The American Journal of
Engineering and Technology

ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 178-184

Check for updates

Psychological Surface Vectors: Mitigating Large Language Model-Driven Social Engineering via Behavioral Anomaly Detection

Dr. Elena V. Rostova,

Department of Computational Security, Moscow, Russia

OPEN ACCESS

SUBMITTED 02 October 2025
ACCEPTED 14 October 2025
PUBLISHED 31 October 2025
VOLUME Vol.07 Issue 10 2025

CITATION

Dr. Elena V. Rostova. (2025). Psychological Surface Vectors: Mitigating Large Language Model-Driven Social Engineering via Behavioral Anomaly Detection. The American Journal of Engineering and Technology, 7(10), 178–184. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/6944

Abstract:

Context: The proliferation of Large Language Models (LLMs) has lowered the barrier to entry for sophisticated social engineering attacks. Adversaries can now automate the inference of psychological traits from user data to generate highly persuasive, targeted phishing content.

Problem: Traditional cybersecurity defenses, such as signature-based Intrusion Detection Systems (IDS) and standard spam filters, are increasingly ineffective against these syntactically perfect and contextually aware AI-generated attacks. They fail to detect the subtle semantic anomalies that characterize algorithmic psychological manipulation.

Method: This study investigates the efficacy of an unsupervised learning framework designed to detect behavioral anomalies in email communications. We simulated an LLM-driven attack campaign that tailors phishing narratives to the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). We then evaluated a hybrid detection model combining Long Short-Term Memory (LSTM) networks for sequence analysis and Isolation Forests for anomaly scoring.

Results: The simulation demonstrated that personality-aligned LLM attacks achieved a theoretical click-through rate 40% higher than generic phishing. However, the proposed behavioral anomaly detection system identified 88.5% of these sophisticated attacks by analyzing deviations in semantic density and communication patterns, outperforming traditional keyword-based filters which detected only 34%.

Conclusion: While LLMs significantly enhance the lethality of social engineering, analyzing the "psychological surface" of communication via unsupervised learning offers a robust countermeasure. Future defense architectures must move beyond content analysis to context and behavioral intent analysis.

Keywords

Large Language Models, Social Engineering, Behavioral Anomaly Detection, Big Five Personality Traits, Cybersecurity, Unsupervised Learning, Phishing Detection.

Introduction

The cybersecurity landscape is currently undergoing a paradigm shift driven by the democratization of artificial intelligence. For decades, the primary vector for network infiltration has been the human element—the susceptibility of individuals to deception. Historically, social engineering attacks were labor-intensive endeavors requiring significant manual reconnaissance to be effective. Attackers relying on generic, "spray-and-pray" phishing templates often failed due to poor grammar, lack of context, or obvious formatting errors. However, the advent of Large Language Models (LLMs) has fundamentally altered this dynamic, introducing a scale and sophistication to digital deception that traditional defenses are ill-equipped to handle.

Recent research indicates that LLMs are not merely text generators; they are capable of analyzing vast amounts of unstructured data to infer sensitive psychological attributes. Studies suggest that models can accurately infer a user's personality, political leanings, and even mental health status from free-form interaction data [11], [12]. When weaponized, this capability allows threat actors to automate "spear-phishing"—the practice of sending highly targeted emails—at a scale previously reserved for generic spam. This creates a threat landscape where every employee in an organization can be targeted with a unique, psychologically tailored narrative designed to bypass their specific cognitive biases.

The challenge is compounded by the limitations of current detection technologies. Traditional Intrusion Detection Systems (IDS) and email gateways rely heavily on signature matching and blacklists [7]. They look for known malicious URLs, specific keywords, or attachment hashes. However, LLM-generated content is polymorphic; it can be rephrased infinitely while retaining the malicious intent, effectively rendering signature-based detection obsolete. Furthermore, the "closed world" assumption in machine learning security—where the training data is assumed to represent the testing distribution—fails when attackers use generative AI to create novel, "zero-day" social engineering vectors [2].

This paper proposes a shift from content-based detection to behavioral anomaly detection. By leveraging unsupervised learning techniques to model the baseline communication patterns and psychological "texture" of legitimate organizational traffic, we can identify the subtle deviations characteristic of AI-generated manipulation [3], [18]. We explore the intersection of the Big Five personality framework [14] and generative AI to understand how attackers exploit psychological vulnerabilities and how defenders can use those same signals to identify threats.

Related Work

AI in Intrusion Detection and Anomaly Detection

The application of machine learning to network security is well-documented. Sommer and Paxson [7] provided early critiques of applying machine learning to network intrusion detection, noting the difficulty of outlier detection in high-dimensional spaces where "normal" traffic is highly variable. However, recent advances have reinvigorated this field. Liu et al. [3] demonstrated the efficacy of unsupervised learning in real-time anomaly detection within eCommerce environments, a domain sharing similarities with email traffic regarding volume and velocity. Furthermore, the use of deep learning to identify complex attack patterns has shown promise, though it remains vulnerable to adversarial examples where small perturbations in input data cause misclassification [10].

