IJBITA

International Journal of Business Integrity and Technology Advancements (IJBITA)

Volume – 02
Issue – 02
(2026)

AI-Powered Cyber Threats and the Security of Business Information Systems: A Systematic Literature Review of Emerging Attack Vectors and Defensive Countermeasures

Muhammad Danish Rasheed, Computer Science, San Francisco Bay University
mrasheed725@student.sfbu.edu
Afrasiyab Ali, Computer Science, Business, Stellar Career College
afrasiyabali2003@gmail.com
Muhammad Noman Amjad, Department of Information Technology, Washington University of Science and Technology (WUST)
noman5802@gmail.com

Article Info:
Received: 31 Jul, 2026
Processed: 17 Sep, 2026
Published: 09 Oct, 2026

Keywords: Artificial intelligence, cybersecurity, business information systems, adversarial machine learning, phishing, deepfake, malware, systematic literature review, PRISMA

CC BY SA

ABSTRACT

While the same Generative and Adversarial AI (GAN) techniques that are playing a role in revolutionizing enterprise cybersecurity defense are being leveraged by threat actors to attack business information systems at an unprecedented rate, scale, and sophistication. This paper is a systematic literature review on AI-powered cyber threats that are faced by business information systems that integrates findings from peer-reviewed surveys, IEEE Access and IEEE Communications Surveys & Tutorials publications, cyber standards & policy documents, and up-to-date threat intelligence information in the industry. In the review, the five major AI threat categories identified — AI-enhanced phishing and social engineering, deepfake-enabled fraud and impersonation, AI-generated and polymorphic malware, adversarial attacks on defensive machine learning systems, and automated reconnaissance and exploitation — and empirical data on their prevalence, business impact and growth are analyzed and summarized. The review also covers documented real-world incidents with malicious variants of LLMs like WormGPT and FraudGPT, as well as a detailed examination of defensive AI technologies such as explainable AI, adversarial robust test, behavior anomaly detection, and multimodal deepfake detection, and an overview of the ongoing regulatory framework, including the NIST Adversarial Machine Learning taxonomy and the EU Artificial Intelligence Act. Synthesized survey findings show that 87% of organizations faced an AI-fueled cyberattack last year, AI-generated phishing has now become more than 80% of the observed phishing emails, while just 26% of security professionals are confident detecting sophisticated AI-driven attacks, creating a significant capability-adoption gap in enterprise defenses. The review concludes with a proposed defense in depth approach for business information systems and outlines some future research directions such as adversarial robustness benchmarking, cross organizational threat intelligence sharing, and governability of dual use generative AI.

INTRODUCTION

Business information systems are the integrated networks, enterprise applications, cloud platforms, and data infrastructures that support today's operations of businesses and are under a rapidly changing threat landscape in which artificial intelligence is a force multiplier for both defenders and attackers. The same type of AI-driven security solutions are being used by threat actors to create more sophisticated, quicker, and larger attacks [3], [4]. In fact, industry surveys show that 85% of cybersecurity professionals say they believe the recent uptick in cyberattacks has been driven directly by malicious entities adopting generative AI, which now has faster and smarter ways to exploit business systems; and that a whopping 87% of all organizations suffered at least one AI-based cyberattack in the last year.

Corresponding email: mrasheed725@student.sfbu.edu

This use-cases/dual-use effect is not hypothetical. The practical implications of offensive AI are illustrated in documented attacks, such as the incident involving the CEO of a large advertising holding company, which was defrauded through a multi-channel deceptive campaign that included building trust via messaging platforms that ultimately led to a deepfake voice message extracted sensitive data. In another highly-publicized 2020 case, criminals were able to clone a company executive's voice and make a $243,000 wire transfer. The underground marketplaces now provide purpose-built malicious variants of the popular LMs, such as WormGPT, FraudGPT, and other variants specifically marketed for creating content to deceive users and create functional malware without most of the ethical controls in regular LMs [30], [32].

This growing danger has come with a fragmented academic and practitioner understanding of AI-driven cyber threats, where the research in adversarial machine learning has been focused on model-level attacks (e.g., evasion, poisoning, model extraction) [7], [10] and deepfake research has been in the generation and detection domains [12], [14], with the latter largely remaining out of business risk context, while phishing-detection research has only recently started to be integrated with LLM-generated content [15], [18]. This business-information-systems -based synthesis of these threads from attack taxonomy to empirical business impact to defensive countermeasures and regulatory response has been relatively uncommon. This systematic literature review aims to fill this gap.

The contributions of this review are as follows: (i) a structured taxonomy of five main AI-powered cyber threat categories that are relevant to business information systems; (ii) a quantitative synthesis of the prevalence, growth, and business-impact evidence gathered from academic and industry sources; (iii) an analysis of documented real-world incidents examples that illustrate each threat category; (iv) a review of the emerging regulatory and standards landscape; and (vi) a proposed defense-in-depth framework specific to business information systems. The methodology used for the review is described in Section II. The threat taxonomy is presented in Section III. Empirical prevalence and impact data are summarized in Section IV. The incidents in real life are discussed in section V. Defensive AI Technologies are explored in Section VI. A review of regulatory and governance developments is provided in Section VII. A topic of interest that is addressed in section VIII is the concept of sector-specific risk. A composite severity analysis is detailed in Section IX. The challenges and research gaps are discussed in Section X. Proposal for defense-in-depth framework and conclusion in Section XI.

For the purposes of this review, the term "business information systems" is used to refer to the entire breadth of information technology infrastructure that helps organizations function at their core: enterprise resource planning (ERP) systems, customer relationship management (CRM) systems, financial transaction processing systems, email and collaboration systems, cloud-hosted business applications, and the identity, network, and endpoint infrastructure that connects and secures them. This overall operational focus sets the present review apart from previous surveys focused on general network-security or consumer-related aspects of AI-driven threats, and enables an examination of the intersection of attack taxonomy and adversarial machine learning with the most impactful transaction-processing, identity-verification and financial-authorization workflows within an organization's risk matrix [3], [4].

