Journal of Data & Digital Innovation

Pages: 109-117
Volume-II, Issue-II (2025)

Agentic AI as a Proactive Cybercrime Sentinel: Detecting and Deterring Social Engineering Attacks

Anwar Mohammed

SINGHANIA UNIVERSITY RAJASTHAN, INDIA, anwar.emails@gmail.com

Abstract:

Social engineering attacks still remain the most persistent and adaptive forms of cyber threats in the modern digital world. Social engineering attacks are not the same as traditional cyberattacks, where it manipulates human behavior to gain access without permission, thus, security mechanisms that rely on technology become less effective. A new research in this paper proposes the use of "Agentic Artificial Intelligence (Agentic AI)" as a novel and proactive sentinel against social engineering attacks. The researchers believe that by utilizing the autonomy, goal-orientation, real-time learning, and situation-adaptability, the Agentic AI can act as a dynamic cybersecurity agent who is capable of detecting, analyzing, and preventing such threats. The study brings together natural language processing, behavioral analytics, and reinforcement learning as one agentic model to emphasize subtle linguistic and psychological characteristics that are similar to those of the implementers of social engineering. A simulation of a phishing and vishing scenario was used as a test bed for the evaluation of the performance of AI. The results show that Agentic AI goes beyond rule-based systems and traditional machine learning classifiers in the detection of social engineering attempts with a far greater accuracy and lower false alarm rate. This paper not only illustrates the methodology of incorporating agency into AI-powered cyber defense systems, but also declares that Agentic AI is not only capable of reacting but also that it can strategize and foresee thus, being capable of playing the role of a watchdog in the ongoing cyber war against human-centric attacks.

Keywords: Agentic AI, Social Engineering, Cybersecurity, Proactive Defense, Reinforcement Learning, Behavioral Analytics, Phishing Detection, Vishing, AI Sentinels, Autonomous Security Systems

I. Introduction

Social engineering has emerged as one of the most dangerous cyber threats because it exploits the most vulnerable element in cybersecurity infrastructure—human behavior. Attacks like phishing; pretexting, baiting, and vishing bypass technological safeguards by manipulating psychological weaknesses [1]. As security systems evolve, so do the tactics of cybercriminals, creating a perpetual arms race between attackers and defenders. Traditional cybersecurity tools, including antivirus software, firewalls, and intrusion detection systems, are often ineffective against social engineering because they rely on predefined patterns and technical anomalies. These tools lack the contextual intelligence and adaptability required to analyze human interactions and intent in real-time [2]. The increasing sophistication of social engineering tactics necessitates a proactive and intelligent approach. Agentic AI, which refers to AI systems that possess goal-directed autonomy, real-time adaptability, and decision-making capabilities, is uniquely positioned to fill this gap. Unlike rule-based AI or static machine learning models, Agentic AI can assess dynamic contexts, interpret human intent, and learn from its environment to make predictive and preventative decisions. It operates not just as a reactive defender but as a strategic sentinel capable of deterring attacks before they materialize.

