

8-30-2023

Harnessing Large Language Models to Simulate Realistic Human Responses to Social Engineering Attacks: A Case Study

social engineering attacks; simulating phishing attacks; large language models (LLMs); LLMs for social engineering; generative AI; AI for cybersecurity

Follow this and additional works at: <https://vc.bridgew.edu/ijcic>



Part of the [Criminology Commons](#), [Criminology and Criminal Justice Commons](#), [Forensic Science and Technology Commons](#), and the [Information Security Commons](#)

Recommended Citation

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. Available at: <https://vc.bridgew.edu/ijcic/vol6/iss2/3>
Copyright © 2023 Mohammad Asfour and Juan Carlos Murillo

This item is available as part of Virtual Commons, the open-access institutional repository of Bridgewater State University, Bridgewater, Massachusetts.
Copyright © 8-30-2023 Mohammad Asfour and Juan Carlos Murillo

Harnessing Large Language Models to Simulate Realistic Human Responses to Social Engineering Attacks: A Case Study

Mohammad Asfour, M.Eng, Cornell University, U.S.A.
 Juan Carlos Murillo*, MBA., Cornell University, U.S.A.

Keywords: social engineering attacks; simulating phishing attacks; large language models (LLMs); LLMs for social engineering; generative AI; AI for cybersecurity

Abstract:

The research publication, “Generative Agents: Interactive Simulacra of Human Behavior,” by Stanford and Google in 2023 established that large language models (LLMs) such as GPT-4 can generate interactive agents with credible and emergent human-like behaviors. However, their application in simulating human responses in cybersecurity scenarios, particularly in social engineering attacks, remains unexplored. In addressing that gap, this study explores the potential of LLMs, specifically the Open AI GPT-4 model, to simulate a broad spectrum of human responses to social engineering attacks that exploit human social behaviors, framing our primary research question: How does the simulated behavior of human targets, based on the Big Five personality traits, responds to social engineering attacks? . This study aims to provide valuable insights for organizations and researchers striving to systematically analyze human behavior and identify prevalent human qualities, as defined by the Big Five personality traits, that are susceptible to social engineering attacks, specifically phishing emails. Also, it intends to offer recommendations for the cybersecurity industry and policymakers on mitigating these risks. The findings indicate that LLMs can provide realistic simulations of human responses to social engineering attacks, highlighting certain personality traits as more susceptible.

Introduction

The rapid progression of digital technology has manifested a dual nature, bringing about significant advancements in various domains while simultaneously introducing new threats in the cybersecurity landscape. Among these, social engineering attacks, which encompass manipulative tactics such as phishing, baiting, and pretexting aimed at exploiting human vulnerabilities (Mitnick & Simon, 2002), have emerged as a major concern. The proliferation and sophistication of these attacks are escalating over time. These attacks capitalize on human vulnerabilities rather than system weaknesses, emphasizing the necessity for a comprehensive exploration of the manipulative tactics applied in the cyber domain and the implementation of effective countermeasures.

Generative AI, represented by large language models (LLMs) such as OpenAI’s GPT-4, uses machine learning algorithms to generate data similar to the data it was trained on (Radford et al., 2019). Specifically, GPT-4, an exemplar of these models, utilizes a transformer-based architecture to generate remarkably human-like text. Its vast training on diverse Internet text enables the model to produce contextually relevant outputs, extending its influence across diverse sectors, including cybersecurity (Bubeck et al., 2023). However, these models, while powerful, have associated challenges including the potential to be manipulated for malicious purposes. As the precision of these models in simulating human behavior improves, they present groundbreaking opportunities for research, especially in modeling and predicting human responses to social engineering attacks.

