International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V74I1P102

Volume 74 Issue 1, 7-11, January 2026

Review Article

Cybersecurity Risks of Social Media Platforms in the Age of Artificial Intelligence

Yash Patel

Capella University, Minnesota, USA.

Corresponding Author : ypatel1@capellauniversity.edu

Received: 21 November 2025 Revised: 29 December 2025 Accepted: 13 January 2026 Published: 29 January 2026

Abstract *- Social media platforms now sit at the center of everyday communication for individuals and organizations. That visibility and reach also make them attractive to attackers. As Artificial Intelligence (AI) becomes embedded in social media ecosystems through content creation tools, recommendation systems, automated interaction, and moderation, well-known threats such as phishing and impersonation are being reshaped rather than replaced. This article reviews double-masked peer-reviewed research on cybersecurity risks tied to social media use, with emphasis on how AI-enabled techniques amplify deception, accelerate malicious content spread, and complicate detection. A thematic analysis synthesizes empirical findings across social engineering, account compromise, malicious link distribution, data leakage, and human factors. The literature indicates that social media security problems are socio-technical: attacks succeed through a mix of platform features, organizational workflows, and user judgment under pressure. The article closes with stakeholder-specific implications for platform developers, organizational security teams, and policymakers, and identifies research gaps that remain unresolved.*

Keywords *- Social Engineering, Phishing, AI-Enabled Threats, Data Leakage, Artificial Intelligence.*

1. Introduction

Social media has moved well beyond casual networking. Organizations use these platforms to communicate with customers, recruit employees, respond during incidents, and maintain public visibility. That reliance creates a predictable outcome: the same channels used for outreach become channels for exploitation. Social platforms differ from internal enterprise systems in ways that matter for security. They are public-facing, fast-moving, and built around social trust. Professional and personal identities often mix, and content can spread rapidly with minimal friction. These conditions are convenient for legitimate engagement, but they also help adversaries. Prior research shows that social media can support credential theft, phishing, malware propagation, and targeted impersonation that carries real organizational costs (Egele et al., 2013; Stringhini et al., 2010).

AI has added another layer to this environment. Text, images, and other media can now be produced at scale with high surface credibility. In practice, this means attackers can craft more convincing pretexts, iterate faster, and tailor messages to a target's context using public profile information. Empirical work indicates that synthetic profiles and AI-assisted artifacts can shift user trust and reduce the effectiveness of certain detection approaches (Feng et al., 2022; Qingying et al., 2024).

Although there is a large body of research on phishing, social engineering, and AI security, the evidence is scattered across subfields. Many studies examine single threats in isolation (e.g., spam campaigns or compromised accounts) or focus on AI at a conceptual level without connecting it to the practical mechanics of social-media-enabled attacks. What is missing is an integrated technical view that links (a) social platform affordances, (b) AI-amplified attack methods, and (c) organizational risk management. This review contributes in three ways. First, it brings together double-masked peer-reviewed findings across social media security, AI-enabled deception, and human factors, rather than treating them as separate topics. Second, it uses thematic analysis to show how threats converge in practice (for example, account compromise supporting malicious link distribution). Third, it separates implications by stakeholder group: platform developers, organizational security teams, and policy makers so that recommendations match responsibility and authority.

2. Literature Review

Social engineering remains one of the most persistent drivers of incidents in social media contexts. A key reason is that social media interactions often arrive already "pre-trusted": the sender appears familiar, the message appears socially situated, and context cues are easy to counterfeit. Foundational work on compromised social accounts

demonstrated that takeovers are often detectable because attackers introduce measurable anomalies, timing shifts, content changes, and altered interaction patterns (Egele et al., 2013). This result matters because it challenges the assumption that a takeover is indistinguishable from normal use; in many cases, it leaves signals. A second thread in the literature connects social media compromise to broader organizational risk.

Credential compromise research indicates that stolen credentials are commonly reused across services and environments (Ho et al., 2017). In other words, a social media compromise can serve as an entry point, not an endpoint. That linkage is operationally important for organizations that treat social accounts as "marketing assets" rather than security assets. Impersonation and authority abuse represent a further escalation. Attackers do not always need full account takeover to cause harm; convincing imitation can be enough to trigger payment requests, urgent data sharing, or workflow exceptions. Social platforms make these attacks easier by allowing adversaries to mimic branding, reuse profile images, or adopt a credible tone and posting patterns. These attacks often succeed through process weaknesses (e.g., inadequate verification steps) as much as through technical gaps.

