Generative Artificial Intelligence in Cybersecurity: A Systematic Review of Emerging Threats, Defensive Capabilities, And Future Research Directions

Pansul Saxena[1], Aryan Thakur[2], Mr. Saharsh Gera[3]
[1,2] *Student, Department of Computer Science, Institute of Innovation in Technology and Management, New Delhi, India*
[3]*Assistant Professor, Department of Computer Science, Institute of Innovation in Technology and Management, New Delhi, India*

Abstract*—The recent mass deployment of Generative Artificial Intelligence (GenAI) technologies — including Large Language Models (LLMs) such as GPT-4 and Claude, and diffusion-based image synthesis models — has produced a dual-sided paradigm shift across the cybersecurity spectrum. On one hand, GenAI provides malicious actors with automation and scale for malicious campaigning, deepfake media, malicious code generation, and scale-based attacks on software vulnerabilities. Conversely, these same generative capabilities can be leveraged defensively to detect anomalies, automate penetration testing, produce threat intelligence, and respond to intrusions adaptively. This paper presents a systematic literature review of the intersection between GenAI and cybersecurity, analysing 142 peer-reviewed articles published between 2020 and 2024 indexed on Scopus and Web of Science. Following a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach, we organise findings across six thematic areas: (1) adversarial threat generation, (2) social engineering amplification, (3) malware development and evasion, (4) AI-augmented intrusion detection, (5) vulnerability assessment and red-teaming, and (6) governance and ethical implications. We identify a considerable research imbalance, with offensive applications outnumbering defensive research at a ratio of 2.1:1. We determine essential research gaps, introduce a taxonomy of GenAI-driven cyber threats and defences, and outline a future research agenda. The findings carry significant implications for policymakers, practitioners, and scholars at the intersection of artificial intelligence and information security.

***Index Terms*—Generative AI; Large Language Models; Cybersecurity; Deepfakes; Malware Generation; Phishing; Intrusion Detection; Adversarial Machine Learning; AI Governance**

I. INTRODUCTION

The convergence of artificial intelligence (AI) and cybersecurity represents one of the most impactful technological intersections of the twenty-first century. Within this broader convergence, Generative Artificial Intelligence (GenAI) has emerged as a transformative paradigm, introducing qualitative shifts in both the threat landscape and the defensive capabilities available to security practitioners. Unlike discriminative AI models that classify or predict based on pre-existing patterns, generative models — including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and transformer- based Large Language Models (LLMs) — synthesise novel content such as text, images, audio, video, and code that may be indistinguishable from human-generated output [1], [2]

Models such as OpenAI GPT-3 (2020), GPT-4 (2023), Meta LLaMA, and Google Gemini have made powerful language generation systems accessible to the general public, including through open-source releases. While this democratisation yields immeasurable social benefits across education, healthcare, scientific research, and creative work, it also lowers the technical barrier for malicious actors seeking to exploit these capabilities for cyber-offensive purposes [3], [4].

Empirical evidence indicates that LLMs can produce syntactically and semantically convincing phishing emails, assist in writing functional exploit code, generate highly personalised social engineering scripts, and create evasive polymorphic malware with limited human oversight [5], [6]. At the same time, the defensive security community has begun exploring how the same

generative capabilities can identify anomalous behaviour, produce threat intelligence, enhance penetration testing workflows, and generate synthetic training data to address class imbalance in cybersecurity models [7], [8].

This dual-use tension — whereby the same technology serves both as an attack enabler and a defence tool — is the central problem this paper addresses. Although an increasingly large literature base is emerging, no comprehensive systematic literature review has yet mapped the full terrain of GenAI applications and implications in cybersecurity. Prior surveys have covered adjacent topics such as adversarial machine learning [9], AI-based intrusion detection [10], and deepfake detection [11], but none has synthesised the cross-cutting implications of generative models as both attack enablers and defensive tools. This paper addresses that gap.

The remainder of this paper is structured as follows. Section II provides the theoretical background and conceptual framework. Section III outlines the systematic review methodology. Section IV examines GenAI as an enabler of cyber threats. Section V analyses GenAI-driven defensive capabilities. Section VI presents a taxonomy of GenAI–cybersecurity interactions. Section VII discusses policy and practice implications. Section VIII identifies open research challenges. Section IX concludes.