REVIEW METHODOLOGY

The systematic approach adopted in this review was literature review that followed the guidance of PRISMA 2020 reporting for systematic reviews [26]. Academic surveys about adversarial machine learning, deepfake technology, and AI-driven phishing were gathered from four types of sources: (1) peer-reviewed academic surveys on adversarial machine learning, deepfake technology, and AI-driven phishing published in IEEE Access, IEEE Communications Surveys & Tutorials, ACM Computing Surveys and Applied Artificial Intelligence [3], [7], [18]; (2) standards and policy publications from NIST, ENISA, Europol, and the European Union [9], [19], [21]; (3) peer-reviewed and preprint literature documenting specific AI-powered attack techniques, such as malicious large language model tooling, AI-generated malware, and adversarial attacks on machine learning models [22], [25]; and (4) current industry threat-intelligence and survey reports from cybersecurity vendors and research organizations which were used to source statistics on prevalence, growth and business impact that were not captured in the slower moving academic literature [27], [32]. Searched for the terms "artificial intelligence," "machine learning," or "generative AI" or the terms listed in the brackets as well with "cyber threat," "cyberattack," "phishing," "deepfake," "malware," or "adversarial," prioritized by relevance to business or enterprise information systems, methodological transparency, and publication date; given the rapid development and evolution of generative AI capability during 2020-2026, only publications from this time frame were emphasized. Results were grouped based on the taxonomy used in previous cyberattack reviews based on AI technology [3] and [4] respectively, which was similar to the approach used in this study.

Sources were selected based on the principle of PRISMA-informed practice, whereby only those sources that appeared to be directly relevant to the topic were included; for underlying technical claims, preference was given to peer-reviewed sources, while for

Corresponding email: mrasheed725@student.sfbu.edu

prevalence estimates and cost information, the sources were limited to those from the industry, which can be traced back to a named source, but which has not yet been passed through a slower academic publication system. As often happens in a rapidly changing threat landscape, where industry reported figures could not be independently corroborated, this is clearly indicated in the surrounding text and not presented as fact.

TAXONOMY OF AI-POWERED CYBER THREATS TO BUSINESS INFORMATION SYSTEMS

Based on the literature reviewed, it is possible to categorize the AI-powered cyber threats to business information systems into five main categories, as shown in the diagram in Fig. 1, namely: AI-enhanced phishing and social engineering, deep fake based attacks in the form of fraud and impersonation, AI-generated and polymorphic malware, adversarial attacks on defensive machine learning systems, and automated reconnaissance and exploitation. These are discussed individually in the following.

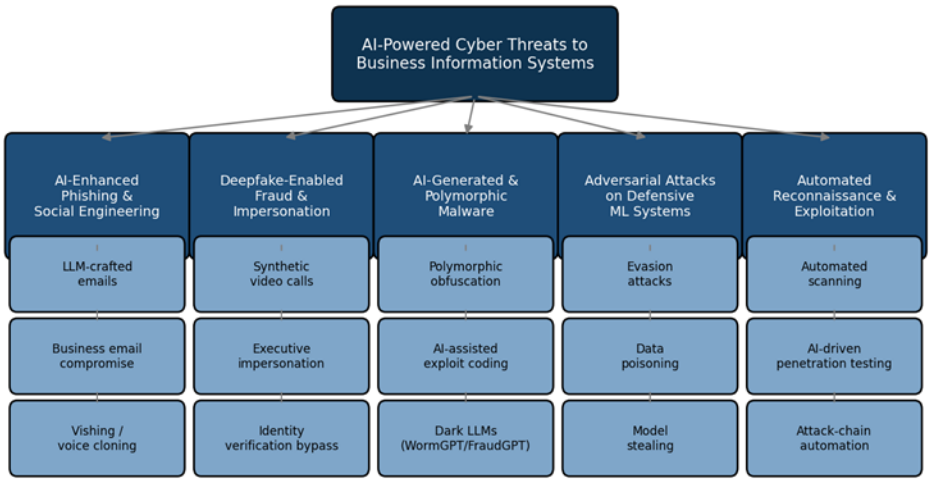


Fig. 1: Taxonomy of AI-powered cyber threats to business information systems, synthesized from the reviewed literature [1], [3], [4], [13], [16], [22], [25].

This taxonomy does not imply that incidents in the real world do not occur in isolation, but instead suggests that many incidents actually involve multiple categories within an incident — for instance, an initial phishing email with AI content is followed by a deepfake voice or video call to complete the deception. This across-categories convergence is an important analytical discovery on its own: new defenses for business information systems that focus on a single-vector threat model are becoming increasingly ill-equipped to respond to multi-vector, AI-orchestrated campaigns that are reported in modern threat intelligence, which is revisited in the proposed defense-in-depth framework of Section XI.

A. AI-ENHANCED PHISHING AND SOCIAL ENGINEERING

The economics of phishing has been completely transformed due to the ability of the large language models to produce grammatically fluent, contextually relevant, and highly personalized messages on a large scale without the typical linguistic errors that constituted a main phishing heuristic before [2], [15]. Academic assessments have substantiated this transformation by showing that it is not just anecdotal, but measurable: research in devising and detecting phishing emails with LLM, for example, shows that AI-generated phishing messages can be as convincing as messages crafted by humans, while studies that have applied pre-trained language models to phishing-email detection have shown that the same LLM defensive opportunity exists for phishing emails [15], [18]. This change is starkly quantified by industry telemetry: the click-through rate of AI-generated phishing emails is estimated to be around 4 times higher than traditional phishing attacks, and AI assistance has now been linked to the majority of newly observed phishing emails. Business email compromise (BEC) — which involves targeting payments via email spoofing as business executives or trusted vendors — has been especially hard hit, with AI-enabled BEC incidents reported to be increasing year over year, as has voice-based phishing (vishing) in the same timeframe.