The Psychology of Deception and OSINT

Social engineering relies on exploiting human cognitive heuristics. The Social Engineering Personality Framework (SEPF) posits that individual susceptibility to persuasion varies significantly based on personality traits [17]. For instance, individuals with high "Agreeableness" may be more compliant with requests for help, while those with high "Neuroticism" may react more impulsively to fear-based appeals.

Open Source Intelligence (OSINT) has traditionally been the method for gathering the data necessary to build these profiles [16]. Szymoniak and Foks [16] highlight that the proliferation of social media has created a "digital exhaust" that, when analyzed, provides a high-fidelity map of an individual's life and psychology.

Generative AI and Personality Inference

The most critical recent development is the ability of LLMs to process this OSINT data. Peters et al. [11] and Peters & Matz [12] have demonstrated that LLMs can infer psychological dispositions from social media users with startling accuracy, often correlating strongly with self-reported Big Five inventories. This implies that an attacker need not manually analyze a target; they can simply feed a target's Twitter or LinkedIn history into an LLM and request a phishing email optimized for the inferred personality type. Schmitt and Flechais [13] describe this as "digital deception," noting that generative AI removes the linguistic cues (typos, awkward phrasing) that previously alerted users to potential fraud.

Methodology

To evaluate the threat and the proposed defense, we constructed a simulation environment known as the "Poly-Phish" framework. This framework consists of three components: the LLM-Attacker (Generation Module), the Target Simulation (Psychological Profiling), and the Behavioral Defense Engine (Detection Module).

The LLM-Attacker Model

We utilized a commercially available frontier LLM (GPT-4 architecture equivalent) to serve as the attacker. The model was prompted to function as a sophisticated social engineer. It was provided with synthetic OSINT profiles derived from the Enron Email Dataset, augmented with simulated social media metadata.

The prompting strategy utilized "persona adoption." For a target identified as having high Conscientiousness (efficient, organized, dutiful), the LLM was instructed: "Draft a spear-phishing email regarding an urgent compliance audit. Use formal language, reference specific policy numbers, and appeal to the target's desire for order and accuracy." Conversely, for a target with high Extraversion, the prompt was: "Draft an email regarding a networking event or team celebration. Use enthusiastic, informal language and appeal to the target's fear of missing out (FOMO)."

Theoretical Framework: The Big Five Integration

We mapped the attack vectors to the Big Five traits [14]:

- Openness: Appeals to curiosity, innovation, and new technology beta tests.
- Conscientiousness: Appeals to duty, compliance, and error correction.
- Extraversion: Appeals to social connection, events, and visibility.
- Agreeableness: Appeals to helpfulness, charity, and mentoring requests.
- Neuroticism: Appeals to urgency, security warnings, and fear of repercussions.

The Defense Architecture

The defense model moves beyond simple NLP classification. Instead of asking "Is this email malicious?", the system asks "Is this email anomalous for this context?"

We implemented an unsupervised learning approach inspired by Liu et al. [3] and Sommers & Barford [8]. The model features two stages:

1. Feature Extraction: We utilize a pre-trained BERT model to extract semantic embeddings from the email body. Additionally, we extract metadata features: timestamp, sender-recipient graph distance, and syntactic complexity scores.

2. Anomaly Detection: These features are fed into an Isolation Forest algorithm. Isolation Forests work by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Since anomalies are few and different, they are easier to "isolate" (require fewer splits) than normal points.

This approach creates a "normality profile" for internal organizational communication. LLM-generated text, while grammatically perfect, often exhibits a distinct "perplexity" signature and semantic density that differs from human-written corporate emails.

Results

Efficacy of Psychologically Targeted Attacks.The simulation results indicated a profound increase in theoretical susceptibility when LLMs utilized personality profiling. In baseline tests using generic phishing templates (e.g., "Password Reset Required"), the estimated Click-Through Rate (CTR) based on historical susceptibility heuristics was approximately 4.2%.

When the LLM tailored the text to the specific Big Five trait of the target, the estimated CTR rose to 18.7% on average. The highest susceptibility was found in the "High Agreeableness" cohort when presented with a "Help Request" scenario (e.g., "I'm locked out and need you to forward this file"). The LLM's ability to mimic the tone of a distressed colleague was indistinguishable from genuine requests in 92% of the sample cases.

Failure of Traditional Detection

We ran the generated emails through a standard open-source spam filter (SpamAssassin) and a keyword-based heuristic engine. These systems detected only 34% of the LLM-generated phishing emails. The failure was largely due to the lack of "spam markers." The LLM did not use broken HTML, excessive capitalization, or known bad domains (as the simulation assumed the use of fresh, zero-day domains). The "cleanliness" of the AI-generated text effectively bypassed filters designed for human-error-prone scams.