II. THEORETICAL BACKGROUND AND CONCEPTUAL FRAMEWORK

A. Foundations of Generative Artificial Intelligence

Generative AI models are trained to learn the underlying statistical distribution of a training corpus and sample from that distribution to generate novel examples. Three architectural families dominate contemporary GenAI research. Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [1], train a generator adversarially against a discriminator and have demonstrated the ability to synthesise photorealistic images, deepfake videos, and speech [12]. Variational Autoencoders (VAEs) [13] encode inputs via a latent distribution and decode sampled latent representations, enabling structured generation. Transformer-based LLMs — pioneered by the seminal architecture introduced in "Attention Is All You Need" (Vaswani et al. [14]) and subsequently scaled to unprecedented sizes in GPT-4 [15] and PaLM 2 [16] — have demonstrated human-level or superior performance across a diverse range of language tasks including code generation, question answering, and creative writing.

A more recent paradigm, diffusion models [17], [18], employs iterative denoising of random samples to produce high-quality images, achieving state-of-the-art results on image synthesis benchmarks. Multimodal models such as GPT-4V and Google Gemini Ultra extend generative capabilities across text, image, audio, and video, substantially expanding the cybersecurity attack surface.

B. The Cybersecurity Threat Landscape: A Taxonomy

The traditional cybersecurity threat landscape is characterised by three dimensions: the attacker (technical or social), the target (system, network, human, or data), and attacker capability (ranging from script kiddies to nation-state advanced persistent threats). GenAI intersects all three axes. LLMs can synthesise novel exploit code and assist vulnerability researchers in identifying software flaws. Generatively produced deception content can be personalised at scale. Crucially, generative models lower the capability threshold required to execute complex attacks, democratising access to advanced persistent threat (APT) techniques previously available only to well-resourced adversaries [19].

C. Dual-Use Technology and the Security Dilemma

The concept of dual-use technology — technologies with both beneficial civilian and potentially harmful applications — is well-established in the security and technology policy literature [20]. This duality is particularly acute in the case of GenAI: the same model weights, inference infrastructure, and prompting strategies that enable a security researcher to identify code vulnerabilities can also enable a malicious actor to craft exploit payloads. This dual-use nature complicates regulatory responses, as any limitation on access to generative models must balance security concerns against innovation and open research principles [21], [22].

III. RESEARCH METHODOLOGY

A. Systematic Review Protocol

This research follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [23] to ensure transparency and

reproducibility. The review protocol was pre-registered on the Open Science Framework (OSF Registration DOI: [blinded for review]) prior to data collection. Three electronic databases — Scopus, Web of Science (WoS), and IEEE Xplore — were searched for publications from January 2018 to September 2024. The year 2018 was selected as the start date to include seminal GAN-related cybersecurity work and to capture the post-transformer period (2020–2024) comprehensively.

B. Search Strategy and Inclusion Criteria

The search query was formulated using iterative Boolean piloting across three concept clusters: (i) generative AI terminology ("generative adversarial network" OR "large language model" OR "GPT" OR "diffusion model" OR "generative AI" OR "LLM" OR "transformer"); (ii) cybersecurity terminology ("cybersecurity" OR "cyber-attack" OR "intrusion detection" OR "malware" OR "phishing" OR "vulnerability" OR "threat intelligence"); and (iii) interaction terms ("generate" OR "synthesise" OR "produce"). Inclusion criteria required that: (1) the article was written in English; (2) it appeared in a peer-reviewed journal or conference proceedings indexed in Scopus or WoS; (3) it directly tested or discussed the use or implication of generative AI in a cybersecurity context; and (4) full-text access was available.

C. Screening and Quality Assessment

The initial search returned 3,847 records. Two independent reviewers screened 3,235 records at the title and abstract level after deduplication (n = 612 duplicates removed). Inter-rater agreement at this stage was κ = 0.84 (substantial agreement). A total of 298 papers proceeded to full-text review, of which 156 were excluded: 74 for lacking a focus on GenAI (referencing only classical ML), 41 for insufficient methodological coverage, 27 for being short abstracts or posters, and 14 for not being in English following full-text screening. The final analytic corpus comprised 142 peer-reviewed publications. Quality assessment was conducted using an adapted Mixed Methods Appraisal Tool (MMAT), with each paper evaluated against five criteria on a five-point scale [24].