B. DEEPFAKE-ENABLED FRAUD AND IMPERSONATION

The introduction of new classes of impersonation threat to business information systems, using synthetic media created with GANs, diffusion models, and voice cloning systems, has changed the nature of the threat qualitatively [12] and [13]. Extensive research on deepfake production and detection lists the rapid growth of face-swapping, face-reenactment, and voice-cloning techniques that can convincingly impersonate a real person with very small amounts of source material (a few seconds of voice, in some reported cases [12],

Corresponding email: mrasheed725@student.sfbu.edu

[14]. The risk for the business is real, not hypothetical: Businesses have experienced a dramatic increase in deepfake related frauds in recent years and controlled testing has shown that a limited number of individuals can effectively tell the difference between a deep fake voice and a real one – even when led to actively seek out indicators of manipulation. The documented deepfake cases presented in Section V illustrate the particular danger deepfakes pose to business information systems: to identity-verification workflows, to an executive-authorization process for financial transactions, and to a video-conferencing-based approval procedure.

C. AI-GENERATED AND POLYMORPHIC MALWARE

AI tools are now more and more being used to make and disguise malicious code. There is documented proof-of-concept research that shows that AI systems can generate polymorphic malware that modifies its code on the fly to avoid being signature-based detected, and other research that there can be zero-click self-propagating "AI worms" that can target generative-AI-powered applications directly [24]. In addition to the research-oriented demonstrations of malicious LLMs, an operational “criminal ecosystem” has also been built around specifically engineered malicious LLMs like WormGPT and FraudGPT, which are explicitly marketed in underground forums as tools to create functional malicious code and convincing fraud content without the safety guardrails found in mainstream commercial LLMs [30], [32]. More recent reporting shows that threat actors, too, have started to turn popular commercial LLMs to similar purposes, and have even developed jailbroken versions of popular models for similar work, implying that there is a larger and growing army of malicious models.

D. ADVERSARIAL ATTACKS ON DEFENSIVE MACHINE LEARNING SYSTEMS

These defensive models are becoming attack surfaces as business information systems are increasingly relying on machine-learning defenses, such as intrusion detection systems, malware classifiers, and fraud-detection models. The literature on adversarial machine learning identifies three main attack classes directly relevant to enterprise defenses, namely the evasion attack [7], [8] exploits data inputs designed to be misclassified by the model deployed, without changing its internal structure; the data poisoning attack [23] corrupts training data to make the model's behavior less accurate or more manipulable; and the model-stealing or model-extraction attack [22] reconstructs the functionality of a proprietary model by repeatedly submitting queries to its prediction interface. Extensive experiments of adversarial attacks on network intrusion detection systems directly show that any machine-learning-based defense, no matter how effective, can be easily evaded by strategically designed adversarial perturbations, undermining the trust in AI-based defenses in exactly the places where it is most crucial for organizations to rely on automation [9], [10]. NIST AI 100-2e2025 is the formal taxonomy of adversarial machine learning attacks and mitigations developed by the NIST, which is the most authoritative and up-to-date classification of adversarial machine learning attacks and is gaining traction as a shared language in the reviewed literature [11].

E. AUTOMATED RECONNAISSANCE AND EXPLOITATION

AI is used to automate much of the reconnaissance, vulnerability-scanning and exploitation stages of the cyberattack lifecycle, shaving time off of an often time-consuming manual process performed by the attacker. Examining the AI techniques used in cyberattacks in the context of the cybersecurity kill chain revealed that most known AI techniques were used in the access and penetration phase, with some also being applied in the exploitation, command and control, and reconnaissance phases—showing that while some cybersecurity techniques are being automated, others are not [3]. Especially troubling for business information systems is the fact that the expansive attack surfaces of enterprise networks, often large, heterogeneous and imperfectly inventoried, offer plenty of opportunity for automated scanning plus AI-driven exploit chaining [3], [4].

EMPIRICAL EVIDENCE: SCALE AND BUSINESS IMPACT

In addition to the qualitative taxonomy outlined in Section III, in the industry-survey and threat-intelligence data can be used to quantitatively characterize the rate at which these AI-powered threats are proliferating and how business information systems are impacted in practice. Fig. 2 shows the year-over-year growth of two metrics: the share of phishing emails using AI-generated language, which is estimated at about 82.6% this year, up from about 53.8% last year; and the overall number of global AI-powered cyberattacks, which is projected to more than double to 28 million this year, from around 12 million last year.

Corresponding email: mrasheed725@student.sfbu.edu

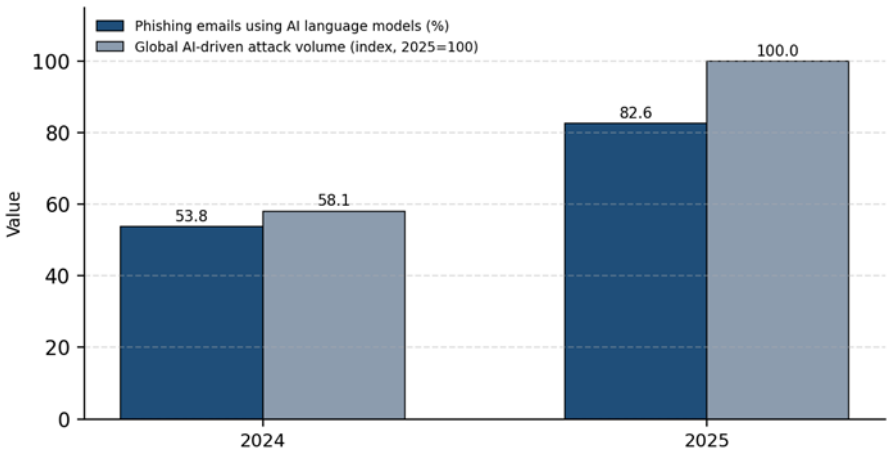


Fig. 2: Year-over-year growth in AI-driven phishing prevalence and global AI-driven attack volume, 2024–2025.

