The American Journal of
Engineering and Technology

ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 252-257

Cognitive Firewalls: Mitigating LLM-Powered Social Engineering Through Personality-Aware Behavioral Analytics and Automated Response Systemscc

Kenjiro S. Watanabe

Independent Researcher, LLM Threat Mitigation & Adaptive Behavioral Modelling, Tokyo, Japan

OPEN ACCESS

SUBMITTED 15 September 2025
ACCEPTED 24 September 2025
PUBLISHED 30 September 2025
VOLUME Vol.07 Issue 09 2025

CITATION
Kenjiro S. Watanabe. (2025). Cognitive Firewalls: Mitigating LLM-Powered Social Engineering Through Personality-Aware Behavioral Analytics and Automated Response Systemscc. The American Journal of Engineering and Technology, 7(09), 252–257. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/6943

Abstract: Background: The rapid advancement of Large Language Models (LLMs) has fundamentally altered the cybersecurity landscape, shifting the paradigm from technical exploitation to cognitive manipulation. Malicious actors now leverage generative AI to automate high-precision social engineering attacks, utilizing Open Source Intelligence (OSINT) to craft hyper-personalized spear-phishing campaigns at scale. Methods: This study analyzes the intersection of personality psychology—specifically the Big Five personality traits—and generative AI capabilities. We examine the theoretical framework of LLM-powered attacks, where attackers predict victim personality traits from digital footprints to tailor psychological triggers. We further evaluate the efficacy of counter-measures rooted in User Behavior Analytics (UBA) and AI-driven instantaneous response systems. Results: Analysis suggests that traditional signature-based detection systems are insufficient against LLM-generated content which lacks typical phishing indicators. However, defense mechanisms that integrate personality-aware baselines and behavioral anomaly detection demonstrate a higher potential for identifying synthetic social engineering attempts. Conclusion: We propose a "Cognitive Firewall" framework that combines psychological resilience training with automated, AI-driven behavioral monitoring. As LLMs lower the barrier for sophisticated attacks, defensive strategies must evolve to protect the human layer through proactive, context-aware algorithmic intervention.

Keywords: Large Language Models, Social Engineering, User Behavior Analytics, Big Five Personality Traits, Cybersecurity Automation, Cognitive Security, AI

Defense Strategies.

1. Introduction

The cybersecurity landscape is currently undergoing a seismic shift, driven by the democratization of Artificial Intelligence (AI) and, more specifically, the ubiquity of Large Language Models (LLMs). Historically, the most disruptive cyber threats were characterized by their technical volume or code complexity, such as Distributed Denial of Service (DDoS) attacks on e-commerce systems [1] or sophisticated malware injection techniques. However, the integration of Generative AI into the threat landscape has elevated the human element to the primary attack surface. The era of generic, bulk phishing emails—often riddled with syntactic errors and easily identifiable by rule-based filters—is rapidly receding. In its place, we face a new generation of "cognitive attacks": campaigns that are syntactically perfect, contextually aware, and psychologically optimized to exploit specific human vulnerabilities.

Recent research indicates that AI-powered threat countermeasures are becoming essential as the sophistication of attacks increases [2]. The core threat posed by LLMs in this domain is their ability to automate the "spear-phishing" process. Traditionally, spear-phishing required a human operator to painstakingly research a target, understand their organizational role, and craft a bespoke message. Today, an LLM can process vast amounts of Open Source Intelligence (OSINT) to infer a target's psychological profile and generate high-efficacy lures in seconds. As noted by Sen, Heim, and Zhu [3], the dual-use nature of AI presents both applications and challenges; while AI strengthens defense, it simultaneously provides adversaries with tools to bypass traditional security perimeters.

This paper addresses the critical gap in current defensive postures: the inability of technical controls to detect semantic and psychological manipulation. While organizations have invested heavily in preventing technical exploits—such as vulnerabilities in mobile banking infrastructure [4]—the defense against psychological exploitation remains reactive and training-focused. This reliance on human vigilance is increasingly untenable as AI-generated content becomes indistinguishable from legitimate human communication.