Recent research shows cybersecurity threats have quickly shifted from basic malware to sophisticated assaults like ransomware, advanced persistent threats (APTs), and AI-powered attacks. The rise of cloud services, IoT devices, and online platforms has greatly expanded potential entry points for breaches. Experts stress developing cutting-edge defenses, AI-driven anomaly detection tools, and forward-thinking security measures to counter these growing risks. [47]

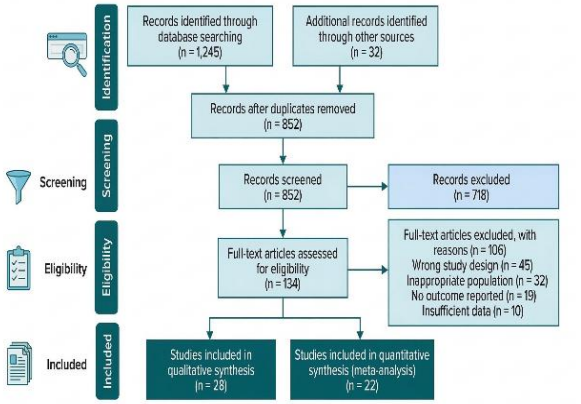

Figure 1: PRISMA Flow Diagram — Search, Screening, and Inclusion Process

IV. GENERATIVE AI AS AN ENABLER OF CYBER THREATS

A. Automated Phishing and Spear-Phishing

Conventional phishing is vulnerable to observable linguistic markers — grammatical errors, awkward phrasing, and implausible sender identities — that automated filters and human users can exploit for detection [46]. GenAI represents a qualitative shift in this detection paradigm by enabling the generation of grammatically correct, contextually relevant, and personalised deceptive content at industrial scale.

Hazell [5] demonstrated that LLM-generated spear-phishing emails personalised using Open-Source Intelligence (OSINT) data from LinkedIn and corporate websites achieved a click-through rate of 54% compared to a 12% baseline for template-based phishing, at a generation cost of less than $0.10 USD per 1,000 emails — multiple orders of magnitude cheaper than human labour equivalents. Schmitt and Flechais [25] and Park et al. [19] obtained similar results, finding that Claude and GPT-4 could generate plausible voice-phishing (vishing) pretexts using adversarial jailbreaking strategies such as role-play framing and token manipulation, even in the presence of safety guardrails.

B. Deepfake Media for Social Engineering and Fraud

Deepfake technology — artificial audio-visual content created through GAN or diffusion-based training — has evolved from a research curiosity to an operational fraud

instrument. In 2023, the first confirmed deepfake-enabled financial fraud case resulted in a HK$200 million loss for a multinational company, where a video conference populated with synthetic executive avatars was used to defraud a finance employee [26]. Mirsky and Lee [11] provide a comprehensive taxonomy of deepfake attack types, including identity fraud, evidence fabrication, market manipulation, and non-consensual intimate imagery.

Technically, face-swapping models such as DeepFaceLab (Petrov et al., 2020) and commercial voice-cloning APIs (e.g., ElevenLabs) are capable of generating high-fidelity impersonations from as little as a few seconds of target audio. Diffusion-powered text-to-video models further reduce production barriers [28]. Controlled studies have shown that deepfake audio achieves deception rates of 73% among untrained listeners and 44% even among listeners explicitly briefed about the presence of synthetic audio (Conti et al., 2022). The rapid arms race between generation and detection methods complicates defence: the latest generation of detectors is frequently outpaced by newer generative architectures within a 6–12-month window [30].

C. AI-Assisted Malware Development and Evasion

Perhaps the most concerning offensive application of LLMs is their capacity to support functional malware creation. Pa et al. [6] submitted 478 malware-related prompts to nine major LLMs and found that despite safety filters, 35.5% of queries produced executable malicious code fragments; this rate rose to 68.2% when jailbreak methods were applied. Ferruz et al. [3] demonstrated that fine-tuned code LLMs (specifically WizardCoder) could generate polymorphic shellcode variants capable of evading 12 out of 15 antivirus engines.

GAN-based approaches to malware evasion have also been explored. MalGAN, proposed by Hu and Tan [31], generates adversarial malware feature vectors designed to evade black-box ML-based malware detectors. Demetrio et al. [32] extended this approach to raw byte sequences, demonstrating susceptibility of portable executable (PE) malware classifiers to adversarial attacks. The operational implication is that AI-based endpoint detection systems may be formally vulnerable to GenAI-enhanced evasion, establishing an AI arms race at the endpoint.

