e-ISSN 3108-026X
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 3, Issue 2, February 2026 | DOI 10.17148/IMRJR.2026.030202

Human Behaviours & Phishing Click-Through Risks Under AI-Generated Content

Sushant Patil[1], Kanchan Patil[2]

Assistant Professor, CSE, Maratha Mandal's Engineering College, Belgaum, 591156, Karnataka, India[1]

Assistant Professor, CSE, Maratha Mandal's Engineering College, Belgaum, 591156, Karnataka, India[2]

Abstract: The emergence of generative artificial intelligence (AI) and its widespread adoption across digital platforms have fundamentally altered the nature of online communication and cybersecurity threats. AI-generated content (AIGC), encompassing synthetic text, images, audio, and multimedia, has become increasingly indistinguishable from human-created material. While these advancements provide significant benefits in productivity and accessibility, they also introduce serious security challenges. One of the most critical concerns is the heightened susceptibility of individuals to phishing and social engineering attacks, as AI-generated content enables adversaries to exploit human cognitive biases with unprecedented realism and scale.

This research paper examines the intersection of human behaviour, phishing click-through risks, and the proliferation of AI-generated content. It explores how cognitive biases, emotional triggers, and information-processing limitations contribute to user vulnerability when exposed to sophisticated AI-crafted phishing messages. The paper further analyses recent advancements in AI-generated content technologies, empirical evidence of their misuse in real-world platforms, and current defensive mechanisms such as watermarking, scalable detection frameworks, and behavioral interventions. By synthesizing contemporary research, this study emphasizes the necessity of integrating human-centric approaches with technical defences to effectively mitigate AI-driven phishing threats.

Keywords: AI-generated content, Phishing attacks, Human behaviour, Cognitive bias, Social engineering, Cybersecurity.

I. INTRODUCTION

The The emergence of generative artificial intelligence (AI) and its proliferation in various domains have revolutionized the creation, dissemination, and consumption of digital content. AI-generated content (AIGC)—ranging from synthetic text and images to multimedia outputs—has become increasingly indistinguishable from human-crafted material, introducing unprecedented challenges for authenticity, attribution, and security. Among the most pressing concerns is the heightened vulnerability of individuals to phishing and other social engineering attacks, as sophisticated AI-generated content can manipulate human behaviors and increase click-through rates for malicious links or fraudulent communications.

With the rapid adoption of advanced generative models such as diffusion-based architectures, large language models (LLMs), and multimodal frameworks, the landscape of digital risks has evolved. This evolution is not only technological but deeply behavioural: attackers can now exploit cognitive biases, heuristics, and information processing limitations through hyper-realistic, contextually relevant AIGC. Meanwhile, defences—such as watermarking, proactive detection, and scalable classification frameworks—struggle to keep pace with adversarial adaptations and the shifting modalities of content generation.

This research paper examines the intersection of human behaviours, phishing click-through risks, and the proliferation of AI-generated content. It synthesizes current advancements in AIGC, analyses the behavioural vulnerabilities exploited by malicious actors, and evaluates the effectiveness of existing and emerging Defense mechanisms. Drawing exclusively on leading-edge research within the provided reference list, this paper aims to provide a comprehensive, critical perspective on the challenges and opportunities presented by AIGC in the context of phishing and broader social engineering risks.

II. THE RISE OF AI-GENERATED CONTENT: CAPABILITIES AND SOCIETAL IMPLICATIONS

A. Generative AI Evolution and Ubiquity

The advent of powerful generative AI models has catalyzed the creation of diverse, high-quality digital artifacts across modalities, including text, images, audio, and video. The diffusion model, in particular, has emerged as a versatile and

e-ISSN 3108-026X
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 3, Issue 2, February 2026 | DOI 10.17148/IMRJR.2026.030202

effective approach for generating coherent and contextually rich outputs, with applications spanning vision, audio synthesis, natural language, time-series data, and even decision-making processes [1]. The deployment of such models on edge devices and within wireless networks—facilitated by collaborative distributed frameworks—has further democratized access to content creation, while also introducing new challenges related to computational efficiency, privacy, and security [1].

AI-generated content is no longer relegated to isolated creative experiments; it now permeates mainstream platforms. For instance, Wikipedia—a longstanding bastion of human-curated knowledge—has seen a marked rise in AI-generated contributions, with recent analyses indicating that over 5% of newly created English articles contain significant AI-generated material [5]. This infiltration extends beyond encyclopaedic domains to social media, press releases, and other influential communication channels, amplifying both the reach and potential impact of AIGC.

