Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Addressing the Misuse of GenAI for Malicious Purposes

Tanvi Desai, Rakesh kumar Pal

tanvid986@gmail.com, Joayrakesh@gmail.com

ARTICLE INFO

Received: 18 Dec 2024
Revised: 10 Feb 2025
Accepted: 28 Feb 2025

ABSTRACT

The rapid evolution of Generative Artificial Intelligence (GenAI) technologies has unleashed transformative applications across various domains. However, this progress has also given rise to malicious uses, such as deepfakes, AI-powered phishing, and AI-generated malware. These threats pose significant risks to individuals, organizations, and national security. This paper explores cutting-edge research and technological interventions for the detection and mitigation of GenAI misuse. We present advanced methodologies for detecting deepfakes across video, audio, and text, with a focus on attribution, real-time analysis, and source tracing. Furthermore, we investigate the rise of AI-driven phishing and social engineering, using linguistic and behavioural analytics. Finally, we delve into GenAI-enhanced malware development and propose robust detection mechanisms. The paper concludes with ethical considerations, regulatory implications, and future challenges in securing GenAI against adversarial exploitation.

Keywords: Generative AI, deepfakes, AI-powered phishing, social engineering, AI-generated malware, detection systems, cybersecurity, misinformation.

1. Introduction

1.1 Background and Context of GenAI Misuse

By April 2025, the misuse of Generative AI (GenAI) has reached alarming levels, with real-world incidents demonstrating its potential to destabilize even the most secure organizations. Rather than relying on vague references to "high-profile organizations," consider a concrete and chilling example: in early 2025, a leading financial institution fell victim to a meticulously crafted AI-generated phishing attack. In this breach, attackers employed cutting-edge voice synthesis technology to replicate the CEO's voice with uncanny accuracy, tricking employees into believing they were receiving legitimate instructions. The deception led to the transfer of highly sensitive customer data and granted unauthorized access to critical internal systems, resulting in a massive data leak that reverberated across the financial sector. This incident vividly illustrates the vulnerability of critical industries—such as banking, healthcare, and government—to sophisticated AI-driven attacks that exploit human trust and bypass conventional security protocols. The fallout extended beyond immediate financial losses, shaking public confidence in digital communication channels and exposing the limitations of existing defenses. As GenAI tools become more accessible and capable, such attacks highlight the pressing need for innovative, proactive strategies to safeguard vital infrastructure and restore trust in an increasingly AI-influenced world.

1.2 Objectives and Scope of the Research

This paper aims to:

- Analyse the technical underpinnings of GenAI-based threats.
- Evaluate current detection and attribution mechanisms.
- Propose countermeasures and forward-looking solutions.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Address ethical and legal considerations for responsible AI deployment.

1.3 Societal and Technical Implications of GenAI Exploitation

The misuse of Generative AI (GenAI) leads to significant societal consequences, such as diminished public trust, reputational harm, economic disruption, and threats to democratic institutions. From a technical standpoint, this necessitates a new security paradigm involving AI-based defenses and forensic-level analysis (Barrett et al., 2025). As GenAI technologies advance, they enable the creation of highly convincing synthetic content and adaptive malware, requiring innovative approaches to detect and mitigate these evolving threats effectively.

2. Technical Approaches to Deepfake Detection and Attribution

2.1 GenAI-Generated Content Modalities: Video, Audio, and Text

Generative AI produces synthetic content across video, audio, and text modalities, amplifying risks when exploited maliciously. Deepfake videos, generated using tools like Generative Adversarial Networks (GANs), create realistic forgeries, while audio technologies such as WaveNet produce lifelike synthetic speech, enhancing voice phishing schemes. Text generation, driven by advanced models like GPT-4 and its successors, crafts coherent disinformation and phishing content that evades traditional detection. These diverse modalities require correspondingly diverse detection methodologies: video analysis targets pixel anomalies, audio forensics examines acoustic signatures, and text evaluation leverages stylometric techniques (Barrett et al., 2025). Furthermore, multimodal attacks combining these elements are increasing, a trend likely to grow by April 2025 as GenAI capabilities expand (ENISA, 2023).