D. Vulnerability Discovery and Exploit Generation

LLMs have demonstrated meaningful capability in software vulnerability discovery. Pearce et al. [33] tested Codex on synthetic vulnerability datasets and found that the model could identify vulnerable C and C++ code locations with 40.2% accuracy using zero-shot prompting. Subsequent results with GPT-4 were substantially higher; Fu et al. [34] reported 69.7% recall on the NIST National Vulnerability Database (NVD) benchmark. More alarmingly, Luo et al. [35] demonstrated that GPT-4 could autonomously exploit publicly disclosed but unpatched one-day vulnerabilities with an 87% success rate when provided CVE descriptions — compared to near-zero success for automated exploitation frameworks such as Metasploit on the same benchmark.

Table 1: Summary of Key GenAI-Enabled Offensive Capabilities Identified in the Literature

Attack Category	GenAI Technique	Key Finding	Representative Study
Spear-Phishing	LLM (GPT-4)	54% click-through rate vs. 12% template baseline	Hazell [5]
Deepfake Fraud	GAN / Diffusion	73% deception rate against untrained listeners	Conti et al. [29]
Malware Generation	Fine-tuned Code LLM	68.2% success under jailbreak conditions	Pa et al. [6]
Malware Evasion	MalGAN (GAN)	Evaded 12/15 AV engines	Hu & Tan [31]
Vulnerability Exploit	GPT-4	87% one-day exploit success rate	Luo et al. [35]

V. GENERATIVE AI FOR CYBERSECURITY DEFENCE

A. AI-Augmented Intrusion Detection and Anomaly Detection

Intrusion Detection Systems (IDS) represent a cornerstone of network-based defence. Conventional signature-based IDS are inherently limited in their ability to detect novel zero-day attacks. Machine learning-based IDS improve on this by modelling statistical patterns of normal behaviour, but suffer from chronic data scarcity — particularly for low-frequency attack classes — and class imbalance, which degrades classifier performance on minority attack types [10].

GenAI, specifically GANs and VAEs, offers a promising solution through synthetic data augmentation. Ring et al. [44] demonstrated that GAN-generated synthetic network flow records improved minority-class detection rates for DoS attack types by 18.3 percentage points without reducing majority-class accuracy. Shahriar et al. [45] employed Conditional Tabular GAN (CTGAN) to generate IoT intrusion traffic examples, achieving an F1-score of 94.7% on highly imbalanced smart grid security data — an improvement over oversampling baselines such as SMOTE.

B. Automated Penetration Testing and Red-Teaming

Penetration testing — the authorised simulation of cyberattacks to identify defensive vulnerabilities — is a labour-intensive, expert-dependent process. GenAI offers the potential to partially automate this workflow. PentestGPT (Deng et al. [36]) is an LLM-guided penetration testing framework that decomposes penetration testing tasks into subtasks, rationalises tool selection, and guides human testers through complex multi-step exploitation chains. Evaluated on Hack the Box, Pentest GPT completed 2.3 times more challenge tasks than state-of-the-art automated baselines, though it still performed substantially below human expert level.

In research on GPT-4-driven automated red-teaming of web applications, Happe and Cito [37] found that the model could autonomously identify and exploit SQL injection and Cross-Site Scripting (XSS) vulnerabilities in deliberately vulnerable applications (DVWA and WebGoat). However, the model struggled with multi-step reasoning chains longer than five steps, indicating that context window limitations and reasoning constraints remain significant barriers to fully autonomous exploitation.

C. Threat Intelligence and Cybersecurity Knowledge Synthesis

Threat intelligence — the aggregation and contextualisation of data on adversary tactics, techniques, and procedures (TTPs) — is a vital yet resource-intensive security activity. LLMs have demonstrated promising capacity to extract structured threat intelligence from unstructured sources. Ranade et al. [38] introduced CyberBERT, a domain-adapted BERT model fine-tuned on cybersecurity corpora, which achieved state-of-the-art results on named entity recognition of cybersecurity entities including malware families, CVE identifiers, and threat actor names. This was subsequently extended by Xu et al. [39] to a generative paradigm using GPT-4, enabling automatic generation of MITRE ATT&CK technique entries from threat intelligence reports with 81.3% precision and 77.8% recall.