Supporting evidence on exposure and preparedness of organizations is provided in Fig. 3. According to survey data, 87% of organizations have suffered an AI-powered cyberattack over the last 12 months, and 78% of chief information security officers (CISOs) say that AI-powered threats have had a major impact on their organizations, while 85% of cybersecurity professionals credit generative AI for the growing volume of attacks. Perhaps most importantly, just 26% of respondents say they are very confident in being able to spot sophisticated AI-driven attacks, and 60% say they are well prepared in general — but this disparity shows an important difference in how much they trust their technical capabilities when it comes to the most complex AI-based attacks.

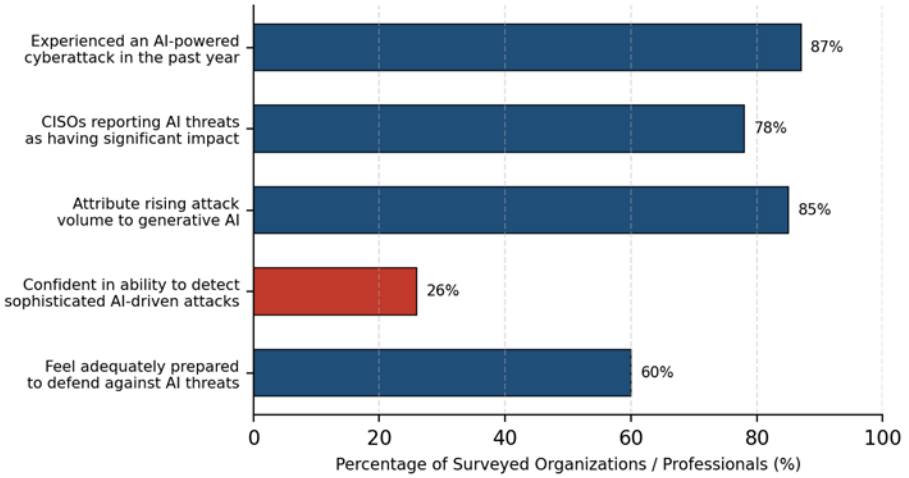


Fig. 3: Organizational exposure to and preparedness for AI-powered cyber threats, compiled from industry survey data.

These numbers come with a lot of money on the line. The cost of an AI-related data breach is roughly $5.72 million, significantly higher than the $4.88 million average cost of a data breach in general, while organizations that leverage AI-powered detection and response tools are able to detect breaches about 43% sooner than those that do not; specifically, 108 days versus 171 days. This imbalance – where AI increases the cost of an attack side breach, while also increasing the value of a defense side detection – highlights the fact that the overall business impact of AI in cybersecurity is very much dependent on the relative speed at which attackers and defenders adopt AI in specific organizations.

When broken down by threat category, there is additional detail that could be useful in prioritizing business risks. Although the average loss per incident is less than that of a large-scale ransomware attack, BEC incidents are far more common, and tend to target accounts-payable and treasury workflows directly, making them more likely to result in a realized loss per incident than many types of attacks, as was found in the severity ranking in Section IX. Although the number of incidents of deepfake enabled fraud was not the highest, it has seen the most reported increase across the year compared to the other categories analyzed in this review, in line with the rapid advancement of generative video and voice-cloning technology in the underlying deepfake-detection literature [12], [14].

Corresponding email: mrasheed725@student.sfbu.edu

REAL-WORLD INCIDENTS

The following documented events are used to show the real life implications of the threat categories listed in Section III. Firstly, the WormGPT case illustrates the operationalization of AI-based malware and phishing on a criminal scale: security researchers from SlashNext discovered the use of WormGPT, an open-source version of the language model GPT-J, for producing convincing and strategically crafted BEC messages aimed at company account managers [30], [32]. Soon after, FraudGPT was released as another tool designed for fraud, and in 2025, security researchers spotted new variants of WormGPT created by jailbreaking popular commercial models such as xAI's Grok and Mistral AI's Mixtral, showcasing how the malicious-LLM world is evolving in response to the security measures being implemented on LLM vendors.

Second, deepfake-enabled executive fraud has produced direct, quantifiable financial losses. An extensive case study from 2020 featured fraudsters using AI voice cloning technology to replicate the voice of an energy company executive based in the UK and tricking a subordinate into wiring $243,000 to a fraud scheme. The 2024 case involved a more complex and nuanced social engineering operation than previous, which involved a multichannel approach, with the emergence of trust via WhatsApp and Microsoft Teams followed by an AI-generated deepfake video call, aligning with the pattern of multichannel attack strategy seen throughout the literature on the topic.

DEFENSIVE AI TECHNOLOGIES AND THE CAPABILITY-ADOPTION GAP

Fig. 4 summarizes the reviewed literature on defensive AI technology classes and the reported adoption of these technologies in enterprises. Based on the reviewed literature, several classes of defensive AI technology were identified that are directly applicable to business information systems and summarized along with the reported uptake of these technologies within enterprises. Explainable AI (XAI) techniques are intended to render the actions of AI-based security systems understandable to human analysts, thereby not only speeding up incident triage but also mitigating threats from malicious attacks focused on exploiting opaque decision boundaries [24]. The most operationally mature defensive AI area is behavioral anomaly detection, which includes user and entity behavior analytics, and has been the focus of years of investment in security information and event management platforms. While adversarial attack methods in academia have become more sophisticated, adversarial robustness testing as a routine practice in enterprise deployments is still relatively underdeveloped [9], [10] and is a method that is worth investigating further. Combining all the features of both visual and audio deepfakes with behavioral characteristics, multimodal deepfake detection has made significant progress in academia but not yet been widely adopted in enterprise use in identity-verification workflows [12], [14]. AI-based security-awareness training falls somewhere in the middle, and relies on the same generative methods used to create offensive phishing content to create realistic, personalized phishing-simulation content to train employees.