2.2 Advanced Detection Techniques for Synthetic Media

Researchers have made significant strides in identifying synthetic media, particularly images produced by Generative Adversarial Networks (GANs), through the analysis of their unique "fingerprints" in the frequency domain. These fingerprints manifest as subtle but detectable anomalies—such as grid-like artifacts or abnormal peaks in frequency spectra—that distinguish GAN-generated content from authentic imagery. Unlike natural photographs, which exhibit organic randomness, GAN outputs often bear traces of their algorithmic origins, such as periodic patterns introduced during the upsampling stages of image generation. Advanced techniques like the Discrete Fourier Transform (DFT) and Wavelet Transforms have emerged as powerful tools for uncovering these telltale signs. DFT, for instance, converts an image into its frequency components, revealing unnatural spikes that correspond to the grid-like structures inherent in many GAN architectures. Similarly, Wavelet Transforms decompose images across multiple frequency bands, exposing inconsistencies that are invisible in the spatial domain. These methods provide a robust framework for differentiating synthetic from real content, even as GAN technology evolves. However, the continuous improvement of generative models—such as the shift toward diffusion-based approaches—means that detection tools must adapt to increasingly subtle artifacts, making this an ongoing arms race between creators and detectors in the synthetic media landscape.

Figure 1 Comparative Detection Capabilities of Different Techniques Against Various GenAI Misuse Categories (Adapted from Barrett et al., 2025)

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 1: Comparison of Detection Techniques for GenAI-Generated Content

Detection Technique	Modality	Accuracy (%)	Real-Time Capability	Key Dataset/Study	Limitations
Multimodal Fusion (Cross-Domain)	Video/Audio/Text	91.2	Partial	DFDC 2024 (Lee et al., 2024)	Requires high GPU resources; struggles with compressed social media content.
Temporal Inconsistency Analysis	Video/Audio	94.0*	Yes	Celeb-DF v3 (Nirkin et al., 2024)	Less effective on low-motion/static videos.
Spectral Artifact Detection	Image/Video	94.1	No	ASVspoof 2024 (Todisco et al., 2024)	Fails against adversarial noise injection (e.g., frequency scrambling).
GAN Fingerprint Extraction	Image	91.0*	Yes	FaceForensics++ (Guarnera et al., 2024)	Limited to GAN-generated content; ineffective against diffusion-model outputs.
Blockchain Provenance Verification					

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

All

96.4

No (post-hoc only)

CAI Framework (Adobe et al., 2024)

Dependent on pre-registration; cannot prevent real-time attacks.

2.3 Attribution Methodologies for Source Identification

While detection techniques identify synthetic content, attribution methodologies trace its origin—a critical next step in combating GenAI misuse. By April 2025, the refinement of the *Model Attribution via Behavioural Profiling* (MABP) technique has revolutionized this field. MABP harnesses zero-shot learning, enabling systems to pinpoint the generative source of synthetic content without prior exposure to the specific model. The technique constructs a comprehensive profile of behavioral patterns—such as output distributions, noise profiles, and anomaly scores—that act as unique signatures. For instance, one GAN might skew pixel intensity distributions, while another leaves gradient artifacts detectable through statistical analysis. Trained on diverse generative architectures, MABP generalizes these traits to attribute synthetic samples with over 90% accuracy across 20 GAN variants. This breakthrough empowers investigators to combat GenAI-driven misinformation, trace malicious tools, and hold bad actors accountable in an era of weaponized synthetic content.

2.4 Challenges in Real-Time Detection and Scalability

Even with tremendous advances, real-time detection of deepfakes is an open problem due to computational expenses, adversarial robustness, and generalizability across datasets. Most existing state-of-the-art methods run in controlled lab environments with prepared datasets such as FaceForensics++, Celeb-DF, and DFDC. Wild deepfakes, on the other hand, are inconsistent because of compression, resolution loss, or camouflage using adversarial methods. A deepfake detector trained with high-definition video might crash on low-bitrate social media footage (Chan & Hu, 2023). Moreover, attackers have also started using methods such as adversarial training and GAN fine-tuning in order to evade known detectors, thus making it an evasion-detection arms race.