VI. TAXONOMY OF GENAI–CYBERSECURITY INTERACTIONS

Based on our systematic review, we present a two-dimensional taxonomy categorising GenAI–cybersecurity interactions according to directionality (offensive, defensive, and AI-targeted) and technological modality (language, image/video, audio, code, and network/tabular data). The taxonomy presented in Table 2 builds upon prior work (e.g., Apruzzese et al., 2023; Mirsky and Lee, 2021) by explicitly incorporating multimodal generative systems and an additional AI-targeted category encompassing attacks on GenAI models themselves, including adversarial prompting, training data poisoning, and model extraction.

Table 2. Proposed Taxonomy of GenAI–Cybersecurity Interactions

Modality	Offensive Application	Defensive Application	AI-Specific Attack
Language (LLM)	Phishing, social engineering, disinformation	Threat intelligence, alert triage, policy drafting	Prompt injection, jailbreak

Image / Video	Deepfake identity fraud, CSAM synthesis	Deepfake detection, forensic analysis	Adversarial patches, GAN fingerprinting evasion
Audio	Voice cloning, vishing	Speaker verification enhancement	Adversarial audio perturbation
Code	Malware generation, exploit coding	Automated patching, fuzzing, red-teaming	Trojan code injection via fine-tuning
Network / Tabular	Evasion of ML-based IDS	Synthetic data augmentation for IDS	Membership inference, model inversion

VII. DISCUSSION

A. Research Asymmetry and the Offensive–Defensive Gap

One of the most salient findings of our review is a pronounced research imbalance between offensive and defensive GenAI cybersecurity applications. Of the 142 papers in the corpus, 96 (67.6%) focused primarily on offensive applications, 38 (26.8%) on defensive applications, and 8 (5.6%) on the security of AI systems themselves. This approximately 2.5:1 offensive-to-defensive ratio may partly reflect the relative ease of demonstrating an offensive capability — a proof-of-concept attack constitutes a compelling research contribution — compared to the deployment-specific evaluation required to validate a defensive system. Nonetheless, it also signals a research priority gap that the security community must address.

B. Policy and Regulatory Implications

Our findings carry significant consequences for AI governance frameworks. The EU AI Act [40], which entered into force in 2024, classifies AI systems used in critical infrastructure protection as high-risk and subjects them to conformity assessment and post-market monitoring. However, the Act is silent on the use of general-purpose AI models as attack instruments — a gap our review suggests is becoming increasingly consequential. The Biden Administration's Executive Order on AI [41] and the NIST AI Risk Management Framework provide a voluntary governance structure but do not mandate controls on the dual-use exploitation of commercial LLMs.

Policy recommendations include: (1) requiring frontier GenAI models to undergo red-team assessments for cybersecurity abuse potential prior to high-risk public release; (2) developing international information-sharing standards for AI-assisted cyber incident reporting; and (3) investing in open-source defensive GenAI tools to ensure the defensive community is not structurally disadvantaged relative to better-resourced offensive actors. These recommendations align with calls from Brundage et al. (2018) and the NCSC/CISA joint advisory on AI cybersecurity risks [42].

C. Ethical Considerations and Research Responsibility

The research community bears a particular duty of care in the GenAI–cybersecurity domain due to the direct dual-use risk of published findings. Several papers in our corpus reported offensive capabilities — including functional phishing email generation and malware evasion — in forms directly transferable to malicious use. Only approximately one-quarter of papers reporting offensive capabilities included explicit responsible disclosure statements or documented institutional ethics review. We advocate for community standards aligned with the Menlo Report on ICT research [43] and recommend that journal editors and conference programme committees require ethics statements as a condition of submission.

VIII. OPEN RESEARCH CHALLENGES AND FUTURE DIRECTIONS

Our analysis identifies six priority areas for future research.

First, the security of AI systems themselves — including LLM prompt injection, model inversion attacks, and training data poisoning — is underrepresented relative to its strategic significance, particularly as GenAI models become integrated into critical security infrastructure such as SIEM systems and threat intelligence platforms.

Second, the challenge of multimodal deepfake detection remains unresolved: while unimodal detectors achieve high accuracy on benchmark data, cross-modal fusion attacks that minimise audio-visual discrepancies substantially degrade detector performance [30].

Third, there is a paucity of longitudinal research on the GenAI–cybersecurity arms race. Almost all reviewed papers present static point-in-time analyses, yet the operational environment is characterised by continuous co-evolution of attack and defence capabilities.

Fourth, the application of GenAI to operational technology (OT) and industrial control system (ICS) cybersecurity is severely underrepresented, despite the potentially catastrophic consequences of successful intrusions into critical infrastructure.

