Chapter 3
Multi-Agent Generative AI Systems for Cyber Risk Evaluation and Control

Syed Mohd Faisal
http://orcid.org/0000-0003-2422-3535
Malla Reddy Technical Campus, Malla Reddy Vishwavidyapeeth (Deemed to be University), Hyderabad, India

Wasim Khan
http://orcid.org/0000-0003-2311-1451
Symbiosis Institute of Technology, Symbiosis International (Deemed to be) University, Pune, India.

Mohammad Ishrat
http://orcid.org/0000-0002-9699-4454
VIT Bhopal University, India

Anwar Ahamed Shaikh
Sanjivani University, Kopergaon, India

ABSTRACT

We present a systems-theoretic framework coupling Generative AI with multi-agent systems to make cyber-risk management a continuous, anticipatory control loop. The architecture: (i) risk-evaluation agents fusing telemetry with a knowledge graph; (ii) generative threat agents (GANs/diffusion/LLMs) synthesizing counterfactual attacks; and (iii) control agents trained with CTDE to allocate defenses under safety constraints. We implement a closed-loop simulation and evaluate ransomware and IoT-botnet cases using AUROC/AUPRC/F1 and time-to-containment. Ablations isolate contributions of generative foresight, coordination, and graph context. Results:

DOI: 10.4018/979-8-3693-8397-1.ch003

higher detection fidelity and faster containment than strong non-generative/non-agentic baselines. We discuss governance, limitations (sim-to-real, MARL stability), and directions in continual learning, risk-sensitive control, and human–AI teaming.

1. INTRODUCTION

The increasing dependence on interconnected digital infrastructures, encompassing enterprise networks, industrial control systems (ICS), cloud computing environments, 5G-enabled services, and Internet of Things (IoT) ecosystems, has amplified both the complexity and severity of cyber risks (Dwivedi et al., 2023). As organizations embrace digital transformation, the attack surface expands across heterogeneous devices, distributed architectures, and multi-layered applications. Consequently, the spectrum of cyber risks now includes ransomware, advanced persistent threats (APTs)(Chen et al., 2014), supply chain vulnerabilities, insider threats, and zero-day exploits, all of which evolve at a pace that often outstrips the ability of conventional defenses to respond. Traditional approaches to cyber risk management have relied heavily on static, rule-based mechanisms such as signature detection, compliance-based risk checklists, and periodic vulnerability scans. While these methods remain useful for addressing known threats, they are increasingly inadequate against the adaptive, stealthy, and automated nature of modern cyber-attacks (Zaidi & Faisal, 2018). Adversaries are progressively leveraging artificial intelligence (AI), automated reconnaissance, and even generative adversarial techniques to craft polymorphic malware, conduct deepfake-driven social engineering, and exploit system misconfigurations with precision (Girhepuje et al., 2024). In such a dynamic and adversarial environment, reactive strategies that rely on historical data or fixed detection patterns are insufficient. Instead, there is a pressing need for intelligent, adaptive, and proactive mechanisms that can continuously evaluate, predict, and control cyber risks in real time. Generative Artificial Intelligence (Generative AI) (Sengar et al., 2025) has emerged as a transformative paradigm in this context. Unlike conventional AI models that primarily classify or detect anomalies, generative models are capable of synthesizing realistic data distributions, simulating adversarial scenarios, and modeling complex threat landscapes. Techniques such as Generative Adversarial Networks (GANs) (Pan et al., 2019), diffusion models, and large language models (LLMs) (Naveed et al., 2025) allow for the generation of synthetic attack patterns, adversarial inputs, and hypothetical exploit trajectories. These capabilities facilitate the proactive anticipation of cyber risks, enabling defenders to test system resilience against a wide variety of threat conditions, including low-probability but high-impact events. However, when applied in isolation, generative models face inherent challenges: they are computationally