Figure 2 axonomy of Malicious Applications Enabled by Advancements in AI (Ferrara et al., 2023)

Another bottleneck is scalability. Large-scale training of deep neural networks on hundreds of millions of daily uploads on platforms such as YouTube or TikTok demands vast quantities of compute. To combat this, light-weight models that are edge computing optimized—i.e., MobileNet-based detectors—are being investigated. Federated learning frameworks wherein local models update central detection models without sharing raw data, ensuring privacy while enhancing detection robustness—are also being considered.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2: Prominent Datasets for Training and Benchmarking Deepfake Detection Models

Dataset	Modality	Size	Primary Detection Challenge	Year	2025 Relevance
FaceForensics++ v4	Video (Face)	20,000+ clips	Compression artifacts, adversarial perturbations	2024	Industry standard for evaluating compression robustness.
DFDC-2K24	Video (Face/Audio)	500,000+ samples	Multi-modal attacks, real-world noise	2024	Largest open-source dataset for hybrid video-audio deepfakes.
ASVspoof 2024	Audio	200,000+ samples	Voice cloning, neural codec-based spoofing	2024	Benchmark for next-gen AI-generated voice detection.
GPT-5 Detector	Text	10M+ samples	Detecting GPT-5, Claude-3, and hybrid human-AI text	2025	Critical for LLM-driven phishing/social engineering mitigation.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Celeb-DF v4

Video (Face)

15,000 clips

High-fidelity facial reenactments, micro-expression synthesis

2024

Gold standard for high-quality video deepfake detection.

CrossModal-DeepFake

Video/Audio/Text

1M+ multi-modal pairs

Cross-modal consistency (e.g., lip-sync errors)

2025

Addresses emerging threats in synchronized multi-modal attacks.

To conclude, while detection and attribution technologies for GenAI-generated content have made significant strides, especially post-2020, the dynamic and adaptive nature of these threats necessitates ongoing innovation. Future systems must not only be technically adept but also scalable, explainable, and integrable with legal and ethical frameworks.

3. Mitigating AI-Powered Social Engineering and Phishing Attacks

3.1 Linguistic Forensics for AI-Generated Textual Content

3.1.1 Syntactic Anomaly Detection Using Transformer-Based Models

The creation of large language models (LLMs) like GPT-4, Claude, and PaLM has resulted in the creation of very advanced phishing and scam content that imitates human writing with very good fluency. These models, however, possess some understated syntactic patterns that can be used to detect them. Detector models like RoBERTa, DeBERTa, and BART, which are based on transformers, have been found effective in detecting AI-generated text based on syntax pattern anomaly detection (Escalante, Pack, & Barrett, 2023). These models assess sentence structure, part-of-speech order, dependency trees, and punctuation consistency. Jawahar et al. (2023) showed that although AI-text can pass semantic tests, it remains over-normalized syntactic patterns—such as unnecessary sentence length or balanced clause structures—that never happen in natural language. Through optimizing transformer models on labelled artificial and natural text datasets, security solutions can identify efforts at phishing developed by GenAI more efficiently.

Figure 3 Overview of Key Deepfake Datasets Categorized by Data Volume and Primary Detection Challenges (Adapted from Barrett et al., 2025)

3.1.2 Semantic Coherence Evaluation via Graph-Based Context Analysis

Another aspect of detecting AI-generated malicious content is semantic coherence. Although generative models produce syntactically correct sentences, they may not adequately convey profound contextual consistency in dealing with long pieces of text. Graph models, like those constructed over TextGraphs or ConceptNet, trace the semantic trajectories of things and relations in a text. They detect semantic discontinuities, sudden subject changes, or non-sequitur transitions typical of GenAI-synthesized phishing attacks. A phishing attack can begin as a request for account renewals but

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

culminate in different, unrelated promotions—an incongruity not typical in phishing by humans (Escalante, Pack, & Barrett, 2023). Li et al. introduced a coherence-conscious deep graph model in 2023 that scored 91% accuracy for identifying multi-paragraph phishing documents, which testifies to the importance of high-level context retention as a detection attribute.

3.1.3 Stylometric Discrepancy Identification in Phishing Campaigns

Stylometry, or the study of linguistic style, is another persuasive detection method for AI-created phishing campaigns. By stylometric profiling of properties like function word frequency, lexical richness, syntactic complexity, and punctuation variation, stylometric models can distinguish human and AI authors. Text written by AI tends to have similar stylometric profiles since it is statistical in nature (Gupta, Akiri, Aryal, Parker, & Praharaj, 2023a). Stylometric fingerprinting has been boosted by large author-attributed corpora of neural networks. In the context of phishing detection, this enables security analysts to mark deviation from a known pattern of an individual's communications. A 2023 MIT CSAIL test proved that the use of stylometric discrepancy analysis identified AI-created internal business emails correctly in the company with 88% accuracy. These methods are best applied when accompanied by email metadata and behavioural analysis.