We propose a shift toward "Cognitive Firewalls"—defensive architectures that utilize User Behavior Analytics (UBA) and automated AI response systems to detect the subtle anomalies associated with social engineering. By synthesizing insights from personality psychology, specifically the Big Five traits, with advanced machine learning detection, we aim to define a defense strategy that protects the human user from the cognitive siege of weaponized LLMs.

2. Literature Review: The Convergence of AI and Psychology

The theoretical underpinnings of this research lie at the intersection of social psychology, cybersecurity, and artificial intelligence. To understand the threat of LLM-powered social engineering, one must first understand the mechanisms of influence and how they are currently being automated.

2.1 The Psychology of Influence and Deception

Social engineering relies heavily on the manipulation of cognitive biases. Cialdini's seminal work on the principles of social influence—reciprocity, commitment, social proof, authority, liking, and scarcity—remains the bedrock of understanding how attackers compromise human targets [5]. In a pre-AI context, applying these principles required human intuition. However, recent studies in deception detection have shown that personality traits play a significant role in an individual's susceptibility to these influence tactics. An et al. [6] demonstrated that deep learning techniques could be used for personality recognition to detect deception, suggesting a link between linguistic patterns and psychological states.

2.2 Personality Trait Profiling via Digital Footprints

The effectiveness of a social engineering attack is often correlated with how well it aligns with the victim's personality. The Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) provide a framework for this alignment. Azucar, Marengo, and Settanni [7] conducted a meta-analysis confirming that these personality traits can be predicted with high accuracy from digital footprints left on social media.

This capability creates a dangerous vector for attackers. By analyzing a target's public data, an algorithm can determine, for instance, that a target scores high in "Agreeableness" and "Conscientiousness." Anawar et al. [8] analyzed phishing susceptibility through this lens, finding that certain personality types are significantly more prone to specific types of manipulation. An attacker knowing this can instruct an LLM to frame a request that appeals to the victim's desire to be helpful (Agreeableness) and their adherence to rules (Conscientiousness), drastically increasing the probability of compromise.

2.3 Generative AI and the Automation of Malignancy

The capabilities of LLMs extend beyond mere text generation; they can simulate specific personas and emotional tones. Asfour and Murillo [9] harnessed LLMs to simulate realistic human responses to social engineering, highlighting how these models can be used to train attackers or, conversely, to automate the attacks themselves. Furthermore, the issue of toxicity and persona assignment in models like ChatGPT has been explored by Deshpande et al. [10], indicating that LLMs can be "jailbroken" or prompted to adopt malicious personas that bypass safety filters.

This automation capability enables what Evangelista et al. [11] describe as the systematic application of OSINT with Artificial Intelligence. The attacker no longer needs to manually parse a LinkedIn profile; the AI ingests the data, builds the psychological profile, and generates the attack vector, creating a scalable, automated threat pipeline.

2.4 Current Defense Paradigms: From Static to Behavioral

Traditional defense mechanisms, such as static blacklisting or signature-based detection, are largely ineffective against this dynamic threat. Liu, Zhang, and Wang [12] discussed solutions for DDoS attacks, but the logic of traffic volume does not apply to the low-volume, high-impact nature of social engineering. The industry is consequently moving toward User Behavior Analytics (UBA). Luo et al. [13] highlight that UBA is critical for identifying the subtle deviations in user activity that signal a compromised account or an insider threat.

Moreover, the speed of AI-driven attacks necessitates an equally rapid response. Matthias, Gadepalli, and Jain [14] argue for AI for instantaneous response to cybersecurity incidents. Relying on human analysts to triage alerts introduces a latency that automated attackers exploit. Sengupta, Kambhampati, and Chaudhuri [15] further support this, proposing AI-based response automation to mitigate cyberattacks in real-time. The consensus in the literature is clear: to fight AI-driven threats, defense systems must possess AI-driven autonomy.