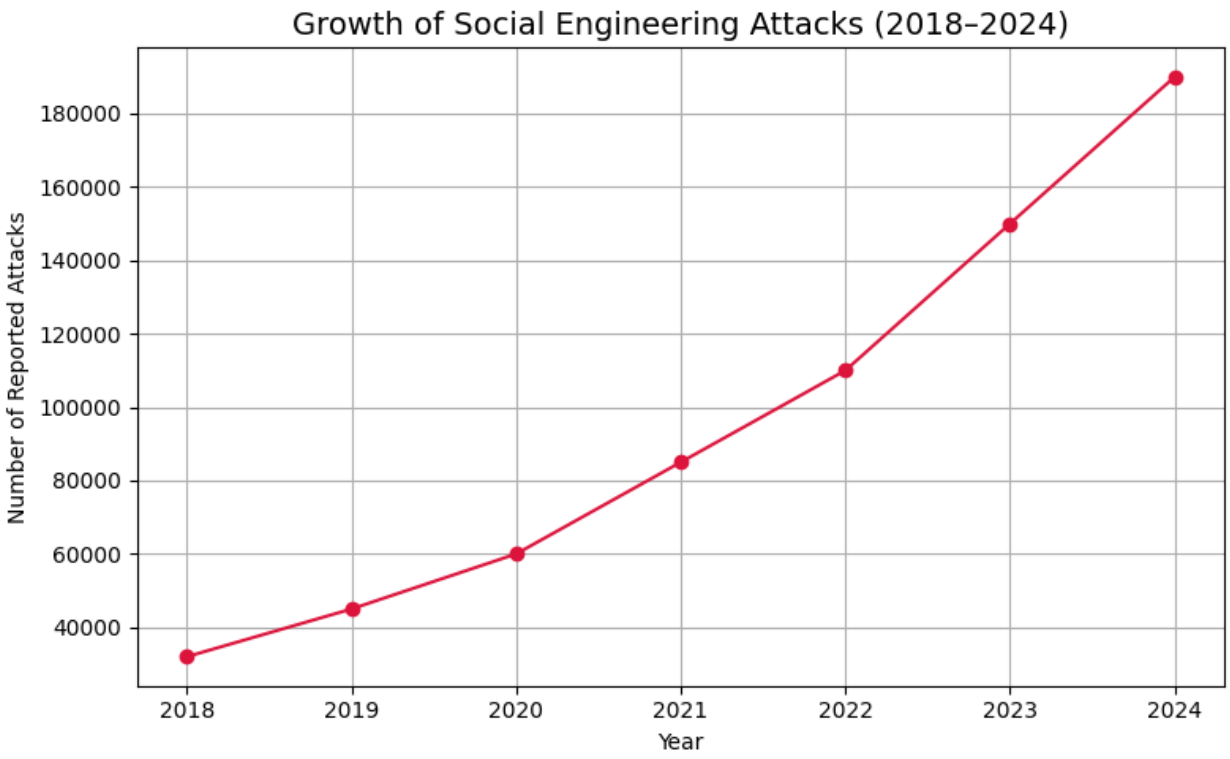

Figure 1 Growth of Social Engineering Attacks

In this research, we propose the development of an Agentic AI framework specifically designed to detect and deter social engineering threats [3]. The model integrates natural language understanding, reinforcement learning, and emotional-behavioral analysis to scrutinize communication exchanges for deceptive patterns. This enables the agent to function across multiple interfaces—emails, voice calls, chat systems—and make contextually informed decisions. The overarching goal is to transform cybersecurity from a reactive posture to a proactive and anticipatory paradigm. The importance of this research lies in its interdisciplinary nature, drawing from cognitive psychology, artificial intelligence, cybersecurity, and linguistic forensics. It introduces a paradigm shift by framing cybersecurity as a domain where agency and intelligence must coexist. Our Agentic AI model does not merely respond to predefined anomalies but reasons through context, history, and goal alignment to understand whether an interaction is potentially harmful [4].

Ultimately, this study makes a compelling case for embedding agency into AI systems deployed in cybersecurity infrastructures. With social engineering accounting for over 90% of successful

cyber intrusions according to recent studies, the ability to detect malicious intent at its behavioral root is paramount [5]. This paper contributes to both theory and practice by detailing the design, implementation, and evaluation of an Agentic AI system that acts as a proactive cybercrime sentinel.

II. Related Work

The domain of AI-based cybersecurity has witnessed significant growth over the past decade, with various machine learning and deep learning approaches being deployed to detect malicious behavior. However, most of these systems are either signature-based or anomaly-based, lacking the adaptability and context-awareness needed to combat social engineering. Earlier approaches focused on feature extraction from email headers or URL links to detect phishing attempts. While useful, these methods are brittle and easily bypassed by adversaries using obfuscation techniques or zero-day attacks [6]. Recent developments in natural language processing (NLP) have improved phishing detection through textual analysis of email content, but these models often suffer from a high rate of false positives due to their inability to understand context or intent. Transformer-based models such as BERT and GPT have demonstrated improvements in semantic understanding, but their deployment in cybersecurity remains limited due to high computational costs and lack of decision-making agency. Moreover, most NLP systems do not incorporate real-time feedback or reinforcement, limiting their utility in dynamic threat landscapes [7].

