International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Trust Is Not a Default Control: AI-Powered Social Engineering and the Need to Have New Governance.

Alex Mathew
Bethany College, USA

***Abstract*— The blistering development of artificial intelligence (AI) and machine learning (ML) has fundamentally undermined the old paradigm of information security, which relies on judgment and implicit trust. Today's attackers also use generative AI to automate and drive hyper-personalized social engineering attacks, including voice cloning and video deepfakes, bypassing email filters and awareness training. In this paper, it is argued that trust is no longer a default control. Instead, the new governance orientation organisations need to follow is based on the principles of Zero Trust Architecture (ZTA), AI-based defensive surveillance, and specialised incident response measures to establish real-world digital resilience during a time of mechanical persuasion.**

***Keywords*— Artificial Intelligence (AI), Social Engineering, Zero Trust Architecture (ZTA), Deepfakes, Cybersecurity Governance, Digital Resilience, Synthetic Media.**

I. INTRODUCTION

Keywords: Information security frameworks have existed for decades, with an underlying implicit rule: you can be in control, or you can trust once, and this is established through authentication and authorization. Employees are relied on to detect phishing. Trust is placed in systems after preliminary validation. This premise, however, is being systematically undermined by the rapid development of artificial intelligence (AI) and machine learning (ML). Trust as a default control will no longer be possible in the age of AI-based social engineering. Human psychology is being manipulated to an extent never before seen, with hyper-personalized, automated, and scalable threats that require a radical shift in governance, risk, and compliance strategies.

II. THE NEW THREAT LANDSCAPE: BEYOND GENERIC PHISHING

Conventional social engineering is based on generalized lures. This is turned into a strike in surgery by AI. Twenty thousand social media posts, hundreds of objects in a professional network, data breaches, and other corporate messages can be analyzed by generative AI models to produce highly individualized content.

Such evolution shifts the disinformation that previously was viewed as a socially marginal factor into a central axis of cybersecurity threats, serving as a direct technical attack vector (Caramancion et al., 2022). It is not just an email that is spoofed by a "CEO". It is a voice clone of an authoritative voice, a video deepfake of a friend, or an email subtly mentioning an internal project, with the correct jargon and at the right time, based on the behavioral results of the target being communicated with (ISACA, 2025). This is a systematic attack process that, as illustrated in Figure 1, attacks in an automated, cyclical manner.

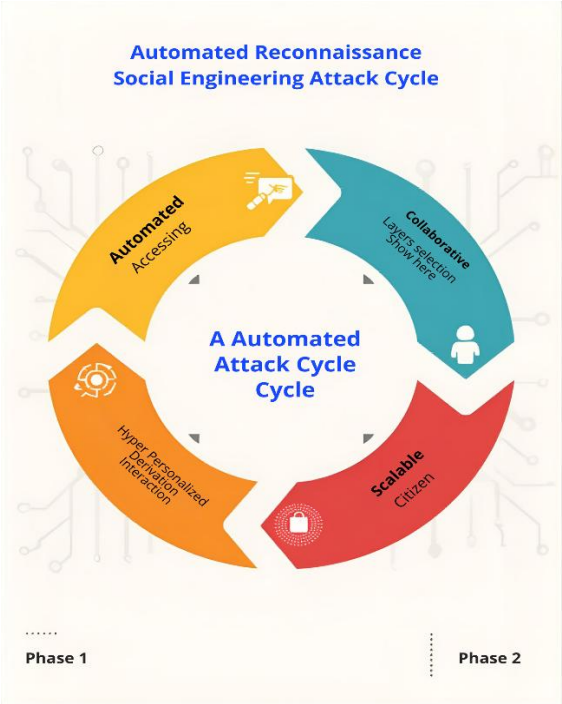

Figure 1: AI-Controlled Social Engineering attack cycle (Original art designed on the occasion of this article)

Source: *Created in Canva*

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

The scale is also a cause of concern. The reconnaissance, crafting, and deployment steps in attacks are automated using AI, allowing threat actors to conduct targeted campaigns against thousands of people at once and provide them with a unique, believable story (Achuthan et al., 2024; Mohamed, 2023). This undermines the effectiveness of awareness training aimed at recognizing generic red flags, since the attacks no longer have any red flags.

III. Why Existing Controls Are Insufficient

Even existing security measures are, in many cases, incompatible with this threat due to audits of already developed facilities, such as higher learning institutions. In spheres such as identity controls and awareness education, the sophisticated psychological nature and technical expertise of AI-enabled attacks are often not considered, creating an essential governance vacuum (Sabillon et al., 2024).

- *Awareness Training:* Difficult to maintain in AI-mediated, synthesized media sophistication.
- *Email Filters:* Could not identify new and personalized text, which contains no suspicious links or attachments.
- *Multi-Factor Authentication (MFA):* MFA is almost essential, but can be compromised by having a real-time interaction between AI and the target (e.g., a deepfaked voice call to coerce a target into accepting an MFA push notification)**Error! Bookmark not defined.**.