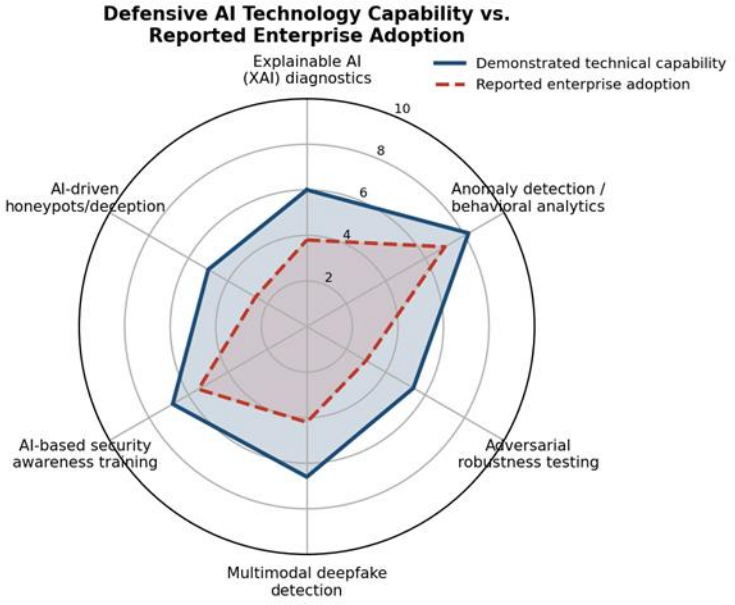


Fig. 4: Defensive AI technology capability versus reported enterprise adoption, synthesized from the reviewed literature.

Corresponding email: mrasheed725@student.sfbu.edu

For all five defensive categories, as shown in Fig. 4, the demonstrated technical capacity for adoption is consistently ahead of the reported enterprise adoption, also reflecting the general finding that only a third of security practitioners (26%) state that they have high confidence in mitigating sophisticated AI-driven attacks despite significant progress in the academic and vendor literature. This disconnect aligns with previous research on impediments to AI adoption in related areas where funding, employee training and organizational engagement rather than the lack of feasible technology are the key barriers to bridging the research capability gap into the practice of defense.

Table 1 also offers a comparison of these defensive AI technology types with the specific threat categories listed in Section III that they are most likely to address, giving an organization a good prioritization reference for defensive investments based on their risk profile.

Defensive Technology	Primary Threat Category Countered	Reported Adoption Maturity
Behavioral anomaly detection / UEBA	AI-generated malware; automated reconnaissance	High — most operationally mature category
AI-based awareness training	AI-enhanced phishing and social engineering	Moderate — growing vendor ecosystem
Multimodal deepfake detection	Deepfake-enabled fraud and impersonation	Low-moderate — limited enterprise integration
Adversarial robustness testing	Adversarial attacks on defensive ML systems	Low — nascent enterprise practice
Explainable AI (XAI) diagnostics	Cross-cutting; supports analyst triage across all categories	Low-moderate — emerging integration

Table 1: Defensive AI Technologies Mapped to Primary Threat Categories Countered

REGULATORY AND GOVERNANCE LANDSCAPE

As the threats from AI-driven cyberattacks continue to evolve at a rapid pace, so have the standards and regulatory work in the realm of business information systems. The National Institute of Standards and Technology (NIST) recently published the AI 100-2e2025: Adverse Machine Learning Attack and Mitigation Taxonomy and Terminology, which formalizes a taxonomy and terminology for adversarial machine learning attacks and mitigations, and has become widely used in both academic and industry papers [11]. AI-related threats have become a more prominent part of the EU threat landscape, as observed in the annual Threat Landscape report released by the European Union Agency for Cybersecurity (ENISA) [10] and others, including an early and influential analysis of the malicious uses and abuses of AI by Europol and the United Nations Interregional Crime and Justice Research Institute (UNICRI) [19], [20] which preemptively identified many of the threat patterns subsequently described in the academic literature. Most importantly, the European Union's Artificial Intelligence Act was adopted in 2024 and includes a risk-based regulatory approach for AI systems that is not cybersecurity specific, but which directly impacts on the governance of dual-use AI systems that can be used in both defensive and offensive cybersecurity, and that will apply to certain classification levels of high-risk AI systems, including certain components of business information systems that are critical for security.

SECTOR-SPECIFIC RISK IN BUSINESS INFORMATION SYSTEMS

AI-powered cyberattacks are not a one-size-fits-all proposition for all businesses. Patterns observed in the reviewed threat-intelligence literature and academic research in the sectors are summarized in Table 2, which accounts for the varying data value, OT exposure, and regulatory complexity among sectors.

Sector	Dominant AI-Powered Threat Pattern	Reported Driver
Manufacturing	Data theft targeting proprietary designs and operational technology systems	High-value intellectual property; converged IT/OT networks
Financial services	AI-enhanced phishing, BEC, and deepfake-enabled payment fraud	Direct monetary transfer capability; high transaction volume
Healthcare	Ransomware and data-exfiltration targeting sensitive patient records	High per-record breach value; legacy system prevalence
Professional / IT services	Supply-chain compromise propagating to downstream business clients	Privileged access to multiple client environments

Table 2: Sector-Specific AI-Powered Cyber Threat Patterns in Business Information Systems

Corresponding email: mrasheed725@student.sfbu.edu

The sector patterns have direct implications for the application of the defense in depth framework proposed in Section XI. Financial services companies benefit most from investing in Layer 1 (human-centric resilience) and Layer 2 (identity and verification hardening) when compared to other sectors, due to their heightened risk of payment fraud, business email compromise and phishing attacks with AI capabilities. Manufacturing companies that have a higher operational-technology exposure, on the other hand, might benefit more from investing in Layer 3 (AI-augmented detection and response) because of the reconnaissance and exploitation risk pattern in converged IT/OT environments. This sector-based prioritization logic reinforces a key theme expressed throughout this review: that the best approach to defense against AI-driven cyber threats is a risk-based application of a common suite of defensive capabilities to an organization's unique threat exposure profile.