Performance of Behavioral Anomaly Detection

The proposed Isolation Forest model performed significantly better. It achieved a detection rate of 88.5% on the LLM-generated dataset. The model successfully flagged the emails not because they contained "malicious" words, but because their semantic structure and request patterns deviated from the baseline.

For example, an email sent to a junior employee (High Conscientiousness) demanding an urgent wire transfer usually triggers a metadata anomaly. However, the LLM attack mitigated this by mimicking a "Compliance Officer." The unsupervised model still detected it because the semantic embedding of the request did not match the historical communication cluster of actual compliance officers within the training data.

Discussion

5.1. The Mechanism of Influence: Why Personality Profiling Works

To understand the severity of the threat, we must analyze the mechanism by which LLMs achieve such high persuasion rates. The effectiveness of these attacks lies in the "Lexical Hypothesis," which suggests that the personality traits most important in a person's life eventually become a part of their language. LLMs, having been trained on internet-scale text, have implicitly learned these correlations.

When an LLM generates a phishing email targeting a "High Openness" individual, it does not merely insert keywords; it alters the syntactic complexity and metaphorical density of the text. It mimics the cognitive style of the target. A user with high Openness tends to value intellectual engagement and novelty. The LLM creates an attack vector disguised as a "unique opportunity" or an "innovative beta test." The user’s cognitive defense mechanism—which typically scans for threats—is bypassed because the incoming message aligns perfectly with their intrinsic motivations. This is a form of "confirmation bias" weaponized. The user wants the message to be true because it validates their self-concept.

Furthermore, the "High Agreeableness" vector exploits the social contract. In corporate environments, cooperation is a survival trait. An LLM that generates a message mimicking a stressed superior asking for a "quick favor" triggers a distinct psychological pressure. Traditional phishing often fails here because the tone is too aggressive or the request is too absurd. The LLM, however, can modulate the "temperature" of the request to be just firm enough to induce compliance but polite enough to avoid suspicion. This nuance is what Peters et al. [12] refer to when discussing the inference of psychological dispositions; the model generates the inverse of the inference to create a "lock and key" fit between the deception and the target.

Recent incidents analyzed by Siddiqui et al. [6] confirm that NLP-based phishing detection struggles with this high-level semantic mimicry. The algorithms are trained to detect "urgency" or "threats," but they are not trained to detect "excessive alignment with personality traits." This is where the gap lies. The LLM is not just generating text; it is performing a "cognitive empathy" attack. It predicts what the user feels is a reasonable request.

The "Black Box" of AI vs.AI The conflict between LLM-based attackers and ML-based defenders represents a new frontier in the "adversarial arms race." We are entering a phase of "AI vs. AI" conflict. As noted by Liu et al. [4], attackers are beginning to use reinforcement learning (RL) to probe defense systems. In our context, an advanced attacker could use an RL agent to iteratively send emails to a defense system, learning which personality triggers bypass the Isolation Forest and which are flagged.

This leads to the issue of "model explainability" raised by Szegedy et al. [10]. While our unsupervised learning model proved effective (88.5% detection), the Isolation Forest is somewhat opaque. It tells us that an email is anomalous, but not always why. In a security operations center (SOC), an analyst needs to know if an email was flagged because of its timestamp, its sentiment, or its semantic incongruity. If the system flags a legitimate email from a CEO who is simply in a bad mood (creating a sentiment anomaly), it creates "alert fatigue." Therefore, the future of this defense lies not just in accuracy, but in interpretability.

The Privacy-Security Paradox

The implementation of such deep behavioral profiling for defense introduces significant ethical concerns, echoing the work of Shokri and Shmatikov [5] on privacy-preserving deep learning. To detect a deviation from "normal behavior," the system must know intimate details about what constitutes "normal" for every employee.

This creates a paradox: to protect employees from psychological manipulation by external AI, the organization must subject them to psychological profiling by internal AI. Monitoring an employee's communication style, work hours, and sentiment shifts to build a baseline for the Isolation Forest borders on surveillance. If the model learns that Employee A usually writes short, angry emails on Mondays, and suddenly writes a long, polite one (flagging an anomaly), the system has effectively inferred a behavioral trait. Organizations must navigate this by implementing privacy-preserving techniques, such as Federated Learning, where the model is trained on decentralized data without raw emails leaving the user's device, or by using differential privacy noise to mask individual contributions to the baseline model.

Regulatory and Compliance Implications

Srinivas et al. [9] reviewed security threats in cloud computing, emphasizing the role of compliance. As AI-driven social engineering becomes prevalent, regulatory frameworks (like GDPR or CCPA) may need to evolve. If an organization fails to implement AI-specific defenses and suffers a breach via an LLM-generated email, could they be liable for "negligent security" given the known state of the art?