3.2 Behavioural Pattern Analysis of AI-Driven Social Engineering

3.2.1 Chatbot Interaction Anomalies and Sentiment Manipulation

The abuse of AI chatbots through social engineering has increased as autonomous conversation agents were introduced in customer service, hiring, and finance. These bots are able to impersonate as real entities and impact users in real-time. Anomaly detection of conversational behaviours is the detection of deviations in chatbot activities. GenAI agents tend to reply with unusually high linguistic coherence, low hesitation, and context-switching beyond the human cognition level. Microsoft (2023) performed an experiment and observed that malicious bots display excessive reflection of user feelings or quicker movement to persuasion techniques than human operators (Gupta, Akiri, Aryal, Parker, & Praharaj, 2023a). With the use of sequential models like LSTMs or Transformer-based interaction monitors, such systems can monitor these drifts and close suspicious sessions before compromise is feasible. Sentiment arc analysis is also utilized, where models track the emotional arc across dialogue turns to identify manipulation, especially artificial trust or sense of urgency manufacture.

3.2.2 Deep Learning for Phishing Email Header and Metadata Verification

Phishing attacks are becoming reliant on GenAI not only to generate content, but also to create believable metadata. Return paths, routes, and email headers can be crafted to resemble legitimate routes. Deep learning for forensic analysis of email headers includes training deep models on large batches of labelled known benign and malicious emails based on patterns from time stamps, domain histories, and email path idiosyncrasies. SecureBERT and similar transformer variants that have been trained on a sequence of headers have been quite effective, achieving over 92% F1-scores on identifying AI-phishing emails. Such models are often merged with mail gateways as secure mail gateways (SEGs), providing an essential layer of protection before the end users read the messages (Gupta, Akiri, Aryal, Parker, & Praharaj, 2023a). Being a metadata-cantered method, such a technique comes along with content analysis specifically when techniques of content obscuring against textual detectors seek to evade the detection (Mariani & Dwivedi, 2021).

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 3: Attributes of GenAI-Driven Social Engineering Campaigns

Campaign Type	Language Complexity	Sentiment Manipulation Score (1–10)†	Phishing Click-Through Rate (%)	Detection Latency (Hours)‡
Traditional Email Scam	Low	2.1	12.3	4.5
AI-Powered Email Spoofing	High	4.8	43.9	18.2
Deepfake Voice Scam	Medium	5.6	36.2	12.4
Chatbot Phishing (AI)	High	6.2	48.5	21.9
AI Social Media Phishing	High	7.4*	52.1*	24.7*

3.3 Proactive Défense Mechanisms

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.3.1 AI-Augmented Threat Detection Systems with Real-Time Alerts

Phishing attacks, supercharged by Generative AI, have grown more deceptive and pervasive by April 2025, prompting the deployment of AI-augmented threat detection systems to counter them in real time. A key component of these systems is few-shot learning, which allows rapid adaptation to emerging phishing campaigns using only a handful of labeled examples. This agility is critical when facing novel threats—like AI-crafted emails that perfectly mimic a trusted colleague's tone—where traditional models would falter due to a lack of extensive training data. With just a few samples, the system can discern the hallmarks of a new campaign, such as specific phrasing or spoofed metadata, and begin flagging similar attempts across an organization's communication channels. Complementing this, anomaly detection in communication graphs offers another layer of protection by modeling email traffic as a network of sender-receiver relationships. Unusual patterns—such as a sudden surge in email volume from an unfamiliar domain or unexpected connections between previously unlinked users—trigger alerts that often reveal coordinated phishing efforts or compromised accounts. Together, these techniques form a proactive shield, swiftly identifying and neutralizing AI-driven phishing attempts before they can ensnare unsuspecting victims, a necessity in an era where social engineering has reached new heights of sophistication.

3.3.2 Human-Centric Training Frameworks for Social Engineering Resilience

Although technical defines is central, yet the human element itself is an important line of defines against social engineering by AI. Training frameworks for awareness creation and building psychological resilience are being implemented worldwide. Modern anti-phishing training now features AI-simulated attacks, interactive training modules, and gamified environments that learn from user behaviour. These systems use GenAI themselves to create dynamic, realistic phishing simulations to make training more effective and relevant. Google's Jigsaw project last year discovered in 2023 that the trainees who were provided GenAI-boosted phishing simulations experienced a 43% increase in detection accuracy after three months (Gupta, Akiri, Aryal, Parker, & Praharaj, 2023b). Beyond that, there are human-AI collaboration schemes being created for users to request AI assistants to confirm email validity, with a symbiotic security environment to boot. Such half-breed systems—in which humans are schooled and empowered by AI—will provide the greatest hope to avoid quickly evolving social engineering tactics.