The literature repeatedly shows that social media can function as a high-throughput delivery mechanism for malicious content. Early measurement studies documented how coordinated campaigns exploit trending topics, link shorteners, and automated accounts to distribute harmful links and scams (Gao et al., 2010). The core insight is simple: velocity reduces the value of traditional defences. When malicious URLs appear and spread quickly, blocklists and delayed takedowns often lag in exposure. Related work explored how suspicious URLs propagate across social streams and how attackers engineer distribution to maximize reach before detection (Gao et al., 2012). Later studies reinforced that moderation and detection remain imperfect in practice; malicious links can persist long enough to reach many users, especially when campaigns adapt their infrastructure or rotate domains (Chaudhary et al., 2021). A recurring pattern is the use of compromised legitimate accounts for delivery. Links posted by trusted accounts are more likely to be clicked, shared, or treated as legitimate. As a result, account compromise and malicious content distribution reinforce each other: takeover provides credibility, and credible distribution increases the payoff from takeover.

AI changes the economics of deception. It lowers the cost of producing plausible content and raises the ceiling on personalization. Empirical work suggests that profiles generated using modern techniques can receive higher trust ratings than older-style fake accounts, especially when they maintain consistent interaction history and realistic media (Feng et al., 2022). That result is not merely about aesthetics; it affects how quickly an attacker can build rapport and how easily a target dismisses warning signs. AI also intersects with detection. Work on evasion shows that generative models can undermine certain protective systems, including image-based defences, by producing artifacts that defeat detectors while remaining convincing to human viewers (Qingying et al., 2024). This is important because it shifts defender burden: the challenge is not only stopping "known bad" content, but handling content that is designed to look normal. The overall pattern is amplification. AI does not invent social engineering; it accelerates it, personalizes it, and makes it harder to filter using static rules.

Social platforms produce a steady stream of public signals, posts, likes, affiliations, connections, photos, and metadata. Earlier research on disclosure and privacy showed that users routinely reveal sensitive information without intending to do so, often because the platform incentives Favor sharing and social visibility (Gross & Acquisti, 2005). Over time, these small disclosures can become meaningful. For adversaries, they provide context for targeted messaging, timing, and credibility. Even when direct disclosure is limited, inference remains possible. Public activity patterns, location hints, organizational relationships, and role details can support detailed targeting. In practical terms, this turns "privacy exposure" into "security exposure" because attackers can craft messages that match the target's environment and expectations.

Many social media attacks succeed because they exploit how decisions are made in context. Behavioural research on phishing victimization shows that users often rely on fast heuristics, especially under time pressure or when the message appears socially legitimate (Luo et al., 2013). This does not mean users are careless; it reflects normal cognition in a high-volume information environment. Training and awareness can help, but the literature suggests limits when training is treated as the primary control. If users are asked to compensate for weak workflows, weak verification procedures, or weak authentication, training becomes a partial patch rather than a durable defence. Stronger outcomes appear more plausible when organizations pair education with practical friction-reducing controls (e.g., reporting mechanisms, verification steps for sensitive requests, and protective authentication).

3. Methodology

A structured narrative review method was used to synthesize double-masked peer-reviewed research in cybersecurity and privacy. Studies were selected from established venues that use anonymized review processes, alongside relevant peer-reviewed journals. The selection focused on literature that examined social media threats, AI-enabled deception, malicious content distribution, and

privacy-to-security escalation in social platforms. A qualitative thematic analysis was applied. Each article was coded for threat type, attack pathway, enabling platform feature, human factors component, and proposed mitigations.

Themes were iteratively refined by comparing codes across studies and consolidating recurring patterns into broader categories. This methodology was selected because the evidence base is distributed across multiple sub-areas. A single empirical dataset would not capture the breadth of the threat surface addressed in this review.