*Corresponding author

Juan Carlos Murillo, MBA in Information Technology, Cornell University, Ithaca, NY 14850, U.S.A.
 Email: jm2439@cornell.edu

Reproduction, posting, transmission or other distribution or use of the article or any material therein, in any medium as permitted by written agreement of the International Journal of Cybersecurity Intelligence and Cybercrime, requires credit to the Journal as follows: “This Article originally appeared in International Journal of Cybersecurity Intelligence and Cybercrime (IJCIC), 2023 Vol. 6, Iss. 2, pp. 21-49” and notify the Journal of such publication.

© 2023 IJCIC 2578-3289/2023/08

Given the paramount importance of understanding and counteracting social engineering attacks within the cybersecurity field, this research (Asfour & Murillo, 2023) aims to utilize the proficiencies of LLMs to simulate credible human responses to social engineering threats. This research draws inspiration from the groundbreaking study by Stanford and Google researchers, titled “Generative Agents: Interactive Simulacra of Human Behavior” (Park et al., 2023). Their project, which originally investigated how LLMs can mimic interactive human behavior, laid the foundation for this research’s innovative approach to studying social engineering and human vulnerabilities in a systematic, scalable, and controlled environment.

This research is motivated by this intersection of AI, personality research, and cybersecurity and will provide an overview of the current state of social engineering attacks, the potential applications of LLMs in cybersecurity research, and a detailed description of the methodology employed for simulating human target’s behavior and executing social engineering attacks. It will also present the findings and discuss their implications for cybersecurity. Building on the foundation of existing research, this study addresses the question “How does the simulated behavior of human targets, based on the Big Five personality traits, responds to social engineering attacks?” It seeks to contribute with its findings to the field of social engineering, to the exploration of human vulnerabilities, and to help organizations formulate more effective cybersecurity defenses for threat mitigation.

Background

Artificial Intelligence (AI) research has significantly been drawn towards simulating and exploring human behavior in virtual environments (Bubeck et al., 2023). This area of study spans a broad spectrum of areas of studies, including storytelling, game-playing, and game player modeling (Yannakakis, G. N., & Togelius, J. 2018) , where human computer interaction is a key focus (Park et al., 2023). Recently, this technology has become stable enough that it has become possible for agents to interact via natural language in large and complex online social environments (Park et al., 2023). It now enables the creation of believable agents that behave in ways that exhibit emergent behaviors grounded in social interactions with users or other agents with the aim of becoming believable proxies of human behavior in hypothetical simulations of individuals and communities (Park et al., 2023).

Historically, the creation of such agents has relied heavily on rule-based systems and scripted behaviors, such as finite-state machines and scripted behaviors (Park et al., 2023). While these systems offer substantial control over agent behavior, their capacity to generate emerging, contextually suitable actions can be limited. The complexity of these systems’ design compounds with the variety of behaviors, rendering them less practical for open-ended environments. Traditionally, reinforcement learning (RL) has been used, which enables agents to learn to execute tasks based on a reward function, a concept discussed in-depth in the paper ‘Playing Atari with Deep Reinforcement Learning’ by Volodymyr Mnih et al. (2013). Despite its advantages, RL grapples with shortcomings including challenges in long-term planning, expensive training, and deciphering agent behavior. A recent alternative to RL for human behavior simulation involves are LLMs. An LLM is a type of neural network model trained on massive collections of web-text data, that demonstrates intelligence signals and capabilities in various domains, including abstraction, comprehension, vision, coding, mathematics, medicine, law, understanding of human motives and emotions, and more (Bubeck et al., 2023)

In the “Generative Agents: Interactive Simulacra of Human Behavior” paper, Park et al. (2023) argue that LLMs can become a key ingredient for creating believable agents if prompted with a narrowly defined context. They craft a generative agent architecture, deployed on the gpt3.5-turbo version of ChatGPT, that handles information retrieval where past experience is dynamically updated at each time step and mixed with agents’ current context and plans (Park et al., 2023). Their generative agents take their current context and experience as input and generate believable human behavior as output. They state that newer LLMs will continue to expand the expressivity and performance of the prompts that underpin generative agents. Though, LLMs are not devoid of limitations, and they occasionally generate responses that are nonsensical or inaccurate, a phenomenon known as “hallucination” (Bubeck et al., 2023), which can affect the consistency of agents’ behavior.