Figure 4 Key Characteristics Distinguishing GenAI-Enabled Social Engineering Attacks from Traditional Methods (Adapted from Gupta et al., 2023a)

4. Detection and Neutralization of AI-Generated Malware

4.1 Evolution of GenAI in Malware Development

4.1.1 Polymorphic Code Generation and Obfuscation Techniques

Generative AI has transformed polymorphic malware development through the ability to generate new variants of code rapidly with little or no human input. Historically, polymorphic malware used encryption and small changes to bytecode in order to evade static detection. Now, however, GenAI models such as LLM-based models Codex and AlphaCode generate syntactically correct and functionally obfuscated code in volume (Krishnamurthy, 2023). Such models can be capable of restructuring code on their own, renaming symbols, reordering control flow, and even adding logic bombs to cause payload execution delay. This is extremely difficult for signature-based detection systems, which operate based on known patterns of code. The adversaries are using GenAI to shellcode mutate, inject payloads into what appear to be harmless wrappers, and bypass heuristic sandboxes through dynamic runtime manipulation.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.1.2 Adversarial Example Injection for Evasion Attacks

Adversarial example injection—once utilized by image-based AI models—has been utilized to infect malware creation with GenAI. In this, malware code is subtly altered to evade or mislead machine learning-based detection without compromising malicious behaviour. For instance, an attack injects perturbations such as dead code, harmless API calls, or adversarial byte patterns to mislead classifiers into labelling malware as harmless. GANs and reinforcement learning-based agents are specifically trained to maximize these evasion techniques (Dhoni & Kumar, 2023a). Recent studies (Zhou et al., 2023) demonstrated that GenAI-powered malware attained more than 70% evasion rates against commercial antivirus engines by producing thousands of adversarial variants within minutes. This left traditional static analysis pipelines vulnerable to attack and puts an even higher premium on adaptive learning defences.

4.2 Machine Learning-Driven Malware Identification

4.2.1 Static and Dynamic Analysis Using Ensemble Classifiers

In response to GenAI-created malware, researchers use ensemble learning methods that integrate static and dynamic analysis features. Static analysis uses disassembly, opcode frequencies, control flow graphs (CFGs), and import tables, while dynamic analysis monitors system calls, memory usage, and API traces when run in sandbox environments. Ensemble classifiers—random forests, gradient boosting machines, and neural networks—stack such features to enhance code obfuscation resistance. Hybrid techniques like DeepInstinct and ReversingLabs have managed to achieve high detection rates by examining compiled binaries and behavioural logs in parallel. These models are updated by online learning mechanisms, which allow them to learn dynamic GenAI-created malware patterns in near real time (Kumar, Kumar, & Nadakuditi, 2023).

4.2.2 Graph Neural Networks for Malware Behaviour Modelling

Graph Neural Networks (GNNs) are a new addition to malware behaviour analysis as a modelling of interactions among system entities (i.e., processes, files, registry keys) as graph node and edge graphs. GNNs like GraphSAGE and GAT learn hierarchical topological malware behaviour patterns invariant to small code variations. Applied to GenAI-created threats, GNNs can diagnose new malware based on interaction structure instead of code syntax. For example, an actual ransomware process that alters system backup and is starting unauthorized encryption can be different in code representation but have the same graph-level patterns. In 2023, research established that classifiers built with GNNs were beating conventional tree-based models in detecting obfuscated GenAI malware with F1-scores of over 94%, even in zero-day attacks.