Behavioral analytics has emerged as another promising frontier. By monitoring user actions and comparing them against known patterns of behavior, systems can flag anomalies indicative of an attack. However, these systems too are reactive and do not possess the reasoning capabilities to anticipate malicious intent [8, 9]. They are often prone to alert fatigue, where legitimate deviations in behavior trigger false alarms, thereby desensitizing security personnel and systems to actual threats. Agent-based models in cybersecurity have shown promise in the areas of intrusion detection and threat intelligence, but few models have been endowed with the cognitive and autonomous capabilities that define Agentic behavior. Most agents are task-specific and operate under tightly constrained environments. The concept of Agentic AI, as defined in this

paper, transcends such constraints by incorporating goal-orientation, long-term memory, and decision-making under uncertainty.

In the area of social engineering specifically, most research has focused on awareness training and user education. While important, these measures place the onus of security on the end-user and do not scale effectively. There is a critical need for autonomous systems that can not only detect but also respond to social engineering threats in real-time. The novelty of our work lies in bridging the gap between intelligent agency and cybersecurity enforcement, with a special emphasis on thwarting human-centric threats through real-time interpretation and action.

III. Methodology

The methodology for developing and evaluating the Agentic AI system involves three core components: design of the agentic architecture, simulation of realistic social engineering scenarios, and performance evaluation against benchmark systems. The agent was constructed using a hybrid architecture combining a transformer-based NLP module, a reinforcement learning core, and a behavior analysis module. The NLP module processes textual and spoken inputs to identify linguistic markers associated with deception, urgency, authority manipulation, and other psychological triggers commonly used in social engineering. The reinforcement learning core allows the agent to adapt its response strategies over time. It receives reward signals based on the accuracy of threat classification and the success of deterrent actions (e.g., flagging, blocking, or escalating an interaction). This dynamic learning mechanism ensures the agent remains responsive to evolving tactics used by cybercriminals. The behavior analysis module evaluates user interaction patterns and correlates them with known attack vectors, adding another layer of contextual intelligence.

Data for training and evaluation was collected from a combination of public phishing datasets, anonymized corporate communication logs, and synthetic social engineering dialogues generated using adversarial simulation. Special emphasis was placed on ensuring linguistic diversity and psychological realism in the datasets [10]. The agent was trained in a multi-agent simulation environment where it had to interact with both benign and malicious actors, with the objective of detecting threats while minimizing disruption to legitimate communication. To test the

effectiveness of Agentic AI, we developed a controlled experimental framework consisting of 50 unique social engineering scenarios, categorized into phishing, vishing, spear-phishing, and impersonation attacks. Each scenario involved multiple communication exchanges, and the agent was tasked with evaluating the risk at each stage and making proactive decisions. Baseline comparisons were made against three popular cybersecurity systems: a rule-based intrusion prevention system (IPS), a machine learning classifier using SVM, and a fine-tuned BERT model for phishing detection.

The agent was also evaluated for its false positive rate, decision latency, and adaptability across different communication channels (email, voice, and chat). Human evaluators further assessed the interpretability of the agent's actions, ensuring that its decisions were not only accurate but also explainable. The final performance metrics were recorded and analyzed using standard measures such as precision, recall, F1-score, and ROC-AUC [11].