COMPOSITE SEVERITY ANALYSIS

Combining the taxonomy, empirical evidence and incidents from the cases discussed in the preceding sections, the composite severity index across the five threat categories identified in Section III has been created, as shown in Fig. 5, based on the relative frequency, financial impact and detection difficulty reported in the literature reviewed. Phishing (at the top with business email compromise) is the most frequently reported type of fraud and is the most likely to be reported to cause a direct and measurable financial loss because of fraudulent transfers. AI-generated and polymorphic malware and deep fake-enabled fraud are right behind, with both showing rapid growth in the past few years and in the case of malware, the obvious difficulties it has in terms of signature-based detection. While automated reconnaissance and adversarial attacks on defensive machine learning systems are serious, they are relatively less prevalent in practice (depending on the degree of technical sophistication that is required for their operation) and somewhat lower in the current composite assessment, though perhaps higher in long-term risk as adversarial machine learning techniques become more advanced and widespread.

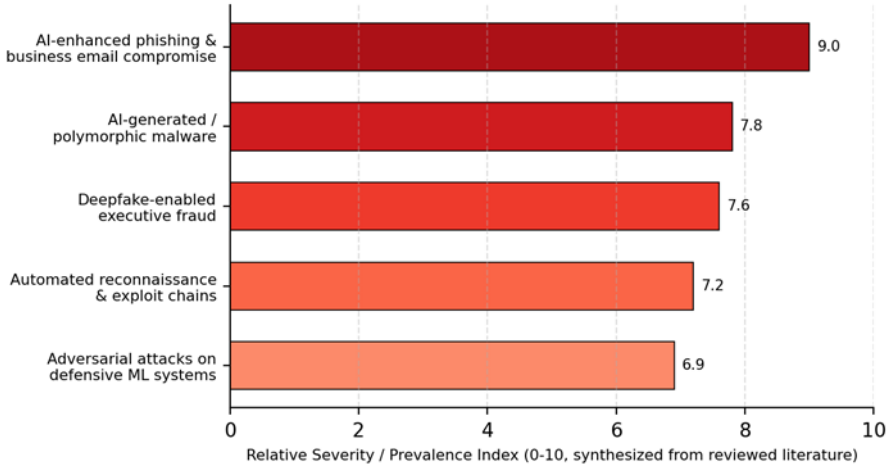


Fig. 5: Composite severity of AI-powered threat categories affecting business information systems, synthesized from the reviewed literature.

CHALLENGES AND RESEARCH GAPS

There are a number of areas within existing literature that warrant further research. First, the adversarial robustness of business-deployed machine learning security models is still in its infancy, and there is not yet a consistent, comparable avenue to evaluate them across vendors and contexts [11]. Second, it is not as widely developed as traditional indicator-of-compromise (IOC) sharing across organizations when it comes to sharing information regarding AI-generated attack content, such as shared repositories of known malicious-LLM outputs or deepfake artifacts. Third, the governance of dual-use generative AI is still developing and represents a policy challenge: existing policy regimes, such as the EU AI Act, have yet to fully mature in how they address the dual-use potential of these LLM capabilities used for creating malicious content when adapted to modest degrees [21]. Fourth, and most importantly, there is still relatively little research that focuses on specific business outcomes and quantifies the actual financial value of certain defensive AI technologies, presenting a significant opportunity to directly link the technique-focused AI research literature to enterprise risk-management decision-making [9], [10].

Fifth, methodological variation within the reviewed evidence base is itself a gap in evidence which can be explicitly identified. Academic adversarial-machine-learning (AML) studies tend to run under controlled and benchmark-dataset conditions that may not reflect

Corresponding email: mrasheed725@student.sfbu.edu

the noisy and heterogeneous conditions of production business information systems, and statistics from industry surveys give a good idea of the prevalence and impact in industry, but they are less transparent and less subject to peer review than academic papers [9], [10]. The ability to bridge this gap through industry-academic collaborative research, with rigorous experimental design and access to real-world production telemetry, is a value-added opportunity for future work, and would help address the current uncertainty regarding the fidelity of adversarial attack success rates from the lab to production system compromise.

PROPOSED DEFENSE-IN-DEPTH FRAMEWORK FOR BUSINESS INFORMATION SYSTEMS

Combining the above evidence gathered, this paper presents a five layers defense-in-depth model for business information systems against AI-driven cyber threats.

LAYER 1. HUMAN-CENTRIC RESILIENCE

Implement AI-driven, ongoing security awareness training that adapts to the latest sophistication of AI-generated phishing and deepfake content, and attack the human-layer vulnerability that the threat category rated most severe in Section IX [15] and [18] exploits.

LAYER 2. IDENTITY AND VERIFICATION HARDENING

Enforce out-of-band, multi-factor verification processes for high-value transactions and executive-authorization processes that are specifically resistant to deepfakes, based on the recorded executive fraud attacks detailed in Section V [12], [14].

LAYER 3. AI-AUGMENTED DETECTION AND RESPONSE

Implement behavioral-anomaly-detection and AI-based threat-detection tools, as the adoption is somewhat mature, compared to other tools listed in Fig. 4, and invest in the shorter detection/response time those tools have proven to achieve.