3. Theoretical Framework: The Automated Social Engineering Lifecycle

To effectively counter LLM-powered social engineering, we must deconstruct the attack lifecycle. We posit a three-stage model that attackers utilize to weaponize generative AI: Automated Reconnaissance, Psychological Profiling, and Semantic Weaponization.

3.1 Automated Reconnaissance and OSINT Aggregation

The initial phase of any targeted attack is information gathering. In the context of AI-threat countermeasures [2], it is observed that LLMs are adept at synthesizing disparate data points. An attacker utilizes scripts to scrape public repositories—social media, corporate directories, and news articles. Unlike previous iterations of scraping, which sought email addresses or phone numbers, this phase seeks context. It ingests the target's writing style, their recent professional achievements, and their interaction patterns with colleagues. This data forms the raw material for the "context window" of the LLM.

3.2 Psychometric Mapping and Vulnerability Assessment

Once the data is aggregated, the attacker applies psychometric algorithms. Utilizing the methodologies described by Azucar et al. [7] and Danner et al. [16], the system processes the textual data (e.g., tweets, blog posts) to generate a Big Five profile.

For example, a target who posts frequently about industry innovations and uses complex vocabulary may be scored high on "Openness." A target who posts rigid schedules or policy updates may score high on "Conscientiousness." This profiling is not merely academic; it determines the attack vector. A high-Openness target might be targeted with a "novel opportunity" or "exclusive invite" lure, whereas a high-Neuroticism target might be targeted with a "security alert" or "urgent compliance warning" lure.

3.3 Semantic Weaponization and Execution

The final phase is the generation of the content. The attacker prompts the LLM with the context and the psychological profile. The prompt might look like: "Generate an email to [Target], who has high Conscientiousness and low Extraversion. Reference their recent project on [Project Name]. Use a tone of professional urgency but low social pressure. Request a password reset due to a policy update."

The LLM generates a message that is statistically likely to bypass the target's skepticism. Because the language model has been trained on vast corpora of corporate email, the syntax, tone, and structure are indistinguishable from legitimate internal communications.

4. Methodology

Having established the threat landscape, we propose a defensive architecture designed to intercept these cognitive attacks. This architecture does not rely on "detecting AI writing"—which is becoming increasingly difficult—but rather on detecting contextual anomalies and intent mismatches through User Behavior Analytics (UBA).

4.1 Data Ingestion and Baseline Establishment

The foundation of the Cognitive Firewall is comprehensive visibility. The system must ingest data from multiple streams: email gateways, instant messaging platforms (Slack, Teams), and endpoint logs. The goal is to establish a behavioral baseline for every identity within the organization.

This baseline is not static. It evolves using Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to understand the temporal sequence of a user's communication. We must know not just who the user emails, but how they email. What is their typical sentiment? What is their average sentence complexity? Do they typically respond to urgent requests immediately or after a delay?

4.2 The Detection Layer: Hybridizing NLP and UBA

The detection engine operates on two parallel tracks:

Track A: Semantic Anomaly Detection

This track utilizes Transformer-based models (e.g., BERT or RoBERTa) fine-tuned on the organization's internal corpus. It analyzes incoming messages for "semantic drift." If an email claiming to be from the CFO arrives, the model compares the linguistic fingerprint of the email against the CFO's historical corpus. Even if the LLM mimics the CFO's style, subtle deviations in topic distribution or lexical choice can trigger a low-confidence alert.

Furthermore, this layer analyzes the "intent" of the message. Using Natural Language Understanding (NLU), it categorizes the email's request (e.g., "Transfer Funds," "Reset Password," "Share File"). If a request for "Transfer Funds" is associated with a high-pressure urgency score (derived from Cialdini's scarcity principle [5]), the risk score increases.

Track B: Behavioral Contextualization

This track applies the principles of UBA [13]. It looks at the recipient's relationship with the sender. If the recipient has never interacted with the sender before, or if the interaction timing is anomalous (e.g., Sunday at 3 AM), the system flags the communication. Crucially, this layer integrates the "Phishing Susceptibility" metrics [8]. If the recipient is known to be highly susceptible (based on past simulation failures or personality profiling), the threshold for flagging suspicious content is automatically lowered for that specific user.