B. Societal Risks: Disinformation, Bias, and Trust Erosion

The proliferation of AIGC brings substantial benefits—enhanced productivity, accessibility, and creative expression—but concurrently magnifies risks related to disinformation, bias amplification, and the erosion of public trust in digital content [5]. The indistinguishability of AI and human-generated outputs complicates efforts to validate authenticity, particularly in high-stakes contexts such as political communication, journalism, and online education.

The ease with which AIGC can be weaponized for malicious purposes—such as generating persuasive phishing emails, deepfake videos, or fabricated news articles—underscores the urgency of robust detection and attribution methods. Furthermore, the presence of AI-generated content in training sets for future AI models presents a recursive risk: unchecked resampling of synthetic data can degrade model performance and propagate artifacts or biases at scale [5].

III. HUMAN BEHAVIORAL VULNERABILITIES IN THE AGE OF AI-GENERATED CONTENT

A. Cognitive Biases and Social Engineering

Phishing and related social engineering attacks fundamentally exploit human cognitive processes—relying on heuristics, trust assumptions, and emotional triggers to elicit desired behaviors, such as clicking on malicious links or divulging sensitive information. AIGC significantly enhances attackers' ability to craft contextually relevant, personalized, and convincing messages, thereby increasing the likelihood of successful exploitation.

The psychological mechanisms at play include familiarity bias (the tendency to believe frequently repeated statements), authority bias (deference to perceived authoritative sources), and confirmation bias (selective acceptance of information that aligns with preexisting beliefs) [5]. AI-generated content can be tailored to exploit these biases at scale, with generative models producing phishing emails that mimic legitimate communication styles, corporate branding, or even the writing idiosyncrasies of known contacts.

B. The Click-Through Dilemma: Realism and Persuasion

Empirical studies highlight the growing difficulty individuals face in distinguishing between AI- and human-generated content, especially as generative models approach or surpass human-level fluency and coherence [5], [7]. This blurring of boundaries increases click-through rates for phishing attempts, as recipients are less likely to question the authenticity of highly polished, contextually appropriate messages.

Furthermore, the scalability and adaptability of generative AI enable attackers to continuously refine their tactics, testing message variants and optimizing for maximum behavioral impact. The feedback loop between generated content and observed human responses—facilitated by analytics, tracking, and social network signals—creates a dynamic adversarial environment in which phishing campaigns can rapidly evolve.

C. Automated Content Creation and Targeting

The collaborative distributed diffusion-based AIGC framework exemplifies the potential for decentralized, large-scale content generation, where devices within wireless networks collaborate to execute AIGC tasks efficiently [1]. While this approach optimizes resource allocation and enhances user experience, it also lowers the technical barriers for attackers, enabling the automated creation of targeted phishing content that is customized for individual recipients or demographic groups.

The convergence of AI-powered content generation with real-time data analytics, social media mining, and behavioral profiling facilitates micro-targeted social engineering attacks. Attackers can generate messages that reference recent events, personal interests, or organizational details, further increasing the plausibility and effectiveness of phishing attempts.

e-ISSN 3108-026X
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 3, Issue 2, February 2026 | DOI 10.17148/IMRJR.2026.030202

IV. DETECTION AND ATTRIBUTION: TECHNOLOGICAL COUNTERMEASURES

A. Watermarking Approaches and Their Limitations

Watermark-based detection has emerged as a key strategy for identifying AI-generated content and mitigating its misuse [2], [3], [4]. The basic principle involves embedding a unique, often imperceptible, watermark into AI-generated outputs at the time of creation. Subsequent analysis can then detect and, in some cases, attribute content to specific users or generative models.

There are two primary categories of watermarking methods: non-learning-based approaches, which rely on hand-crafted heuristics and transformations (e.g., frequency domain encoding), and learning-based approaches, which utilize neural networks for encoder-decoder training [2], [3]. Learning-based methods, particularly those employing adversarial training, have demonstrated enhanced robustness against common post-processing operations such as JPEG compression, Gaussian blur, and brightness/contrast adjustments [3], [4].