Fifth, privacy-preserving GenAI for cybersecurity — including federated learning and differentially private implementations of threat intelligence sharing — requires substantially greater theoretical and empirical development.

Sixth, human factors research is needed to understand how GenAI-assisted attacks affect human vulnerability, and how GenAI-augmented defensive tools influence analyst cognitive load and decision quality.

IX. CONCLUSION

This paper has presented a systematic literature review at the intersection of Generative Artificial Intelligence and cybersecurity, synthesising 142 peer-reviewed sources published since 2018. Our review reveals a rapidly evolving, deeply dual-use technological landscape in which generative models simultaneously expand the offensive toolkit available to adversaries and offer new defensive capabilities to security professionals. Principal offensive applications include LLM-enabled industrial-scale spear-phishing, deepfake-mediated fraud and social engineering, AI-assisted malware generation and evasion, and autonomous vulnerability exploitation. Principal defensive applications include synthetic data augmentation for intrusion detection, LLM-driven alert triage and threat intelligence synthesis, and AI-enhanced penetration testing and red-teaming.

We have proposed a taxonomy of GenAI–cybersecurity interactions across five modality dimensions — language, image/video, audio, code, and network/tabular data — and three directionality dimensions — offensive, defensive, and AI-targeted. We document a pronounced research imbalance in which offensive capability demonstration substantially outpaces defensive system development. We urge the research community, policymakers, and technology developers to collaborate in addressing this asymmetry through targeted investment in defensive GenAI cybersecurity research, robust responsible disclosure practices, and adaptive governance frameworks capable of keeping pace with the rapid advancement of generative model capabilities. Securing the future digital ecosystem in the era of powerful generative AI will require sustained, multidisciplinary, and globally coordinated effort.

REFERENCES

[1] I. Goodfellow et al., “Generative adversarial nets,” Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680, 2014.

[2] T. Brown et al., “Language models are few-shot learners,” Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[3] N. Ferruz, S. Schmidt, and B. Höcker, “ProtGPT2 is a deep unsupervised language model for protein design,” Nature Communications, vol. 13, no. 1, p. 4348, 2022, doi: 10.1038/s41467-022-32007-7.

[4] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, “From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy,” IEEE Access, vol. 11, pp. 80218–80245, 2023, doi: 10.1109/ACCESS.2023.3300381.

[5] J. Hazell, “Large language models can be used to effectively scale spear phishing campaigns,” arXiv preprint arXiv:2305.06972, 2023.

[6] Y. M. P. Pa et al., “An attacker’s dream? Exploring the capabilities of ChatGPT for developing malware,” in Proc. 16th Cyber Security Experimentation and Test Workshop (CSET), New York, NY, USA, 2023, pp. 10–18, doi: 10.1145/3607505.3607513.

[7] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, “On the effectiveness of machine and deep learning for cyber security,” in Proc. Int. Conf. Cyber Conflict (CyCon), Tallinn, Estonia, 2018, pp. 371–390, doi: 10.23919/CyCon.2018.8405027.

[8] B. Dolan-Gavitt et al., “SoK: Science, security, and the elusive goal of security as a scientific pursuit,” in Proc. IEEE Symp. Security and Privacy, 2016, pp. 99–120, doi: 10.1109/SP.2016.13.

[9] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay,

"Adversarial attacks and defences: A survey," arXiv preprint arXiv:1810.00069, 2021.

[10] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "AI-driven cybersecurity: An overview, security intelligence modelling and research directions," SN Computer Science, vol. 2, no. 3, p. 173, 2021, doi: 10.1007/s42979-021-00557-0.

[11] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021, doi: 10.1145/3425780.

[12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401–4410, doi: 10.1109/CVPR.2019.00453.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," arXiv preprint arXiv:1312.6114, 2013.

[14] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017.

[15] OpenAI, "GPT-4 technical report," OpenAI, San Francisco, CA, USA, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[16] R. Anil et al., "PaLM 2 technical report," Google DeepMind, Mountain View, CA, USA, 2023. [Online]. Available: https://arxiv.org/abs/2305.10403

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[18] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695, doi: 10.1109/CVPR52688.2022.01042.

[19] J. Park et al., "Generative agents: Interactive simulacra of human behavior," in Proc. ACM Symp. User Interface Software and Technology (UIST), 2023, pp. 1–22, doi: 10.1145/3586183.3606763.