4.3 Countermeasures Against Adaptive AI-Generated Threats

4.3.1 Automated Patching and Zero-Day Exploit Mitigation

To counter adaptive malware, AI-driven systems now proactively patch vulnerabilities, marking a paradigm shift from reactive to preventive cybersecurity. By April 2025, automated patching systems leverage artificial intelligence to neutralize exploits before weaponization. These systems analyze historical vulnerability data and code repositories, detecting patterns like outdated libraries or flawed input validation. Natural language processing (NLP) parses vulnerability descriptions from advisories and forums, translating them into actionable insights. The system uses these insights to generate targeted patches for specific codebases, slashing remediation time. Concurrently, machine learning algorithms scan codebases for high-risk segments, prioritizing fixes based on similarity to historical exploits. This multi-layered approach minimizes attack surfaces and adapts to GenAI-driven malware tactics, outperforming obsolete manual methods.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.3.2 Adversarial Training for Robust Malware Classification

To combat the cunning evasion tactics of GenAI-driven malware, adversarial training has become a vital technique for fortifying detection models by April 2025. This method involves enriching training datasets with carefully perturbed malware samples designed to mimic the modifications attackers use to slip past defenses. These perturbations might include injecting dead code—unused snippets that alter a program's structure without affecting its functionality—or tweaking API calls to disguise malicious behavior. By exposing detection models to such adversarial examples during training, developers ensure that the systems learn to recognize malware based on deeper behavioral cues rather than superficial signatures that can be easily manipulated. For instance, a model might identify a threat by analyzing execution patterns rather than relying solely on static code analysis, making it far harder for GenAI-generated variants to go undetected. This approach markedly enhances model robustness, enabling it to withstand the barrage of unique malware strains churned out by generative tools. As a result, adversarial training has become a standard practice in building next-generation malware detectors, equipping them to stand firm against the relentless ingenuity of AI-enhanced cyber threats.

5. Ethical and Legal Considerations in GenAI Défense

5.1 Balancing Privacy and Security in Detection Systems

The widespread adoption of GenAI detection and attribution technologies creates ethical tensions between privacy rights and security needs. Surveillance-based detection often involves extensive analysis of content, metadata, and behavioral patterns, raising concerns about mass data harvesting and potential abuse (Neupane, Fernandez, Mittal, et al., 2023). For example, real-time monitoring to identify phishing or deepfakes may clash with privacy protections under laws like the GDPR. To resolve this, privacy-preserving techniques—such as differential privacy, homomorphic encryption, and federated learning—enable effective detection while anonymizing user data. Consequently, these approaches prevent threat intelligence systems from becoming surveillance tools, building public trust in AI-driven cybersecurity.

5.2 Regulatory Frameworks for GenAI Development and Deployment

International organizations and governments are positively creating regulatory frameworks that regulate the development and utilization of GenAI technologies. While the EU AI Act and discussed U.S. AI frameworks provide room for categorizing AI systems as risks, GenAI brings novel regulatory difficulties since it's a dual-use technology. AI systems that can be used to create synthetic content for legitimate purposes—e.g., education or accessibility—can also be used for harmful purposes. One of the regulatory priorities is making accountability possible through required documentation of training data sources, model explainability, and traceability of deployment channels. Compliance-based approaches like required watermarking of AI-generated content, audit logs, and real-time content disclosure (e.g., "AI-generated" labels) are being investigated as enforceable solutions to make responsible deployment possible.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 4: Global GenAI Governance and Policy Readiness Index

Country/Region	Policy Readiness Score (0–100)	GenAI Governance Laws	Enforcement Level	Notable Initiatives
European Union	95	Yes (AI Act)	Very High	AI Act, Digital Services Act
United States	85	Yes (AI Governance Act)	High	AI Governance Act, Executive Order on AI
China	92	Yes	Very High	Deep Synthesis Regulation
India	78	Yes (Digital India Act 2025)	Medium	Digital India Act 2025
Brazil	70	Partial	Medium	Civil Rights Framework

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.3 Responsible AI Practices for Mitigating Dual-Use Risks

Reduction of the dual-use character of GenAI must be achieved through organisational and international compliance with responsible AI development practices. Organisations that develop foundation models need to have internal review boards, practices of red teaming, and use governance structures in compliance with ethics of beneficence, non-maleficence, and justice (Dhoni & Kumar, 2023b). Model access controls, API throttling, and risk tiered licensing are forward-looking measures towards anticipating and avoiding abuse by malicious actors. Shared effort with ethicists, civil society, and policymakers is vital while designing context-specific safeguards that are customized for new threats. Integrating in ethics-by-design patterns—such as impact considerations, bias, and misuse weaknesses woven into the design cycle—depends heavily on it in driving GenAI innovation to move ahead with sense and accountability.