Despite these advancements, watermarking remains vulnerable to sophisticated adversarial post-processing. Attackers can leverage techniques that introduce subtle, human-imperceptible perturbations to watermarked content, effectively evading detection while preserving visual or textual quality [2]. The WEvade framework, for example, systematically generates adversarial examples that bypass watermark-based detectors in both white-box (decoder known) and black-box (API query limited) settings, underscoring the insufficiency of current watermarking standards [2].

B. User-Level Attribution and Forensic Challenges

Beyond detection, attribution—the ability to trace AI-generated content back to the originating user or service—has become increasingly important for forensic investigations, accountability, and legal compliance [3]. User-level attribution schemes assign unique watermarks to each registered user of a generative AI service, embedding this identifier into all generated content. Upon detection, the system attempts to match the extracted watermark with the user database, facilitating traceability.

However, effective attribution hinges on the selection of sufficiently dissimilar watermarks for each user to minimize false attribution and maximize detection rates. The combinatorial complexity of this problem scales with the number of users and watermark length, rendering brute-force optimization infeasible for large-scale deployments [3]. Recent research has proposed probabilistic analysis and approximate algorithms based on the farthest string problem to improve attribution performance [3]. Nevertheless, attribution inherits the (non-)robustness of the underlying watermarking method: adversarial post-processing can degrade attribution accuracy, especially in scenarios where attackers possess significant knowledge or resources [3].

C. Scalable Classification and Detection Frameworks

Given the rapid evolution and diversity of generative AI models, scalable detection and classification frameworks are essential for maintaining robust defenses against AIGC-driven attacks. The integration of perceptual hashing, similarity measurement, and pseudo-labeling enables the identification and classification of AI-generated content across modalities without necessitating retraining for every new generative model [7]. Such frameworks leverage high-dimensional feature extraction (e.g., via BART Large for text or Swin Transformer for images) and k-nearest neighbor comparisons to distinguish human and AI content and further categorize outputs by their generative source.

Incremental learning and feature augmentation—enabled by pseudo-labelling and dynamic adaptation—allow these systems to remain effective as new generative models and attack vectors emerge [7]. While competitive in benchmark evaluations, these approaches must also contend with adversarial attacks that manipulate feature distributions or exploit model blind spots, highlighting the ongoing arms race between attackers and defenders in the AIGC landscape.

V. THE PHISHING THREAT: EMPIRICAL EVIDENCE AND BEHAVIOURAL ANALYSIS

A. AI -Generated Content in the Wild: Case Studies

Recent empirical analyses of Wikipedia, Reddit, and United Nations press releases reveal a growing prevalence of AI-generated content in public information domains [5]. In English Wikipedia, for instance, over 5% of new articles created in August 2024 were flagged as AI-generated by state-of-the-art detectors, with many exhibiting lower quality, reduced citation density, and increased self-promotion or polarization compared to human-authored counterparts [5]. Manual inspection further identified cases where users engaged in coordinated campaigns to promote specific viewpoints or entities, often leveraging AI to generate or translate large volumes of content.

These trends are mirrored in other domains: Reddit comment sections and press releases from international organizations have also seen spikes in AI-generated contributions, some of which contain persuasive calls to action or political

e-ISSN 3108-026X
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 3, Issue 2, February 2026 | DOI 10.17148/IMRJR.2026.030202

messaging [5]. The implications for phishing and social engineering are profound—attackers can now seamlessly inject credible, high-volume messages into trusted channels, increasing the likelihood that targets will encounter, trust, and act upon malicious prompts.

B. Behavioral Triggers and Click-Through Susceptibility

The success of phishing campaigns depends on more than just the quality of content; it hinges on exploiting specific behavioral triggers that prompt recipients to engage with malicious links or requests. AI-generated phishing emails, for example, can be tailored to evoke urgency ("Your account will be locked in 24 hours"), authority ("This is the IT department"), or personal relevance ("Your recent purchase requires verification"), all of which are well-documented psychological levers [5].

The realism and adaptivity of AIGC further increase click-through susceptibility. As generative models learn to mimic not only linguistic patterns but also emotional tone, formatting, and context, recipients become less able to rely on traditional red flags (e.g., poor grammar, generic greetings) to filter malicious messages. Combined with attackers' ability to rapidly A/B test and iterate on message variants, this leads to a continuous optimization of phishing tactics, maximizing behavioral impact and minimizing detection [5], [7].