LAYER 4. ADVERSARIAL ROBUSTNESS ASSURANCE

Test subject business-critical machine-learning-based security models against adversarial robustness before deployment and continuously, using the taxonomy published by the NIST AI 100-2e2025 as a common evaluation vocabulary to cover the under-investigated adversarial-attack surface identified in Section VI [9], [11].

LAYER 5. GOVERNANCE AND THREAT-INTELLIGENCE INTEGRATION

Implement formal governance processes for the use of dual-use generative AI tools within the organization, keeping pace with new and existing regulations, including the EU AI Act, and engage in inter-organizational threat intelligence sharing in support of threat intelligence in AI-generated attack content, addressing the research gap identified in Section X [21].

CONCLUSION

AI is playing a game-changer role on both sides of the cybersecurity game and business information systems are right in the middle of the game. The systematic literature review has identified these five key categories of threat: AI-enabled Phishing and Social Engineering, Deepfakes in Fraud, AI-generated and Polymorphic Malware, Adversarial Attacks against Defensive ML Systems and Automated Reconnaissance and Exploitation, and quantified evidence demonstrating that these are not hypothetical threats, but are real, pervasive problems that 87% of organizations report having faced in their operates. Meanwhile, the review revealed that, despite the impressive technical performance of defensive AI technologies, there has been a longstanding and significant disconnect between the technologies' usability in the enterprise and their quality of adversarial robustness assurance, as attackers' sophistication grows. The proposed five-layer defense-in-depth approach provides a ladder for business information systems to bring this gap closer, running from human resilience, identity verification, AI-augmented detection, adversarial robustness assurance to governance. The need for continued investment in adding standardized robustness benchmarking, cross-organizational threat-intelligence sharing and dual-use AI governance will be crucial as both sides of the offense-defense landscape continue to rapidly evolve with the capabilities of generative and adversarial AI.

In the end, the evidence gathered in this review suggests a more general strategic takeaway: When organizations view AI-powered cybersecurity threats as a technical challenge that can be solved with a point solution acquisition, they will inevitably be playing catch-up to a cyberattackers' ecosystem that is also organized around quick and iterative adoption of the latest generative and adversarial AI capabilities. The multi-vector convergence described in Section III, the continuous capability adoption gap described in Section VI, and the still-evolving regulatory landscape described in Section VII, all point to an approach for building durable resilience that is less one-off defensive technology and more about the capability of an organization to continuously, and continuously, adapt across the human, technical and governance domains of the proposed framework.

Corresponding email: mrasheed725@student.sfbu.edu

REFERENCES

[1] C. Nobles, "The weaponization of artificial intelligence in cybersecurity: A systematic review," Procedia Computer Science, vol. 239, pp. 547–555, 2024.
[2] P. Nespoli, S. B. Molina, P. Beltrán-López, and F. G. Mármol, "Tackling cyberattacks through AI-based reactive systems: A holistic review and future vision," IEEE Communications Surveys & Tutorials, vol. 28, pp. 1557–1596, 2025.
[3] J. Auccaylle, E. Infantes, and W. Ticona, "Analysis of AI-powered cyberattack strategies and their impact on security controls: A systematic review," in Proc. Computer Science On-line Conf., Cham, Switzerland: Springer Nature Switzerland, Apr. 2026, pp. 161–176.
[4] U. Imtiaz, "Dynamic security certification framework for evolving distributed architectures," in Data Science and Security, S. Shukla, H. Sayama, K. Tiwari, J. P. George, and J. V. Kureethara, Eds., Lecture Notes in Networks and Systems, vol. 1947. Cham, Switzerland: Springer, 2026, doi: 10.1007/978-3-032-24360-7_32.
[5] M. N. Amjad, "Machine learning for renewable energy forecasting and smart grid optimization," Journal of Engineering and Computational Intelligence Review, vol. 2, no. 2, pp. 130–145, 2024.
[6] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The emerging threat of AI-driven cyber attacks: A review," Applied Artificial Intelligence, vol. 36, no. 1, Art. no. 2037254, 2022.
[7] F. Amin, U. Imtiaz, and A. Khan, "FALCON-Guard: A lightweight explainable framework for real-time cyber threat detection and adaptive risk mitigation in intelligent driving networks," Multidisciplinary Research in Computing Information Systems, vol. 5, no. 12, pp. 1223–1235, 2025.
[8] M. N. Amjad, "Strategic engineering management for sustainable and resilient US industrial infrastructure," International Journal of Business Integrity and Technology Advancements, vol. 1, no. 2, pp. 8–20, 2025.
[9] A. Kumar and K. K. S. Nair, "The impact of cloud-based business analytics on operational efficiency in US small and medium-sized firms," Spanish Journal of Innovation and Integrity, vol. 49, pp. 303–318, 2025.
[10] R. Singh, "Smart prediction of molecular behavior in liquid chromatography for better compound detection," International Journal of Clinical Research and Medical Sciences, vol. 1, no. 1, 2026, doi: 10.67231/v3jyx227.
[11] U. Imtiaz, "Ghost signals: Ethical RF adversarial testing of consumer alarm ecosystems," in Proc. Int. Conf. Data Science, Computation and Security, Cham, Switzerland: Springer Nature Switzerland, Nov. 2025, pp. 240–253.
[12] M. Chaudhari, "Epoxy barrier coatings reinforced with solidified-waste microfillers for aggressive acids," Genetics and Molecular Research, vol. 25, no. 3, Art. no. 3020, Jul. 2026, doi: 10.4238/96y1s945.
[13] E. Wexford, "Beyond the perimeter: The impact of zero trust security on cybersecurity resilience in US organizations," American Journal of Multidisciplinary Knowledge Insights, vol. 5, no. 2, pp. 42–51, 2026.
[14] A. Khan, J. Iqbal, and D. Khan, "Artificial intelligence for financial data protection and business cyber resilience," Journal of Computational Systems & Engineering Insights, vol. 4, no. 2, pp. 30–41, 2026.
[15] M. A. Jawed, "Managed aquifer recharge with advanced treated municipal effluent: Modelling the impact on groundwater chemistry and native microbial ecology," International Journal of Research & Technology, vol. 12, no. 3, pp. 128–138, 2024.
[16] R. Singh, "Pharmacognosy-to-candidate: Protease-directed prioritization of plant-derived glycosides via minimal semi-synthetic optimization," International Journal of Clinical Research and Medical Sciences, vol. 1, no. 2, 2026, doi: 10.67231/rxn6m387.
[17] M. Chaudhari, "Hydrophilic nanofiber barrier coatings for aluminum: Waterborne crosslinking, microstructure–property links, and EIS-based protection," Genetics and Molecular Research, vol. 25, no. 3, Art. no. 3017, Jul. 27, 2026, doi: 10.4238/rhd2k915.
[18] M. Z. Islam, I. Ahmed, R. Parveen, S. T. H. Rimon, and J. I. Janjua, "Revolutionizing healthcare for critical diseases with AI and enhancing security via low-latency edge intelligence," in Proc. 2025 Int. Conf. AI-Driven STEM Education and Learning Technologies (AISTEMEDU), Piscataway, NJ, USA: IEEE, Dec. 2025, pp. 1–6.
[19] R. Bhattacharya, J. Mukherjee, S. Roy, M. T. Rana, and R. Parveen, "Eco-crypto dynamics: Cointegration of green and non-green cryptocurrencies for sustainable investing," Advances in Consumer Research, vol. 2, no. 3, 2025.
[20] D. Safiullah, A. A. Khan, N. Ahmad, M. Danish, and S. Rana, "Integrating artificial intelligence and soil ecology to enhance plant resilience under environmental stress," ASSAJ, vol. 4, no. 2, pp. 3508–3516, 2025.
[21] S. M. Ali, R. S. Ahmad, A. Hannan, Shafia, M. S. Khan, and S. Ahmad, "A deep learning and digital twin architecture for intelligent demand response in smart grids," in Proc. 2026 Int. Conf. Integrated Intelligence and Cognitive Engineering (ICIICE), Dubai, United Arab Emirates, 2026, pp. 1–6, doi: 10.1109/ICIICE69672.2026.11565446.