4. Research Findings

Social media platforms amplify trust-based attacks by embedding malicious activity within familiar social contexts. Compromised and impersonated accounts exploit established relationships, significantly increasing attack success rates (Egele et al., 2013; Ho et al., 2017). Threat vectors frequently converge on social media platforms. Account compromise, malicious content dissemination, and impersonation attacks often occur in combination, creating cascading risk scenarios (Gao et al., 2010; Chaudhary et al., 2021).

Across the literature, the same mechanism appears repeatedly: social platforms externalize trust into visible cues such as names, profile images, mutual connections, and prior activity. Attackers exploit those cues. Compromised accounts are especially valuable because they inherit legitimacy and history, often long enough to produce harm before detection (Egele et al., 2013). Credential compromise research reinforces that the consequences can extend beyond the platform when stolen access is reused elsewhere (Ho et al., 2017).

The studies do not describe neat categories in practice; instead, they describe combinations. Spam campaigns evolve into phishing, phishing leads to credential theft, credential theft supports account takeover, and compromised accounts then distribute malicious content. Measurement work on social spam and suspicious links illustrates how coordinated campaigns exploit platform dynamics to sustain reach (Gao et al., 2010; Gao et al., 2012). Evidence of detection gaps further indicates that some malicious links persist long enough to support multi-stage attacks (Chaudhary et al., 2021).

Evidence suggests that modern synthetic profiles can shift user trust and engagement, which changes how quickly attackers can build credibility (Feng et al., 2022). In parallel, evasion research indicates that generative techniques can be adapted to defeat certain defensive tools, including visual detection approaches (Qingying et al., 2024). The practical consequence is not simply "more attacks," but attacks that are harder to triage because messages and media are less templated and less repetitive.

Public information on social platforms contributes directly to targeted attack planning. Disclosures that look harmless in isolation can still support detailed targeting and persuasive pretexting when aggregated (Gross & Acquisti, 2005). This helps explain why organizations sometimes face highly tailored impersonation and phishing attempts that reference real colleagues, real projects, or real events.

User-focused interventions matter, but the literature suggests limits when awareness is used as a substitute for stronger controls. When decisions are made quickly and socially, heuristic processing increases risk (Luo et al., 2013). The more realistic and personalized the content becomes, the less feasible it is to expect users to detect deception reliably without strong organizational safeguards.

5. Discussion

The reviewed evidence points to a consistent conclusion: social media security problems are socio-technical. Platform affordances create opportunities, organizational processes create pathways, and user judgment becomes the final gate. That combination explains why technical solutions alone are often insufficient and why awareness alone is fragile.

The literature supports prioritizing controls that reduce impersonation and slow malicious link propagation. Improved detection of account compromise signals (Egele et al., 2013), faster response to suspicious URL patterns (Gao et al., 2012), and defenses designed with adversarial adaptation in mind become increasingly relevant as generative techniques are used to evade detectors (Qingying et al., 2024). Transparency in enforcement and clearer signals for authenticity can also reduce reliance on user guesswork. Organizations benefit from treating social media accounts as part of the security perimeter. Credential reuse risk and lateral movement pathways suggest that social accounts should receive the same authentication discipline applied to enterprise systems (Ho et al., 2017). Practical measures include phishing-resistant authentication where feasible, well-defined verification steps for sensitive requests, monitoring for impersonation, and simple reporting channels so suspicious interactions can be escalated quickly.

Policy levers may be relevant where platform incentives do not align with security outcomes. The persistence of malicious link activity and the scale of coordinated campaigns suggest a need for accountability and transparency around moderation performance, incident response timelines, and user protection mechanisms (Chaudhary et al., 2021). Policymakers may also consider frameworks that encourage robust reporting, standardized disclosure practices for major platform incidents, and support for protective defaults in privacy settings.

A key cause is overstatement. Not every AI development guarantees a corresponding leap in attacker success, and defenses also evolve. However, available evidence supports the claim that AI increases variation and plausibility, which complicates both user judgment and automated filtering (Feng et al., 2022; Qingying et al., 2024). That is an evidence-based inference, not a prediction detached from observed behavior.