C. Platform-Specific Risks and Amplification

Certain platforms are particularly susceptible to the amplification of AI-generated phishing risks. Publicly curated spaces like Wikipedia provide a veneer of trust and authority, making it more likely that unsuspecting users will click on embedded links or follow calls to action within articles [5]. Social media platforms, with their viral sharing dynamics and personalized recommendation algorithms, can further amplify the reach and impact of AIGC-based phishing campaigns. The translation and propagation of AI-generated content across languages and platforms also complicate detection and response. Machine-translated articles or messages may evade language-specific filters, while the rapid replication of content in different domains increases the surface area for potential exploitation [5]. Attackers may even leverage automated tools to generate and disseminate content across multiple platforms simultaneously, overwhelming manual moderation and traditional security controls.

VI. DEFENSIVE STRATEGIES: OPPORTUNITIES AND LIMITATIONS

A. Enhancing Watermark Robustness and Attribution

Strengthening watermarking techniques remains a priority for mitigating AIGC-driven phishing risks. Recent work advocates for the adoption of learning-based watermarking methods with adversarial training to improve robustness against common post-processing and black-box adversarial attacks [3], [4]. Additionally, restricting access to watermark decoders and detection APIs—limiting them to trusted customers or internal forensic teams—can reduce the risk of decoder compromise and large-scale evasion [3].

Nevertheless, current watermarking approaches remain imperfect. As demonstrated by the WEvade attack, even adversarially trained learning-based watermarks can be circumvented with relatively small, human-imperceptible perturbations, especially in white-box settings where attackers have decoder access [2]. The arms race between detection and evasion continues, necessitating ongoing research into more resilient watermarking schemes, hybrid passive-active detection approaches, and the integration of behavioral analytics into attribution workflows.

B. Scalable and Adaptive Detection Frameworks

The development of scalable, adaptable detection frameworks—capable of integrating new generative models and evolving attack vectors without extensive retraining—is crucial for sustaining an effective defense posture [7]. The use of perceptual hashing and k-nearest neighbor similarity comparisons enables rapid adaptation to novel content types, while pseudo-labeling and incremental learning allow systems to generalize from limited labeled data.

However, these frameworks must be constantly updated to incorporate new features, attack signatures, and adversarial techniques. Attackers may seek to manipulate feature representations, exploit weaknesses in pseudo-labeling algorithms, or craft content that mimics human feature distributions, requiring defenders to maintain a proactive, iterative approach to model refinement and evaluation [7].

C. Human- -Centric Interventions: Education and Behavioral Nudges

While technological defenses are indispensable, addressing the behavioral dimension of phishing risks demands complementary human-centric interventions. User education, awareness campaigns, and the promotion of digital literacy can help individuals recognize and resist social engineering tactics, even as AIGC blurs the lines between authentic and synthetic content [5].

e-ISSN 3108-026X
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 3, Issue 2, February 2026 | DOI 10.17148/IMRJR.2026.030202

Behavioral nudges—such as warning banners, contextual risk indicators, and friction in high-risk interactions—can reduce impulsive click-throughs and encourage more deliberate decision-making. Embedding such interventions within platforms and communication channels adds an additional layer of defense, leveraging cognitive science insights to counteract the psychological levers exploited by attackers.

VII. FUTURE DIRECTIONS AND OPEN CHALLENGES

A. Towards Resilient, Multi-Modal Defense Ecosystems

The future of AIGC defense lies in the integration of multiple, complementary approaches—combining robust watermarking, scalable detection frameworks, user-level attribution, and behavioral analytics into a cohesive ecosystem. Cross-modal detection capabilities (spanning text, images, audio, and video) will be essential as attackers diversify their tactics and exploit the full spectrum of generative AI capabilities [1], [7].

Collaborative distributed frameworks for AIGC—such as the one proposed for wireless networks—offer opportunities for decentralized, privacy-preserving content generation and detection, but also introduce new attack surfaces and coordination challenges [1]. Ensuring the security, accountability, and resilience of such systems will require advances in secure multi-party computation, federated learning, and trust management.

B. Addressing Adversarial Adaptation and Model Degradation

The rapid pace of adversarial adaptation in the AIGC landscape necessitates continuous research into detection and attribution methods that are robust to evolving evasion tactics. This includes the development of certifiably robust watermarking schemes, the use of ensemble or hybrid detection models, and the exploration of new metrics for content authenticity and provenance [2], [3], [7].

