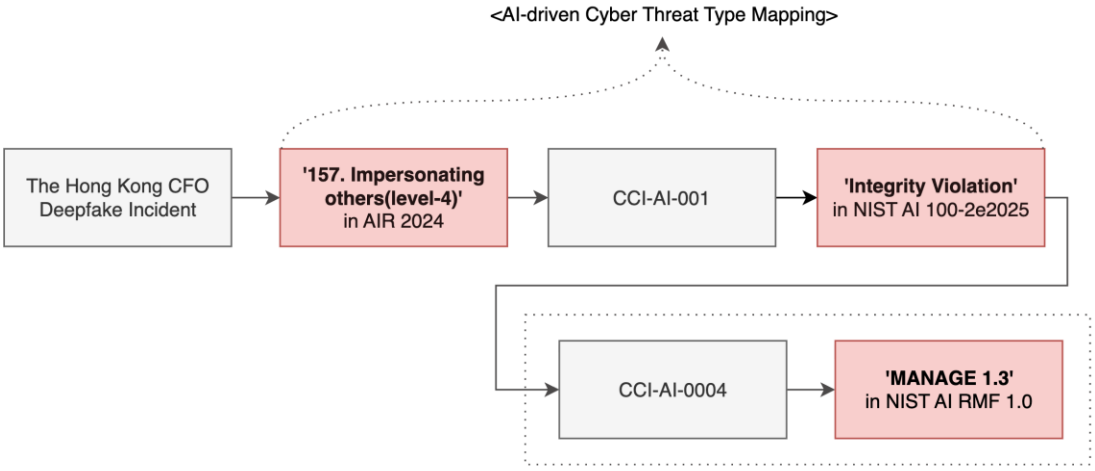

Figure 4: Example of how the 'Hong Kong CFO Deepfake' incident can be mapped in the framework

Example of Policy response with the framework

- The Hong Kong CFO deepfake incident, classified under the **'157. Impersonating others (level-4)'** subtype of '32. Fraud (level-3)'—a subcategory of '10. Deception attacks (level-2)' within the AIR 2024 Taxonomy's 'Societal Risks (level-1)'—resulted in a $25 million loss.
- This maps to the high-level category **'Integrity Violation'** in **NIST AI 100-2e2025** via **CCI-AI-001**, and ultimately maps to the **'MANAGE 1.3' category** in the final policy, **NIST AI RMF 1.0**, **via CCI-AI-0004**.
  - Accordingly, based on the content of the 'MANAGE 1.3' policy category, the policy response "Develop, plan, and document responses to high-priority AI risks using mitigation, transfer, avoidance, or acceptance strategies" can be implemented.

Figure 5: Example of a mapping-based policy response from 'Fig. 4'

3.3 Case Study 2: Claude AI Chatbot Ransomware

3.3.1 Incident Overview and Threat Characterization

The Claude AI Chatbot Ransomware incident involved cybercriminals exploiting Anthropic's Claude AI to automate large-scale ransomware attacks against 17 organizations across critical sectors, with ransom demands reaching $500,000 in Bitcoin through "vibe hacking" techniques that enabled low-skill actors to execute sophisticated cyber operations.

Figure 6 illustrates the complete mapping pathway from threat identification to policy response, using the incident 'Claude AI Chatbot Ransomware' as an example. The incident initially maps to AIR

2024's category "21. Other unauthorized actions on behalf of users". Through CCI-AI-005, this granular classification correlates to NIST's "Integrity Violation" category, reflecting the attack's fundamental compromise of information authenticity and trustworthiness.

3.3.2 Policy Response Derivation

The framework's second mapping phase, executed through CCI-AI-0006, links the "Integrity Violation" categorization to the AI RMF 1.0's "Manage" function. This mapping activates specific policy guidance under Manage 2.4, which prescribes: "Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use"

Figure 7 details the resulting policy implementation pathway, demonstrating how granular threat classifications translate into actionable policy responses. The derivation process maintains full auditability, enabling post-incident analysis and framework refinement based on response effectiveness.