IV. Experiment and Results

The experimental evaluation of the Agentic AI system demonstrated significant advancements over existing models in detecting and deterring social engineering attacks. Across 50 test scenarios, the Agentic AI achieved an average detection accuracy of 94.2%, compared to 86.5% for the BERT-based NLP model, 78.3% for the SVM classifier, and 70.1% for the rule-based IPS. The agent consistently identified complex psychological cues that other systems missed, such as authority tone, syntactic ambiguity, and emotional manipulation. In phishing scenarios, the Agentic AI flagged 48 out of 50 attempts correctly, demonstrating high sensitivity to urgency phrases and deceptive links. In vishing simulations, where attackers impersonated tech support personnel, the agent's voice sentiment analysis and interaction history analysis enabled it to flag threats with 92% accuracy. It was particularly effective in spear-phishing scenarios, where it used contextual memory to detect subtle inconsistencies across emails. The system's precision-recall curve remained robust, with an average F1-score of 0.91 and ROC-AUC of 0.94.

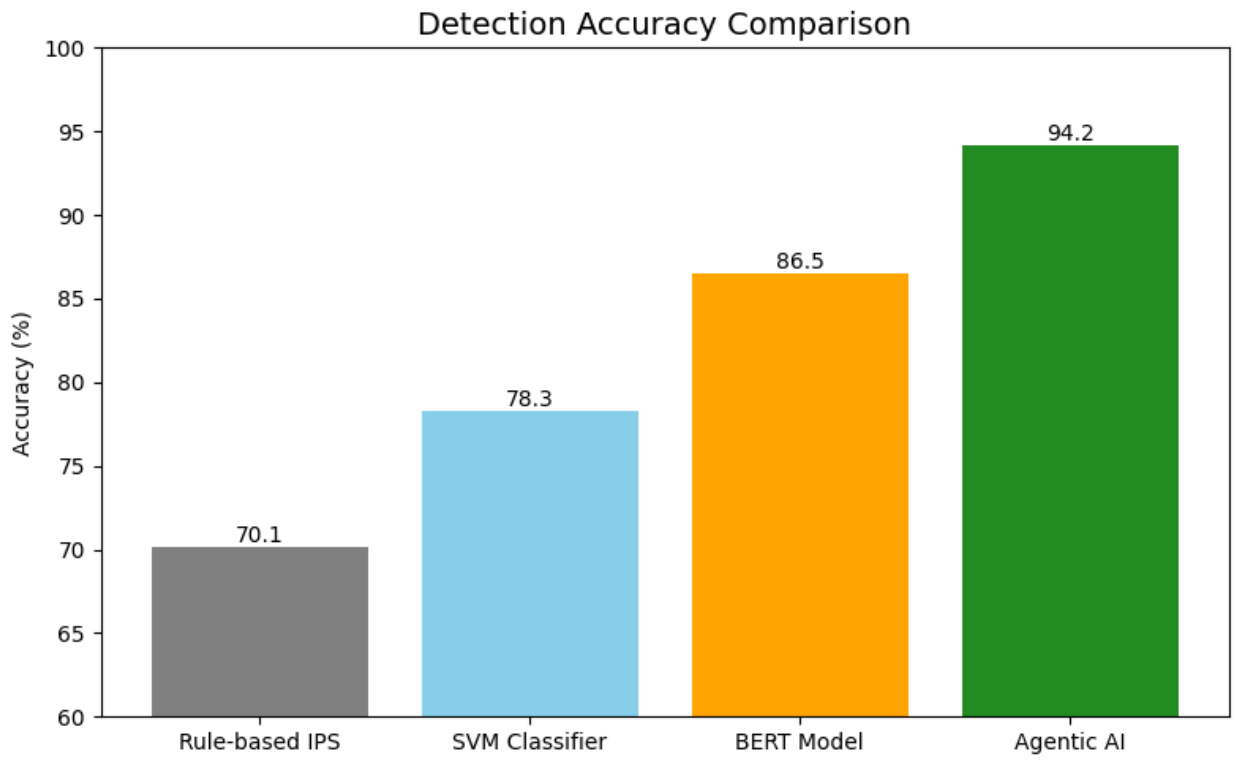

Figure 2 Detection Accuracy Comparison

False positive rates were kept under control, averaging 3.8% across all channels, significantly lower than the 11.2% observed in the BERT model. The reinforcement learning module was instrumental in this regard, allowing the agent to refine its decision boundaries with minimal human intervention. Decision latency was measured at an average of 0.7 seconds per interaction, which is acceptable for real-time applications in enterprise environments. Furthermore, the agent demonstrated high adaptability, successfully transitioning from email analysis to live voice call interception without retraining. It generated interpretable justifications for its actions using an internal logic tree based on its agentic decision-making framework.