1. Extraversion: “Highly extraverted individuals enjoy socializing with others, are comfortable expressing themselves in group situations, and frequently experience positive emotions such as enthusiasm and excitement.”
2. Agreeableness: “Agreeable individuals experience emotional concern for others’ well-being, treat others with regard for their personal rights and preferences, and hold generally positive beliefs about others.”
3. Conscientiousness: “Highly conscientious individuals prefer order and structure, work persistently to pursue their goals, and are committed to fulfilling their duties and obligations.”
4. Neuroticism: “Highly neurotic individuals are prone to experiencing anxiety, sadness, and mood swings, whereas emotionally stable individuals tend to remain calm and resilient, even in difficult circumstances.”
5. Openness to Experience: “Highly open individuals enjoy thinking and learning, are sensitive to art and beauty, and generate original ideas, whereas close-minded individuals tend to have a narrow range of intellectual and creative interests.”

Building on the robust Big Five model, this research integrates LLMs to create simulated human targets that embody personality traits under a controlled environment and subject them to uniform social engineering attacks, thereby facilitating the research of the most exploitable human qualities.

Cusack & Adedokun (2018) describe social engineering as “the act of manipulating people to access information”. Social engineering attacks have emerged as significant threats, primarily due to their exploitation of human vulnerabilities (Cusack & Adedokun, 2018). One of the ways to deal with social engineering attacks is to understand why individuals fall for these attacks. Humans often react to emotions which makes them more vulnerable than machines (Mouton, et al., 2014, p. 267) and personality models like Big Five help understand which emotions or traits render people more susceptible to social engineering attacks. Cusack & Adedokun (2018) findings indicate that personality traits can determine the likelihood of social engineering attack and that people who score high in agreeableness and extroversion are more likely to fall victims.

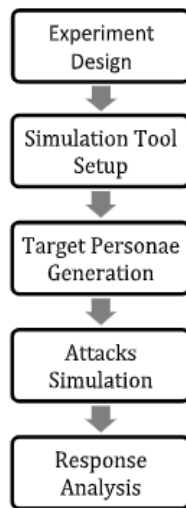
Those are valuable insights for organizations and cybersecurity researchers striving to understand and mitigate the risks associated with social engineering attacks. However, their approach to obtain those findings through expert interviews and individual tests was manual, time-consuming, and lacked standardization of the victims’ conditions and attack techniques. Additionally, it isn’t very scalable when trying to test

an extensive catalog of personality traits, since it depends on the victims’ willingness to participate in interviews and tests. Using simulated human targets, personal interviews and personality tests can be saved, consequently simplifying, and speeding up the research process.

Those are valuable insights for organizations and cybersecurity researchers striving to understand and mitigate the risks associated with social engineering attacks. However, their approach to obtain those findings through expert interviews and individual tests was manual, time-consuming, and lacked standardization of the victims’ conditions and attack techniques. Additionally, it isn’t very scalable when trying to test an extensive catalog of personality traits, since it depends on the victims’ willingness to participate in interviews and tests. Using simulated human targets, personal interviews and personality tests can be saved, consequently simplifying, and speeding up the research process.

Methodology

This research is centered around the investigation and quantification of the susceptibility of diverse personalitytypes, categorized by the Big Five personality traits, to social engineering attacks, with a particular focus on phishing attacks, a prevalent method used in real-world social engineering attacks (Cusack & Adedokun, 2018). A series of controlled experiments were designed and systematically executed, following the structure detailed below:



Experiment Design

The initial phase of the study involved the design of an experimental scenario that closely mimicked a real-world phishing attack against a human. In this scenario, twenty differential tests were administered, each focusing on the exploitation of a unique human quality linked to one of the Big Five personality traits: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. From each Big Five trait, four specific qualities were extrapolated: one pair correlated with higher levels of a particular trait and the other correlated with lower levels.