6. Conclusion

Social media platforms are now a routine part of organizational operations, and that routine use creates predictable security exposure. The reviewed literature shows how social engineering, compromised accounts, malicious link campaigns, and privacy-derived targeting interact in ways that amplify risk. AI intensifies these dynamics by enabling faster content production, richer personalization, and, in some cases, evasion of protective systems. Effective mitigation depends on layered defences. Platform-side protections can reduce exposure, but organizations also need governance, authentication discipline, verification steps for sensitive requests, and training that supports realistic decision-making rather than blame. Continued research is needed to measure how AI-enabled deception performs in real-world settings over time, how platform design choices influence attack success, and which mitigation combinations produce durable results.

Conflicts of Interest

The author(s) declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This research was conducted as an independent research initiative to study security threats with the increasing advancement of AI implementation, and this research was self-funded.

References

[1] Zainab Alkhalil et al., "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Frontiers of Computer Science*, vol. 3, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Yazan Boshmaf et al., "The Socialbot Network: When Bots Socialize for Fame and Money," *Proceedings of the 27th Annual Computer Security Applications Conference,* pp. 93-102, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[3] Yazan Boshmaf et al., "Design and Analysis of a Social Botnet," *Computer Networks,* vol. 57, no. 2, pp. 556-578, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[4] Qiang Cao et al., "Aiding the Detection of Fake Accounts in Large Scale Social Online Services," *9th USENIX Symposium on Networked Systems Design and Implementation*, 2012. [Google Scholar] [Publisher Link]

[5] Sayak Saha Roy, Unique Karanjit, and Shirin Nilizadeh, "What Remains Uncaught? Characterizing Sparsely Detected Malicious URLs on Twitter," *Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb)*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Manuel Egele et al., "COMPA: Detecting Compromised Accounts on Social Networks," *Network and Distributed System Security Symposium,* vol. 13, pp. 83-91, 2013. [Google Scholar] [Publisher Link]

[7] Jaron Mink et al., "DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks," *Proceedings of the 31st USENIX Security Symposium,* 2022. [Google Scholar] [Publisher Link]

[8] Hongyu Gao et al., "Detecting and Characterizing Social Spam Campaigns," *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pp. 35–47, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[9] Hongyu Gao et al., "Towards Online Spam Filtering in Social Networks.," *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, pp. 1-16, 2012. [Google Scholar]

[10] Oana Goga et al., "Exploiting Innocuous Activity for Correlating Users Across Sites," *Proceedings of the 22nd International Conference on World Wide Web*, pp. 447-458, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ralph Gross, and Alessandro Acquisti, "Information Revelation and Privacy in Online Social Networks," *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, pp. 71-80, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[12] Grant Ho et al., "Detecting Credential Compromise in Enterprise Environments," *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security*, 2017. [Google Scholar] [Publisher Link]

[13] Tom N. Jagatic et al., "Social Phishing," *Communications of the ACM*, vol. 50, no. 10, pp. 94-100, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[14] Katharina Krombholz et al., "Advanced Social Engineering Attacks," *Journal of Information Security and Applications*, vol. 22, pp. 113-122, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[15] Xin Luo et al., "Investigating Phishing Victimization with The Heuristic–Systematic Model: A Theoretical Framework and An Exploration," *Computers and Security*, 38, pp. 28-38, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[16] Lukasz Olejnik et al., "The Leaking Battery - A Privacy Analysis of the HTML5 Battery Status API," *Data Privacy Management, and Security Assurance,* pp. 254-263, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[17] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna, "Detecting Spammers on Social Networks," *Proceedings of the 26th Annual Computer Security Applications Conference*, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[18] Gianluca Stringhini et al., "Follow the Green: Growth and Dynamics in Twitter Follower Markets," *Proceedings of the 2013 Internet Measurement Conference*, pp. 163-176, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[19] Gang Wang et al., "Social Turing Tests: Crowdsourcing Sybil Detection," Social and Information Networks, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[20] Biwei Yan et al., "On Protecting the Data Privacy of Large Language Models (LLMs) and LLM Agents: A Literature Review," *High-Confidence Computing*, vol. 5, no. 2, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Qingying Hao et al., "It Doesn't Look Like Anything to Me: Using Diffusion Model to Subvert Visual Phishing Detectors," *Proceedings of the 33rd USENIX Security Symposium*, Philadelphia, PA, USA, 3027-3044, 2024. [Google Scholar] [Publisher Link]