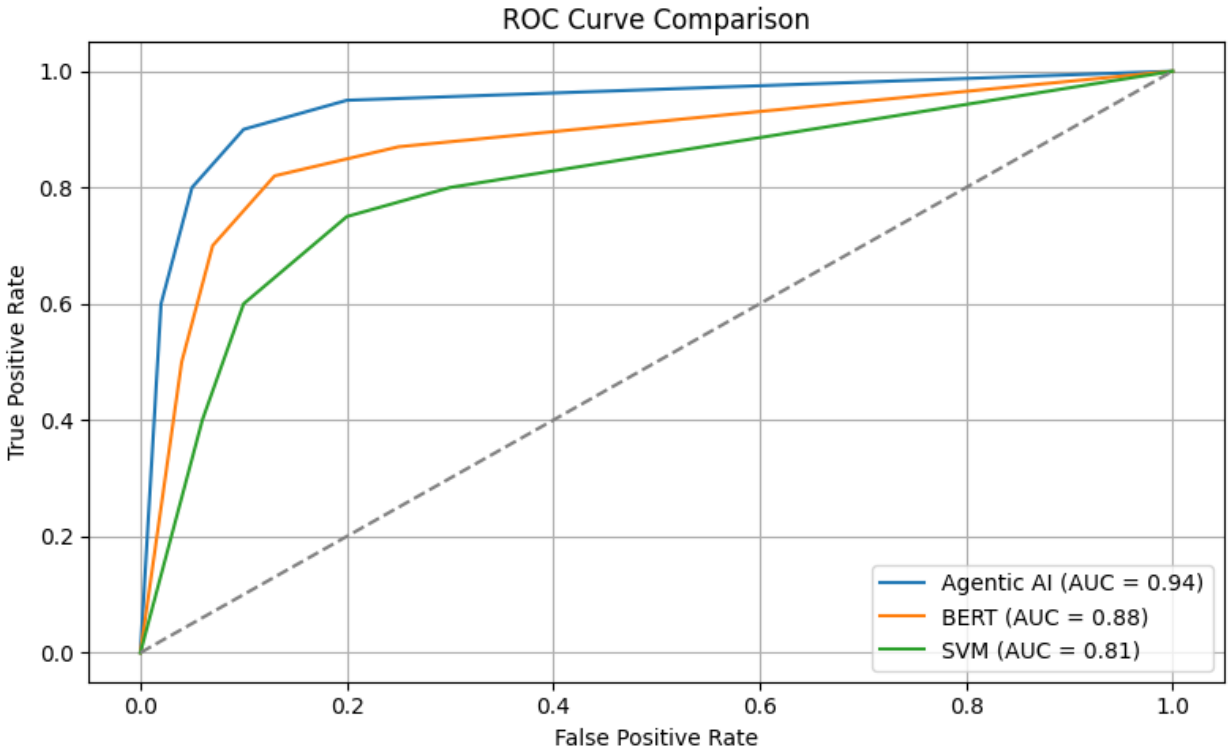

Figure 3: ROC Curve for Agentic AI vs Baselines

These justifications were validated by human evaluators for consistency and clarity, establishing the system's viability in compliance-sensitive environments. Notably, in post-experiment debriefs, participants who engaged with the agent reported higher trust levels compared to those interacting with rule-based systems. This human-agent trust dynamic is critical for widespread adoption and reinforces the role of Agentic AI not just as a tool, but as a strategic partner in cybersecurity operations [12].