4.3 The Response Layer: AI-Driven Mitigation

Detection without response is observation, not defense. Following the insights of Matthias et al. [14] and Sengupta et al. [15], the response layer must be automated.

When a message exceeds a certain risk threshold, the system can execute several actions:

1. Quarantine: The message is held before reaching the user's inbox.

2. Banner Injection: The message is delivered, but with a prominent, dynamic warning banner explaining why it is suspicious (e.g., "This email claims to be from the CEO, but the linguistic style does not match the CEO's history").

3. Interactive Nudging: For ambiguous cases, the system can challenge the user via a separate channel (e.g., a push notification to their mobile app) asking, "Did you expect a file request from [Sender]?"

5. Discussion

The implementation of a Cognitive Firewall represents a significant step forward, but it also highlights the escalating "arms race" between attackers and defenders. As we deploy AI to detect attacks, adversaries will deploy AI to evade detection.

5.1 Evasion Techniques and Adversarial AI

The primary challenge to the proposed architecture is adversarial machine learning. Attackers may attempt to "poison" the baseline data. If an attacker has compromised a user's account, they might spend weeks sending benign emails to gradually alter the system's understanding of the user's "normal" behavior. This "frog-boiling" technique could allow them to eventually send a malicious payload that the UBA system recognizes as normal.

Furthermore, attackers can use "GANs" (Generative Adversarial Networks) to test their phishing emails against open-source detection models before deploying them. If the attacker knows the parameters of the defense model, they can iterate the attack vector until it finds a blind spot. This necessitates that defensive models be continuously retrained and that their exact parameters remain obfuscated.

5.2 Ethical Considerations of Deep Behavioral Monitoring

The proposed solution relies on deep surveillance of employee behavior. Analyzing the "Big Five" personality traits of employees to determine their security risk profile raises significant privacy concerns. While Anawar et al. [8] suggest this is necessary for understanding susceptibility, it risks crossing the line into intrusive psychological profiling.

Organizations must navigate this ethically by ensuring that the data is pseudonymized and used strictly for security purposes. There is a risk that such data could be misused for performance reviews or hiring decisions (e.g., firing employees with high "Neuroticism" scores). Strict governance frameworks and "Privacy-by-Design" principles must be embedded into the UBA architecture to preventing "function creep."

5.3 The Human-in-the-Loop Necessity

Despite the emphasis on automation [15], the human element remains relevant. Total automation of response runs the risk of "false positives" blocking legitimate, business-critical communications. If the AI aggressively quarantines emails from a new client because they deviate from the baseline, it disrupts business operations. Therefore, the system is best viewed as a "Decision Support System" rather than a purely autonomous agent. It augments human decision-making by highlighting risks that are invisible to the naked eye, but it may still require human oversight for edge cases.

5.4 Expansion on Integration with Mobile Ecosystems

It is imperative to recognize that social engineering is not confined to desktop email environments. McCray [4] highlights the vulnerabilities in mobile banking, noting that the mobile form factor (smaller screens, distracted usage) increases susceptibility to fraud. LLM-powered "Smishing" (SMS Phishing) attacks are particularly dangerous because mobile operating systems often hide the full URL path and users are conditioned to respond quickly to text messages.

The Cognitive Firewall must therefore extend to the mobile endpoint. This involves integrating with Mobile Device Management (MDM) solutions to analyze SMS and messaging app metadata. The behavioral analytics engine must account for "cross-channel" attacks, where an attacker sends a phishing email and follows up with an SMS (generated by an LLM) to create a false sense of urgency and multi-modal verification. Defending the mobile perimeter requires a lightweight version of the detection model that can operate with the resource constraints of mobile devices while maintaining synchronization with the central UBA core.