Additionally, the recursive risk of model degradation—where AI models trained on increasingly synthetic datasets lose their ability to distinguish real from fake—poses a long-term challenge for the reliability and trustworthiness of generative AI. Strategies for data provenance tracking, dataset curation, and the exclusion of synthetic content from training pipelines must be prioritized to preserve the integrity of future AI systems [5].

C. Ethical, Legal, and Governance Considerations

The deployment of AIGC detection and attribution technologies raises complex ethical, legal, and governance questions. Balancing the imperative for accountability and security with concerns about privacy, free expression, and potential misuse of detection tools requires careful policy development and stakeholder engagement [3], [5]. The establishment of industry standards, regulatory frameworks, and transparent auditing processes will be critical for fostering trust and accountability in the era of ubiquitous AIGC.

VIII. CONCLUSION

The intersection of human behaviors, phishing click-through risks, and the proliferation of AI-generated content represents a critical frontier in digital security and trust. As generative AI models become more sophisticated, accessible, and pervasive, attackers gain unprecedented capabilities to craft persuasive, contextually relevant phishing messages that exploit cognitive biases and behavioral vulnerabilities. The empirical evidence underscores the growing prevalence of AIGC in influential information domains, elevating the urgency of robust, scalable, and adaptive defense mechanisms.

Watermark-based detection and attribution offer promising avenues for identifying and tracing AI-generated content, but remain vulnerable to advanced adversarial attacks and scalability challenges. Scalable classification frameworks and human-centric interventions provide additional layers of defense, yet must be continually refined to keep pace with the evolving threat landscape. Ultimately, addressing the risks posed by AIGC-driven phishing and social engineering requires a holistic, multi-disciplinary approach—integrating technological innovation, behavioral science, and ethical governance to safeguard the integrity of digital communication in the age of AI.

ACKNOWLEDGMENT

The heading of the Acknowledgment section must not be numbered. It is important to recognize the contributions of those who have supported the work presented in this paper. If applicable, the authors should express gratitude to funding agencies, colleagues, or other parties who have contributed to the research but are not listed as authors.

e-ISSN 3108-026X
International Multidisciplinary Research Journal Reviews (IMRJR)
A Peer-reviewed journal
Volume 3, Issue 2, February 2026 | DOI 10.17148/IMRJR.2026.030202

REFERENCES

[1]. H. Du, R. Zhang, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, X. Shen, and H. V. Poor, "Exploring Collaborative Distributed Diffusion-Based AI-Generated Content (AIGC) in Wireless Networks," *arXiv preprint arXiv*:2304.03446v2, Dec. 2023. [Online]. Available: https://arxiv.org/pdf/2304.03446v2

[2]. Z. Jiang, J. Zhang, and N. Z. Gong, "Evading Watermark based Detection of AI-Generated Content," in Proceedings of the 2023 *ACM SIGSAC* Conference on Computer and Communications Security (CCS '23), Nov. 2023, pp. 1–20. [Online]. Available: https://arxiv.org/pdf/2305.03807v5

[3]. Z. Jiang, M. Guo, Y. Hu, and N. Z. Gong, "Watermark-based Attribution of AI-Generated Content," *arXiv preprint arXiv*:2404.04254v3, Nov. 2024. [Online]. Available: https://arxiv.org/pdf/2404.04254v3

[4]. C. Brooks, S. Eggert, and D. Peskoff, "The Rise of AI-Generated Content in Wikipedia," *arXiv preprint arXiv*:2410.08044v1, Oct. 2024. [Online]. Available: https://arxiv.org/pdf/2410.08044v1

[5]. C. Brooks, S. Eggert, and D. Peskoff, "The Rise of AI-Generated Content in Wikipedia*," arXiv preprint arXiv*:2410.08044v1, Oct. 2024. [Online]. Available: https://arxiv.org/pdf/2410.08044v1

[6]. A.-K. Duong and P. Gomez-Krämer, "Scalable Framework for Classifying AI-Generated Content Across Modalities," *arXiv preprint arXiv*:2502.00375v2, Feb. 2025. [Online]. Available: https://arxiv.org/pdf/2502.00375v2

[7]. A.-K. Duong and P. Gomez-Krämer, "Scalable Framework for Classifying AI-Generated Content Across Modalities," *arXiv preprint arXiv*:2502.00375v2, Feb. 2025. [Online]. Available: https://arxiv.org/pdf/2502.00375v2