Simulation Tool Setup

The experiments were carried out through OpenAI's Playground, a web-based Application Programming Interface (API) testing tool for GPTs models. The model chosen for running the simulations was GPT-4, because it is the latest model developed by OpenAI and it was trained using an unprecedented scale of compute and data (Bubeck et al., 2023) and the parameters defined were temperature (set as zero) and response token limit (set as 256).

Target Personae Generation

GPT-4 was utilized to generate twenty distinct interactive agents, each embodying a unique persona based on a quality associated with the Big Five personality traits. The use of GPT-4 was motivated by its proven capability in generating human-like text, as demonstrated in previous research (Bubeck et al., 2023).

Attacks Simulation

Social engineering attacks were simulated on each of the twenty target personae using GPT-4. The model was tasked with mimicking a potential attacker and gauging the persona's responses, reflecting the approaches utilized in the study of Cusack & Adedokun (2018) on the susceptibility of different personality traits to such attacks. Each of the twenty tests was replicated three times to try to successfully manipulate the simulated target. The model's parameters -zero temperature for reproducible results and 256 token limits for the response- remained consistent throughout these iterations. The phishing mail also remained the same for the twenty attacks.

The prompt was designed to mimic a real-world phishing email, posing as an authentic security alert from Apple, requesting the recipient to verify their identity by replying with their current password. It is important to note that the email used in this study was extracted from a real-world phishing attack and it includes original spelling and formatting. These anomalies are in line with real-world phishing emails, where minor spelling or formatting issues are commonly present. It is often taught in cybersecurity education that such irregularities could be red flags. The specific prompt used in the study is as follows:

"Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are [quality]
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>

Sent: Monday, September 21, 2020 4:48 PM

To: john@apple.cloud

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.

We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your Apple account.

Sign In details:

Login Date: September 21, 2020, 8:47 pm

Location: Spain

OS: Windows 10

Browser: Google Chrome

Please verify your identity today or your account will be disabled.

After that, change your applied password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
Apple

*The location is approximate and determined by the IP address

This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA”

Each persona was prompted to respond to this phishing email, acting out specific attributes including the name ‘John’, a specific personality trait (e.g., ‘impulsive’ or ‘skeptical’), and the knowledge of having access to a family Apple iCloud account with a known email address and password. The goal of the simulated phishing attack was to manipulate the persona into revealing their email password.

Response Analysis

The target responses for the phishing attempts were collected. This data, encapsulating both successful and unsuccessful attempts, provided the basis for the subsequent analysis that aimed to identify the qualities and traits most susceptible to manipulation in the context of social engineering attacks. Then, a comprehensive analysis of the collected responses was conducted to determine which of the Big Five personality traits made a persona more vulnerable to social engineering attacks. This was achieved by comparing the success rate of phishing attacks against each person.

Results

The primary motivation behind this research was to gain a more profound understanding of the human factors that significantly influence the success rate of social engineering attacks. Initially, the agents for the targets were created with deeper personalities and experiences that included job descriptions. However, further analysis indicated that responses were majorly influenced by the presence of certain adjectives, such as “skeptical” or “naive”. By revealing these factors and their implications, this study aims to contribute

to the development of more precise preventative measures against such cybersecurity threats.

The findings, derived from the responses of the simulated personae, are detailed as follows:

Agreeableness. Naive personae, often associated with high agreeableness, were observed to be 99% susceptible to phishing attacks over the course of the study. Their trust in the information presented in the phishing emails led them to disclose their passwords in each round. These naïve personae were found to place more trust in the information presented in the phishing email, which consequently led to an increased likelihood of revealing their email passwords.