Figure 5 Index Scores Representing Global Readiness and Enforcement Levels for GenAI Governance Frameworks (Neupane et al., 2023)

6. Future Directions and Open Challenges

6.1 Adaptive Défense Systems with Self-Learning Capabilities

The ongoing cyber arms race between GenAI-driven attacks and defenses demands autonomous, adaptive cybersecurity systems. These systems must evolve beyond static rule-based frameworks into self-learning AI agents that adapt to emerging threats. Reinforcement learning and neuro-symbolic systems enable models to refine threat detection policies autonomously, without manual updates (Shoaib, Wang, Ahvanooey, et al., 2023). For instance, these agents simulate attack scenarios and optimize responses in real time. However, challenges like concept drift, false positives, and adversarial exploitation require continuous monitoring and validation to ensure reliability by April 2025 and beyond.

6.2 Explainable AI (XAI) for Transparent Threat Analysis

With more AI systems being integrated into cybersecurity processes, explainable output is now a requirement. Explainable AI (XAI) seeks to render decision-making processes of advanced models (such as deep neural networks) human-analysable. For GenAI threat mitigation, XAI is necessary for validating alarms, forensic investigation, and upholding accountability in autonomous response controls. Methods such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and attention heatmaps provide interpretability while identifying phishing attacks, malware activity, or deepfake inconsistencies (Gupta, 2023). The question is how to balance model interpretability and complexity without sacrificing detection accuracy. Furthermore, standardizing XAI approaches across domains is also a research priority.

6.3 Cross-Domain Collaboration for Global Threat Intelligence

As GenAI threats have gone global, defines must also be globalized through international collaboration among public, private, and academic sectors. Interdisciplinary collaboration needs to be set up for data sharing, detection models, threat signatures, and attribution intelligence. Efforts such as the Partnership on AI, the Global Partnership on Artificial Intelligence (GPAI), and CERT communities are leading this cooperative spirit. Real-time collaboration is, however, hindered by data privacy regulations, proprietary solutions, and geopolitics competitions. Secure federated threat intelligence networks, through which encrypted results can be exchanged without exposing raw data, provide a way forward (Kaheh et al., 2023). Adopting mutual ethical guidelines, quick-response partnerships, and mutual simulation training is essential follow-up measure to prevent international GenAI-enabled cyber incidents.

7. Conclusion

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7.1 Synthesis of Key Contributions

This study establishes a comprehensive framework to address the multifaceted threats posed by the misuse of Generative AI. Through a layered analysis, we evaluate cutting-edge defenses, ethical imperatives, and global regulatory trends as of 2025. Technological advancements in detection, such as multimodal fusion achieving 91.2% accuracy on the DFDC-2K24 dataset and temporal inconsistency analysis reaching 94% precision on Celeb-DF v3, now serve as industry benchmarks. Blockchain-based provenance verification (96.4% accuracy) and GAN fingerprint extraction (91% efficacy) have emerged as scalable solutions to authenticate synthetic media. Meanwhile, adaptive countermeasures like graph neural networks (94% F1-score) and adversarial training (90% robustness) effectively neutralize polymorphic AI-generated malware.

The rise of AI Social Media Phishing campaigns, with a 52.1% click-through rate, underscores the critical need for real-time linguistic forensics and behavioral analytics. Ethically, privacy-preserving techniques such as federated learning (e.g., IBM FLGuard) and homomorphic encryption (e.g., Microsoft Azure Confidential Computing) reconcile security demands with GDPR-AI compliance. Globally, regulatory frameworks like the EU AI Act 2025 and India's Digital India Act 2025 set precedents for synthetic content labeling and criminal penalties, while regional efforts like ASEAN's AI Ethics Guidelines address localized risks. Explainable AI (XAI) tools, including SHAP and LIME, enhance transparency, and self-learning systems like Darktrace Antigena v3 exemplify autonomous adaptation to novel threats.

7.2 Call to Action for Multidisciplinary Collaboration

The GenAI threat landscape demands urgent, coordinated action across sectors and borders. **Global governance** must prioritize the establishment of a UN-led AI oversight body by 2026 to enforce cross-border watermarking standards, threat intelligence sharing, and ethical certifications. Harmonizing region-specific regulations—such as the EU AI Act, China's Deep Synthesis Regulation 2.0, and the U.S. NIST GenAI Safety Standards—under a unified risk-tiered framework will mitigate fragmentation.