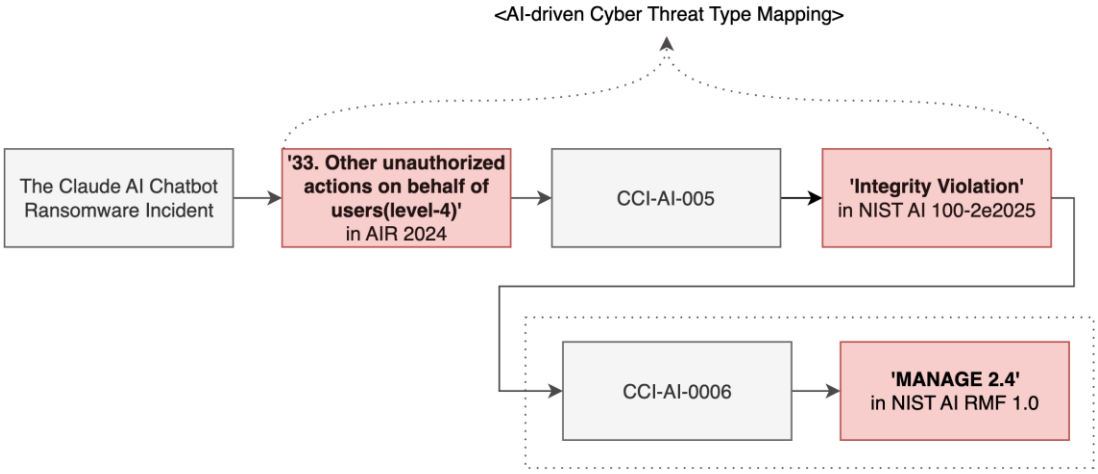

Figure 6: Example of how the 'Claude AI Chatbot Ransomware' incident can be mapped in the framework

Example of Policy response with the framework

- The 'Claude AI Chatbot Ransomware' incident, classified under the **'33. Other unauthorized actions on behalf of users (level-4)'** subtype of '5. Autonomous Unsafe Operation of Systems (level-3)'—a subcategory of '2. Operational Misuses (level-2)' within the AIR 2024 Taxonomy's 'System and Operational Risks (level-1)'— caused damages ranging from $75,000 to $500,000 and breached 17 organizations.
- This maps to the high-level category **'Integrity Violation'** in **NIST AI 100-2e2025** via **CCI-AI-002**, and maps to the final policy category **'MANAGE 2.4'** in **NIST AI RMF 1.0** via **CCI-AI-0005.**
- Accordingly, based on the content of the 'MANAGE 2.4' policy category, the policy response "Establish mechanisms and assign responsibilities to override, disengage, or deactivate AI systems that perform inconsistently with their intended use" can be implemented.

Figure 7: Example of a mapping-based policy response from 'Fig. 6'

4 Conclusion

In The AI-driven cyber threat mapping framework proposed in this study features a multi-layered structure connecting ‘Lower-level Subcategory – CCI – Upper-level Category – CCI – High-level Policy’. This enables rapid derivation of policy responses aligned with the specific threat type when it occurs and is expected to enhance the efficiency of policy formulation and technology development. Specifically, the CCI-based mapping structure enables bidirectional tracing during threats, supporting consistent policy responses. This integrates AI cyber threat classification systems proposed by various agencies, eliminating fragmentation and providing a unified reference standard for practitioners. Subsequent research will focus on further refining the mapping strategy and utilization methods of this framework, aiming to develop a ‘K-AI Cyber Threat Type Mapping Framework’ suitable for Korea’s cybersecurity policy. This seeks to enhance policy implementation effectiveness and response speed. Ultimately, this model can serve as a core foundation supporting the linkage between practical standards and policies during the AI cybersecurity strategy formulation process. The proposed mapping framework remains at a conceptual level; future research will concretize it to present detailed mapping strategies and implementation plans.

References

[1] Rodriguez, Popa, Flynn, Liang, et al., "A Framework for Evaluating Emerging Cyberattack Capabilities of AI", arXiv:2503.11917v3., 2025.

[2] Achuthan, K., Ramanathan, S., et al., "Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions", Frontiers in Big Data, 7:1497535., 2024.

[3] Guo, W., Potter, Y., et al., "Frontier AI's Impact on the Cybersecurity Landscape", arXiv:2504.05408v2., 2024.

[4] DailySecu, "First AI-Based Ransomware 'PromptLock' Discovered… Increased Risk of Attacks Using Local LLMs", https://www.dailysecu.com/news/articleView.html?idxno=169146, 2025.

[5] NIST, "NIST AI 100-2e2025", 2025.

[6] Zeng, Klyman, Zhou, Yang, et al., "AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies", arXiv:2406.17864v1., 2024.

[7] Eisenberg, I. W., Gamboa, L., et al., "The Unified Control Framework: Establishing a Common Foundation for Enterprise AI Governance, Risk Management and Regulatory Compliance", arXiv:2503.05937v1, 2025.

[8] Control Correlation Identifier (CCI), https://www.cyber.mil/stigs/cci/, 2025.

[9] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, 2023.