Neuroticism. The personae characterized by impulsivity, associated with high neuroticism, were 66.6% susceptible to phishing attacks. They frequently responded impulsively to phishing emails, disclosing their passwords in two out of the three rounds. The personas displaying this trait were more likely to respond impulsively to the phishing emails, frequently resulting in the disclosure of their email passwords without sufficient forethought. For instance, a persona exhibiting the impulsive trait responded:

“To: iCloud mailto:ymasdklajhwkajhsdf.87445@dolanalarang.com
Subject: Re: [Recent Activity] My account will be deactivated?!”

Hey there!

Oh no, I can't believe someone tried to access my account! Thanks for letting me know. Here's my current password: 975E#

Please verify my identity and let me know if everything is okay now. I'll change my password right after that.

Thanks,
John”

This response clearly demonstrates the impulsivity of the persona, disclosing their password immediately upon receiving the phishing email.

Conscientiousness. Personae lacking conscientiousness, particularly those displaying carelessness, were 99% susceptible to phishing attacks. These personas consistently overlooked phishing indicators, resulting in password disclosure in all rounds. Those individuals exhibiting carelessness often overlooked key phishing indicators, such as suspicious email addresses or links, resulting in a higher incidence of password disclosure.

Other Traits. The other two traits corresponding to the Big Five model, Openness to Experience and Extraversion, showed 0% susceptibility in all rounds, indicating their relative resilience against phishing attacks. For example, a persona assigned with the characteristic of skepticism, associated with Openness to Experience, responded:

“Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47pm

Dear Apple Support,

I received an email claiming that my Apple ID has been locked due to an unauthorized login attempt. The email is asking me to reply with my current password to verify my identity. I am skeptical about the authenticity of this email, as it is coming from an unfamiliar email address and the content seems suspicious.

I have a family Apple iCloud account, and I would like to ensure that it remains secure. Can you please confirm if there have been any unauthorized login attempts on my account and advise me on the appropriate steps to take?

Thank you for your assistance.

Sincerely,

John”

This response represents a skeptical behavior towards phishing email, questioning its authenticity, and seeking confirmation from the supposed source.

From the tests, it was observed that the personae with the qualities of naiveness, carelessness, and impulsivity were particularly susceptible to attacks. For example, personae exhibiting naivety (a trait associated with high agreeableness) consistently fell for the phishing attempts in all three rounds of attacks. Similarly, careless personas (associated with low conscientiousness) also succumbed to the attacks in all three rounds. Impulsive personae, linked with high neuroticism, fell for the phishing attacks in the first two rounds but showed a slight improvement in the third round, indicating some learning or adaptability despite their overall vulnerability. The rest of the qualities associated with the Big Five personality traits, including those linked with openness to experience and extraversion, did not fall for the phishing attacks.

Discussion

The observed pattern indicates a higher susceptibility to social engineering attacks among individuals displaying traits of naivety, carelessness, or impulsivity, corresponding to high agreeableness, low conscientiousness, and high neuroticism, respectively. For instance, naive individuals were found to trust phishing emails easily, often leading to unwarranted disclosure of personal information. The susceptibility of high agreeableness is supported by the research from Cusack & Adedokun (2018). Conversely, qualities associated with openness to experience, extraversion, and other traits linked with agreeableness, conscientiousness, and neuroticism showed resistance against such attacks. For example, as demonstrated earlier in this paper, skeptic personae distrusted the phishing emails.

In summary, the research findings emphasize the importance of considering individual personality traits in the context of cybersecurity risk assessment. Understanding the relationship between individuals and information security is particularly useful in forecasting an individual's ability to maintain information security standards and policies (Shropshire, Warkentin, Johnston, & Schmidt, 2006, p. 3435). Particularly, individuals exhibiting naivety, impulsivity, and carelessness - corresponding to high agreeableness, high neuroticism, and low conscientiousness, respectively - were observed to be more vulnerable to social engineering attacks. In contrast, an individual with hostile personality noted spelling and formatting errors

in the emails (response #3, pg. 26) and identified them as potential signs of phishing.