The fundamental weakness lies in retaining the idea of human judgment as the last resort for attacks designed to exploit it.

IV. A Governance Framework For The AI-Driven Social Engineering Era

The new governance orientation of ensuring that companies reconfigure a new trust must be carried out, but verify, never trust without explicitly stating this. It requires decoupling technical controls from heavyweight procedural and human controls, and developing the concept of digital resilience (ISACA, 2024).

1. Principle of Least Privilege (PoLP) & Zero Trust Architecture (ZTA): They must proceed with network-based considerations to finalise identity and transaction verification. A separate channel of trust not associated with the request vector should be used to authenticate any sensitive access request. This is the opposite extreme of default trust control, as defined in ZTA: never trust, always verify (Sabillon et al., 2024).
2. Artificial Intelligences With Plus Defense and Continuing Check: AI vs. AI. Adopt AI-monitored security systems that analyse communication patterns, detect deviations in voice or writing style, and assess the circumstances of the request in real time, which is one of the key trends in modern cyber defence (Alanezi & AL-Azzawi, 2024; Mohamed, 2023). Use continuous authentication systems that assess users' actions following the process.
3. Use of AI and Data Footprint Governance: Data governance is becoming a critical security issue. Companies need to scan and downplay the digital presence of significant personalities. The use of generative AI tools by employees should be regulated by policies aligned with global efforts toward responsible AI, as sensitive data can fuel future attacks (Mohamed, 2023).
4. Advanced, Simulated training: Shift theoretical training to AI-driven simulation, expose employees to hyper-realistic, personal phishing and vishing cases, and develop practical resilience.
5. Incident Response with Synthetic Media: Update incident response plans with procedures for confirming communications using established pre-planned out-of-band code words or channels in case of a suspected synthetic media attack (ISACA, 2025).

V. Conclusion

AI-based social engineering is a paradigm shift, and not the technical escalation of the matter. It focuses on the human layer, a strength powerful enough to make the conventional trust models a forgotten thing. For audit, risk, and security professionals, it is pretty straightforward: the governance models should be re-modeled to remove the concept of trust as a control in its own right (Schreiber & Schreiber, 2025). By adopting Zero Trust, defensive AI, data exhaust management, and human-layer hardening through sophisticated simulation, companies can establish a defensible posture and foster real-world digital trust (ISACA, 2024). In the era of mechanical persuasion, the best way to stay resilient is no longer to assume unthinkingly, but to evaluate in an innovative, ongoing, and contextual way.

REFERENCES

[1] Achuthan, K., Ramanathan, S., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions. *Frontiers in big data*, *7*, 1497535. https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1497535/full?utm_source=perplexity

[2] Alanezi, M., & AL-Azzawi, R. M. A. (2024). AI-powered cyber threats: A systematic review. *Mesopotamian Journal of CyberSecurity*, *4*(3), 166-188. https://mesopotamian.press/journals/index.php/CyberSecurity/article/view/648

[3] Caramancion, K. M., Li, Y., Dubois, E., & Jung, E. S. (2022). The missing case of disinformation from the cybersecurity risk continuum: A comparative assessment of disinformation with other cyber threats. Data, 7(4), 49. https://doi.org/10.3390/data7040049

[4] ISACA. (2024). What is resilience, and how does it promote digital trust? ISACA Journal, 4. https://www.isaca.org/resources/isaca-journal/issues/2024/volume-4/what-is-resilience-and-how-does-it-promote-digital-trust

[5] ISACA. (2025). The rise of deepfakes: A deep dive into synthetic media and its implications. ISACA Journal, 1. https://www.isaca.org/resources/isaca-journal/issues/2025/volume-1/the-rise-of-deepfakes-a-deep-dive-into-synthetic-media-and-its-implications

[6] Mohamed, N. (2023). Current trends in AI and ML for cybersecurity: A state-of-the-art survey. Cogent Engineering, 10(2), 2272358. https://doi.org/10.1080/23311916.2023.2272358

[7] Sabillon, R., Higuera, J. R. B., Cano, J., Higuera, J. B., & Montalvo, J. A. S. (2024). Assessing the Effectiveness of Cyber Domain Controls When Conducting Cybersecurity Audits: Insights from Higher Education Institutions in Canada. Electronics, 13(16), 3257. https://doi.org/10.3390/electronics13163257

[8] Schreiber, A., & Schreiber, I. (2025). AI for cyber-security risk: harnessing AI for automatic generation of company-specific cybersecurity risk profiles. *Information and Computer Security*, *33*(4), 520-546. https://www.sciencedirect.com/org/science/article/abs/pii/S2056496125000066

[9] Sharma, R. S., Loucif, S., Kshetri, N., & Voas, J. (2024). Global initiatives on "safer" and more "responsible" artificial intelligence. Computer, 57(11), 131-137. https://ieeexplore.ieee.org/abstract/document/10718669/