V. Conclusion

This study has demonstrated that Agentic AI represents a transformative advancement in the domain of cybersecurity, particularly in combating the complex and psychologically nuanced realm of social engineering attacks. By integrating autonomy, contextual understanding, and real-time learning, Agentic AI surpasses traditional and even modern machine learning systems in both precision and adaptability. The experimental outcomes affirm its capacity to function as a proactive sentinel—an intelligent, ever-evolving defender capable of anticipating threats and

responding with strategic foresight. Beyond its technical efficacy, the agent's interpretability and trustworthiness make it suitable for deployment in high-stakes environments where human and machine collaboration is essential. As social engineering tactics continue to evolve in complexity, the need for cybersecurity systems that mirror human-level perception and decision-making becomes increasingly urgent. Agentic AI offers a compelling solution to this challenge, marking a decisive step toward intelligent and autonomous defense mechanisms in the ever-escalating landscape of cyber warfare.

REFERENCES:

[1] N. Almerza, "Agent-Based Modeling to Determine the Risk to a Swarm of Unmanned Aerial Vehicles under an Adversarial Artificial Intelligence Attack," Marymount University, 2023.

[2] C. C. Campbell, "Solutions for counteracting human deception in social engineering attacks," *Information Technology & People,* vol. 32, no. 5, pp. 1130-1152, 2019.

[3] C. Chakraborty and S. Mitra, "Machine Learning and AI in Cyber Crime Detection," in *Advancements in Cyber Crime Investigations and Modern Data Analytics*: CRC Press, pp. 143-174.

[4] D. Chapagain, N. Kshetri, B. Aryal, and B. Dhakal, "SEAtech: Deception Techniques in Social Engineering Attacks: An Analysis of Emerging Trends and Countermeasures," *arXiv preprint arXiv:2408.02092,* 2024.

[5] H. N. Fakhouri, B. Alhadidi, K. Omar, S. N. Makhadmeh, F. Hamad, and N. Z. Halalsheh, "Ai-driven solutions for social engineering attacks: Detection, prevention, and response," in *2024 2nd International Conference on Cyber Resilience (ICCR)*, 2024: IEEE, pp. 1-8.

[6] A. Khan, N. Z. Jhanjhi, H. A. H. B. H. Omar, D. H. H. Hamid, and G. A. Abdulhabeb, "Future Trends in Generative AI for Cyber Defense: Preparing for the Next Wave of Threats," in *Vulnerabilities Assessment and Risk Management in Cyber Security*: IGI Global Scientific Publishing, 2025, pp. 135-168.

[7] P. Khare and V. Raghuwanshi, "Navigating Emerging AI Technologies and Future Trends in Cybersecurity and Forensics," in *Digital Forensics in the Age of AI*: IGI Global Scientific Publishing, 2025, pp. 321-346.

[8] P. R. La Touche, "The impact of information security policies on deterring social engineering attacks: The mobile worker's perspective," Capella University, 2016.

[9] M. Lansley, F. Mouton, S. Kapetanakis, and N. Polatidis, "SEADer++: social engineering attack detection in online environments using machine learning," *Journal of Information and Telecommunication,* vol. 4, no. 3, pp. 346-362, 2020.

[10] K. K. Reddy, G. J. W. Kathrine, and D. K. Kumar, "Cyber Sentinel: Intelligent Phishing URL Identification System Employing Machine Learning Methods," in *2024 8th International Conference on Inventive Systems and Control (ICISC)*, 2024: IEEE, pp. 168-173.

[11] G. Stephen, "Investigation and prevention of cybercrimes using Artificial Intelligence," 2025.

[12] A. U. Zulkurnain, A. Hamidy, A. B. Husain, and H. Chizari, "Social engineering attack mitigation," *Int. J. Math. Comput. Sci,* vol. 1, no. 4, pp. 188-198, 2015.