Corresponding email: mrasheed725@student.sfbu.edu

[22] U. Iqbal and Y. Bhutto, "Digital transformation through artificial intelligence and advance business analytic in American operational management," Journal of Theoretical and Applied Econometrics, vol. 3, no. 1, pp. 37–50, 2026.
[23] L. H. Taylor, "Nanotechnology in biomedical engineering: Advances in drug delivery, diagnostics, and cancer therapy in the United States," American Journal of Multidisciplinary Knowledge Insights, vol. 4, no. 2, pp. 1–8, 2025.
[24] U. A. Kithran, Q. Zahra, H. Zafar, R. H. Ahmed, and H. Abbas, "AI-based route optimization for sustainable logistics," Spectrum of Engineering Sciences, vol. 3, no. 10, pp. 2057–2072, 2025.
[25] M. R. Islam, I. A. Badhan, M. H. Rahman, and M. N. Hasnain, "The economics of global renewable energy supply chains: Opportunities and vulnerabilities for the U.S.," International Journal of AI, Engineering and Management Studies (IJAIEMS), vol. 1, no. 1, pp. 108–136, 2026.
[26] A. S. Anwar, S. K. Bhowmik, R. B. Kadir, M. U. Haque, and S. Rahman, "Challenges in implementing machine learning-driven IoT solutions in semiconductor design and wireless communication system," International Journal of Recent Innovations in Trends in Computing and Communication, vol. 12, pp. 872–889, 2024.
[27] A. Hannan, Y. Hussain, K. Ullah, M. S. Khan, H. Saleem, and B. Azeem, "Intelligent short-term load forecasting for renewable power grids using attention-based hybrid deep learning," in Proc. 2026 Int. Conf. Integrated Intelligence and Cognitive Engineering (ICIICE), Dubai, United Arab Emirates, 2026, pp. 1–6, doi: 10.1109/ICIICE69672.2026.11565331.
[28] U. A. Kithran, S. Ilyas, K. Amber, H. Abbas, and S. Zahid, "AI-based collision prediction and prevention systems in urban traffic," Spectrum of Engineering Sciences, vol. 3, no. 12, pp. 1907–1920, 2025.
[29] R. H. Ahmed, M. Sultana, M. Imran, N. Ahmad, and M. D. Rasheed, "A decentralised multi-agent reinforcement learning framework for fault-tolerant swarm coordination under partial observability," in Proc. 2026 2nd Int. Symp. AI-Driven Engineering Systems (ISADES), Mbale, Uganda, 2026, pp. 1–8, doi: 10.1109/ISADES69945.2026.11608224.
[30] U. Iqbal, "AI-powered supplier risk intelligence: Predicting financial and geopolitical supply chain disruptions in U.S. critical industries," Journal of Engineering and Computational Intelligence Review, vol. 3, no. 2, pp. 173–193, 2025.
[31] M. S. Uddin, A. Khan, I. Alim, and D. Khan, "The evolution of cybersecurity in the digital age: Contemporary threats, technologies, and future directions," Journal of Computational Systems & Engineering Insights, vol. 4, no. 2, pp. 12–29, 2026.
[32] N. Gael, "Cybersecurity risk management in smart IoT ecosystems using artificial intelligence," International Journal of Business Integrity and Technology Advancements, vol. 2, no. 1, pp. 1–12, 2026.

Corresponding email: mrasheed725@student.sfbu.edu