Furthermore, the use of employee data to train these defense models falls under strict data processing scrutiny. The "purpose limitation" principle of GDPR requires that data collected for one purpose (business communication) not be used for another (behavioral profiling) without consent. This legal friction may slow the adoption of the very tools needed to stop these advanced attacks.

Limitations and Zero-Day Threats

Our study relied on synthetic data and simulations. While the "Poly-Phish" framework is robust, real-world human behavior is noisier than the Enron dataset suggests. Real humans have "mood swings" that might trigger false positives in an anomaly detection system.

Additionally, Liang and Zhao [2] discuss "Zero-Day" threat detection. We must acknowledge that LLMs are evolving faster than defense models. A new generation of "agentic" AI could perform multi-stage social engineering—engaging in a long email thread to build trust before delivering the payload. Our current anomaly detection model looks at individual emails or short sequences. It may struggle to detect a "slow-burn" attack where the deviation from the norm is introduced so gradually that the unsupervised model adapts to it, effectively being "poisoned" by the attacker.

Conclusion

The democratization of Large Language Models has fundamentally altered the threat landscape of social engineering. By automating the analysis of OSINT and the generation of psychologically targeted narratives, threat actors can now launch spear-phishing campaigns with the scale of spam and the efficacy of human spies. The results of this study demonstrate that traditional signature-based defenses are insufficient against these "psychological surface vectors."

However, the very reliance of these attacks on psychological manipulation provides a new avenue for detection. By utilizing unsupervised learning algorithms like Isolation Forests to model the behavioral and semantic baselines of organizational communication, we can detect the subtle anomalies introduced by generative AI. While this "AI vs. AI" approach raises significant privacy and ethical questions, it appears to be the only viable path forward. Future research must focus on adversarial hardening of these defense models and the development of privacy-preserving architectures that can secure the human element without compromising human rights.

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. Liang, X., & Zhao, J. (2020). "Towards Better Zero-Day Threat Detection." IEEETransactions on Information Forensics and Security, 15, 1381-1392.

3. Liu, F., Huang, X., & Zhang, Y. (2020). Real-time anomaly detection in eCommerce usingunsupervised learning. IEEE Transactions on Industrial Informatics, 16(8), 5435-5442.

4. Liu, Q., Yang, Y., Ding, M., Guo, W., Wang, Q., & Jin, S. (2022). Reinforcement learning anddeep learning-based attacks on network intrusion detection systems. Journal of Network andComputer Applications, 210, 103512.

5. Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In Proceedings of the22nd ACM SIGSAC conference on computer and communications security (pp. 1310-1321).

6. Siddiqui, M. A., Alam, M., & Raza, M. (2019). "Detecting Phishing Emails Using AI and NLPTechniques." Cybersecurity and AI, 6(4), 297-309.

7. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning fornetwork intrusion detection. In 2010 IEEE Symposium on Security and Privacy (pp. 305-316).

8. Sommers, J., & Barford, P. (2012). Analyzing network traffic anomalies. Communications ofthe ACM, 55(9), 57-64.

9. Srinivas, M., Reddy, G. R., & Govardhan, A. (2019). "A Review on Security Threats andVulnerabilities in Cloud Computing." Journal of Cyber Security and Mobility, 8(3), 345-367.

10. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R.(2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

11. Peters, H., Cerf, M., Matz, S.C.: Large language models can infer personality from free-form user interactions. arXiv preprint arXiv:2405.13052 (2024)

12. Peters, H., Matz, S.C.: Large language models can infer psychological dispositions of social media users. PNAS nexus 3(6), pgae231 (2024)

13. Schmitt, M., Flechais, I.: Digital deception: Generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review 57(12), 1–23 (2024)

14. Soto, C.J., John, O.P.: The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. Journal of Personality and Social Psychology 113(1), 117 (2017)

15. Stachl, C., Au, Q., Schoedel, R., Gosling, S.D., Harari, G.M., Buschek, D., Völkel, S.T., Schuwerk, T., Oldemeier, M., Ullmann, T., et al.: Predicting personality from patterns of behavior collected with smartphones. Proceedings of the National

Academy of Sciences 117(30), 17680–17687 (2020)

16. Szymoniak, S., Foks, K.: Open source intelligence opportunities and challenges–a review. Advances in Science and Technology. Research Journal 18(3) (2024)

17. Uebelacker, S., Quiel, S.: The social engineering personality framework. In: 2014 Workshop on Socio-Technical Aspects in Security and Trust. pp. 24–30. IEEE (2014)Uebelacker, S., Quiel, S.: The social engineering personality framework. In: 2014 Workshop on Socio-Technical Aspects in Security and Trust. pp. 24–30. IEEE (2014)