5.5 The Future of "Defensive AI" in Personality Computing

Looking ahead, we anticipate the rise of "Defensive AI Personas." Just as attackers use LLMs to simulate trusted figures, defenders can use LLMs to simulate "decoy" employees. These decoys can engage with attackers, wasting their resources and gathering intelligence on their tactics. If an attacker attempts to profile a decoy, the decoy can emit "poisoned" personality signals, confusing the attacker’s profiling algorithms. This proactive deception turns the table on the attacker, using the very principles of personality computing to disrupt the social engineering kill chain.

6. Conclusion

The integration of Large Language Models into the cybercriminal toolkit has rendered traditional social engineering defenses obsolete. The ability of adversaries to automate the generation of syntactically perfect, psychologically targeted attack vectors at scale represents a critical threat to organizational security. As we have explored, these attacks exploit the cognitive biases of human users, leveraging OSINT and personality profiling to bypass skepticism.

This paper argues that the only viable defense against AI-driven cognitive attacks is an AI-driven cognitive defense. By implementing "Cognitive Firewalls" that utilize User Behavior Analytics, Natural Language Understanding, and automated response protocols, organizations can detect the subtle anomalies that characterize synthetic influence operations. While challenges regarding adversarial evasion and privacy ethics remain, the necessity of this shift is undeniable.

We are entering an era where the battle is no longer just code against code, but cognition against cognition. To secure the future, we must fortify the human mind with the speed and insight of machine intelligence. The integration of personality-aware analytics into the security operations center is not merely an enhancement; it is the requisite evolution for survival in the age of the weaponized LLM.

References

1. Rajgopal, P. R. . (2025). AI Threat Countermeasures: Defending Against LLM-Powered Social Engineering. International Journal of IoT, 5(02), 23-43. https://doi.org/10.55640/ijiot-05-02-03
2. Liu, X., Zhang, X., & Wang, C. (2020). "DDoS Attacks on E-Commerce Systems: Vulnerabilities and Solutions." Journal of Cybersecurity, 8(2), 109-119.
3. Luo, X., Brody, R., Seazzu, A., & Burd, S. (2018). User behavior analytics: Applications and challenges in cybersecurity. Computers & Security, 74, 93-111.
4. Matthias, G., Gadepalli, N., & Jain, A. (2021). "AI for Instantaneous Response to Cybersecurity Incidents." Journal of Cybersecurity and Privacy, 5(2), 101-115.
5. McCray, K. L. (2023). Vulnerabilities and Threats in Mobile Banking that Financial Institutions Must Understand to Reduce Mobile Banking Fraud (Doctoral dissertation, Marymount University).
6. Sen, R., Heim, G., & Zhu, Q. (2022). Artificial intelligence and machine learning in cybersecurity: Applications, challenges, and opportunities for mis academics. Communications of the Association for Information Systems, 51(1), 28.
7. Sengupta, S., Kambhampati, S., & Chaudhuri, B. (2020). "Cyberattack Mitigation with AI-Based Response Automation." Journal of Network and Computer Applications, 164, 102706.
8. An, G., Levitan, S. I., Hirschberg, J., & Levitan, R. (2018). Deep personality recognition for deception detection. In INTERSPEECH, pp. 421–425.
9. Anawar, S., Kunasegaran, D. L., Mas'ud, M. Z., Zakaria, N. A., et al. (2019). Analysis of phishing susceptibility in a workplace: a big-five personality perspectives. J Eng Sci Technol, 14(5), 2865–2882.
10. Asfour, M., & Murillo, J. C. (2023). Harnessing large language models to simulate realistic human responses to social engineering attacks: A case study. International Journal of Cybersecurity Intelligence & Cybercrime, 6(2), 21–49.
11. Evangelista, J.R.G., Sassi, R.J., Romero, M., Napolitano, D. (2021). Systematic literature review to investigate the application of open source intelligence (osint) with artificial intelligence. Journal of Applied Security Research 16(3), 345–369.
12. Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and Individual Differences, 124, 150–159.
13. Cialdini, R.B. (1995). Principles and techniques of social influence. Advanced Social Psychology, 256, 281.
14. Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C., & John, O.P. (2016). Die deutsche version des big five inventory 2 (bfi-2). Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).
15. Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335.