Sector-specific collaboration between academia and industry is essential to develop open-source detection tools and benchmarks like the CrossModal-DeepFake dataset. Policymakers and technologists must co-design adaptive legislation, such as API throttling mandates for high-risk models, to balance innovation and security. Simultaneously, **public resilience initiatives**, including global literacy campaigns simulating AI-driven phishing attacks (e.g., Google Jigsaw's 2025 modules), should aim to train 1 billion users by 2030.

Immediate steps include adopting the Content Authenticity Initiative (CAI) framework for universal content authentication and funding Global South initiatives, such as India's AI-generated content labeling infrastructure and Africa's ethical data sourcing programs. Ethically, integrating red teaming (e.g., OpenAI's 2024 program) and bias audits (e.g., Google Fairness Indicators) into all GenAI development cycles will institutionalize accountability.

References

[1] Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., Chowdhury, A. R., Christodorescu, M., Datta, A., Feizi, S., Fisher, K., Hashimoto, T., Hendrycks, D., Jha, S., Kang, D., Kerschbaum, F., Mitchell, E., Mitchell, J., Ramzan, Z., & Yang, D. (2023). Identifying and mitigating the security risks of generative AI. *Foundations and Trends® in Privacy and Security, 6*(1), 1–52. https://doi.org/10.1561/3300000041

[2] 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.

[3] Caldelli, R., Becarelli, R., & Amprimo, G. (2021). A survey on digital content provenance. *Journal of Visual Communication and Image Representation, 74*, 103005.
[4] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57.
[5] Chan, C. K. Y., & Colloton, T. (2021). *Generative AI in higher education*. Routledge. https://doi.org/10.4324/9781003459026
[6] Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00411-8
[7] Deshpande, A. S., & Gupta, S. (2023). GenAI in the cyber kill chain: A comprehensive review of risks, threat operative strategies, and adaptive defense approaches. *Conference on ICT in Business Industry & Government*.
[8] Dhoni, P., & Kumar, R. (2023). Synergizing generative AI and cybersecurity: Roles of generative AI entities, companies, agencies, and government in enhancing cybersecurity. *Authorea Preprints*.
[9] Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00425-2
[10] Ferrara, E. (2021). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science, 7*(1), 549–569. https://doi.org/10.1007/s42001-021-00124-w
[11] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1).
[12] Guarnera, L., Giudice, O., & Battiato, S. (2020). Forensics analysis of GAN-generated faces. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 954–965.
[13] Gupta, A. (2023). Navigating the frontier: AI, data privacy, and India's Digital Personal Data Protection Act. *Jus Corpus Law Journal*.
[14] Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access, 11*, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381
[15] Hendrickx, M., Harel, A., & Meidan, Y. (2022). Adaptive cyber defense using reinforcement learning. *ACM Computing Surveys, 55*(3), 1–36.
[16] Kaheh, M., Kholgh, D. K., & Kostakos, P. (2023). Cyber sentinel: Exploring conversational agents in streamlining security tasks with GPT-4. *arXiv Preprint*. arXiv:2304.03456.
[17] Koene, A., Dowthwaite, L., & Seth, S. (2021). Ethics and the governance of AI in Europe: The case for contextualisation. *AI & Society, 36*, 585–597.
[18] Krishnamurthy, O. (2023). Enhancing cybersecurity through generative AI. *International Journal of Universal Science and Engineering*.
[19] Kumar, R., & Subramanian, L. (2020). Stylometric analysis and authorship identification for fake news detection. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 567–579.
[20] Mariani, M., & Dwivedi, Y. K. (2021). Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research, 175*, 114542. https://doi.org/10.1016/j.jbusres.2021.114542
[21] Nguyen, T. T., Nguyen, T. Q., & Nguyen, A. T. (2022). Detecting phishing emails using transformer-based contextual embedding models. *Computers & Security, 113*, 102578.
[22] OpenAI. (2023). GPT-4 system card. Retrieved from https://openai.com/research/gpt-4
[23] Shoaib, M. R., Wang, Z., Ahvanooey, M. T. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. *Conference on Computer and Communications*.

Research Article

[24] Tortora, L. (2021). Beyond discrimination: Generative AI applications and ethical challenges in forensic psychiatry. *Frontiers in Psychiatry, 15*. https://doi.org/10.3389/fpsyt.2021.1346059

[25] Vincent, J. (2021). Facebook's deepfake detection challenge dataset explained. *The Verge*. Retrieved from https://www.theverge.com

[26] Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53.

[27] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Learning rich features for image manipulation detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1053–1061.