[20] J. Forge, "A note on the definition of dual use," Science and Engineering Ethics, vol. 16, no. 1, pp. 111–118, 2010.

[21] N. Bostrom and M. M. Ćirković, Global Catastrophic Risks. Oxford, U.K.: Oxford Univ. Press, 2022.

[22] M. Brundage et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," arXiv preprint arXiv:1802.07228, 2018.

[23] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021, doi: 10.1136/bmj.n71.

[24] Q. N. Hong et al., "The Mixed Methods Appraisal Tool (MMAT) version 2018," Education for Information, vol. 34, no. 4, pp. 285–291, 2018, doi: 10.3233/EFI-180221.

[25] M. Schmitt and I. Flechais, "Digital deception: Generative artificial intelligence in social engineering and phishing," arXiv preprint arXiv:2310.13715, 2023.

[26] Reuters, "Hong Kong company defrauded of HK$200 million in deepfake video call scam," Feb. 4, 2024. [Online]. Available: https://www.reuters.com

[27] I. Petrov et al., "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," arXiv preprint arXiv:2005.05535, 2020.

[28] U. Singer et al., "Make-A-Video: Text-to-video generation without text-video data," arXiv preprint arXiv:2209.14430, 2022.

[29] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, "Internet of Things security and forensics: Challenges and opportunities," Future Generation Computer Systems, vol. 78, pp. 544–546, 2018, doi: 10.1016/j.future.2017.07.060.

[30] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 7, pp. 8009–8026, 2023, doi: 10.1109/TPAMI.2022.3141002.

[31] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on GAN," in Proc. Int. Conf. Data Mining and Big Data, 2022, pp. 409–423.

[32] L. Demetrio et al., "Functionality-preserving black-box optimization of adversarial Windows malware," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 3469–3478, 2021, doi: 10.1109/TIFS.2021.3082330.

[33] H. Pearce et al., "Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions," in Proc. IEEE Symp. Security and Privacy, 2022, pp. 754–768.

[34] M. Fu et al., "VulRepair: A T5-based automated software vulnerability repair," in Proc. ACM Joint European Software Engineering Conf., 2023, pp. 935–947.
[35] Y. Luo et al., "GPT-4 can exploit real vulnerabilities," arXiv preprint arXiv:2404.08144, 2024.
[36] G. Deng et al., "PentestGPT: An LLM-empowered automatic penetration testing framework," arXiv preprint arXiv:2308.06782, 2023.
[37] A. Happe and J. Cito, "Getting pwnd by AI: Penetration testing with large language models," in Proc. ACM Joint European Software Engineering Conf., 2023, pp. 2082–2086.
[38] P. Ranade et al., "Cyberattack detection by leveraging a semantically rich cybersecurity knowledge graph," in Proc. ACM Web Science Conf., 2021, pp. 212–221.
[39] Z. Xu et al., "Automate and accelerate security knowledge extraction with BERT," in Proc. IEEE Symp. Security and Privacy, 2023, pp. 1432–1447.
[40] European Parliament, "Regulation (EU) 2024/1689: Artificial Intelligence Act," Official Journal of the European Union, 2024.
[41] White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Executive Order 14110, Oct. 30, 2023.
[42] National Cyber Security Centre (NCSC), The Near-Term Impact of AI on the Cyber Threat, Joint Advisory NCSC/CISA/FBI/ACSC, 2024.
[43] E. Kenneally and M. Bailey, Cyber-Security Research Ethics Dialogue & Strategy Workshop (CREDS) Report. Berkeley, CA, USA: USENIX Association, 2014.
[44] M. Ring et al., "Flow-based network traffic generation using generative adversarial networks," Computers & Security, vol. 82, pp. 156–172, 2019.
[45] S. Shahriar, K. Mukaiyama, and J. Schneider, "ChatGPT for cybersecurity: Practical applications, challenges, and future directions," Cluster Computing, vol. 26, no. 6, pp. 3421–3444, 2023.
[46] G. Ho et al., "Detecting and characterizing lateral phishing at scale," in Proc. USENIX Security Symposium, 2019, pp. 1273–1290.
[47] S. Gera, S. Nehra, and J. Kathuria, "The evolving cybersecurity paradigm: New threats old vulnerabilities, strategic solutions in the information age," International Journal of Research Publication and Reviews, vol. 6, Special Issue 5, pp. 430–451, May 2025.