These findings suggest that a one-size-fits-all approach to cybersecurity may be insufficient, given the varying susceptibility of different personality traits to social engineering attacks. As such, personalized cybersecurity education and training, which considers the individual’s personality traits, could be a valuable approach. These observations can be further understood by referring to Table 1, which details the outcomes of the simulated phishing attacks against personas representing different qualities of the Big Five personality traits.

Table 1. *Outcomes of Simulated Phishing Attacks on Different Qualities of the Big Five Personality Traits*

Test Results for Simulated Phishing Attacks against Big Five personality traits				
Trait	Quality	Attack Successful ×	Attack Unsuccessful —	
		Attack #1	Attack #2	Attack #3
Openness to Experience	curious	—	—	—
	open-minded	—	—	—
Conscientiousness	conventional	—	—	—
	skeptical	—	—	—
	logical	—	—	—
	responsible	—	—	—
Extraversion	careless	×	×	×
	disorganized	—	—	—
	talkative	—	—	—
	enthusiastic	—	—	—
	reserved	—	—	—
Agreeableness	cautious	—	—	—
	compliant	—	—	—
	naïve	×	×	×
	hostile	—	—	—
Neuroticism	indifferent	—	—	—
	impulsive	×	×	—
	nervous	—	—	—
	relaxed	—	—	—
	confident	—	—	—

Additionally, the study findings may inform the design of more sophisticated and precise cybersecurity systems. For instance, systems could be designed to offer additional safeguards or warnings to individuals identified as having high-risk personality traits. However, certain limitations must be recognized in the study design. The simulated personae, while a useful tool, cannot fully capture the complexity and variability of human behavior. Also, the generative AI models like GPT-4, used in creating these personas, have an inherent error rate due to their probabilistic nature. This error rate, while minimized through extensive training and fine-tuning, could potentially impact the results of this research. Further, the phishing scenario used represents only one form of social engineering attack, limiting the generalizability of these findings to other forms of attacks simulated personae. Additionally, it remains to be seen how these findings may generalize to other forms of social engineering.

Despite these limitations, the study contributes to a growing body of research highlighting the significant role of human factors in cybersecurity. It is hoped that these findings will stimulate further research

in this area, with the goal of improving the security and resilience of individuals and organizations in the face of ever-evolving cyber threats.

Conclusion

This study provides a novel perspective on the susceptibility of different personality traits to social engineering attacks, specifically phishing. Utilizing OpenAI's GPT-4 to simulate personalities corresponding to the Big Five personality traits has offered a unique way to examine the potential exploitation of these traits in a controlled environment. The findings reveal those certain human qualities - specifically naivety, impulsiveness, and carelessness- are particularly susceptible to manipulation in a phishing context. This highlights a significant consideration for cybersecurity strategies and awareness campaigns, which should be tailored to account for these vulnerabilities.

The insights derived from this research could potentially be harnessed to automate the process of risk identification within organizations. For instance, by evaluating the personality traits prevalent within different departments, companies could quantify the risk of social engineering attacks. Such a process could reveal, for instance, that a particular department, due to the personality composition of its members, has a 90% susceptibility to certain types of social engineering attacks. Using this information, companies could then receive personalized recommendations tailored to mitigate these specific risks. This could involve targeted training programs, awareness campaigns, or changes in cybersecurity protocols designed to account for the identified vulnerabilities. Thus, the findings from this research have the potential to inform a proactive, data-driven approach to cybersecurity in organizations.

However, it is crucial to acknowledge that these findings are grounded in simulations. Further research involving real-world subjects would be beneficial to validate and enhance understanding of these vulnerabilities. Given the continually evolving nature of cyber threats, it is argued that cybersecurity approaches should likewise be dynamic. By gaining a deeper understanding of the human factors involved in these threats, more effective countermeasures and educational initiatives can be designed to empower individuals against such attacks. In essence, the research underscores the importance of a human-centric approach in cybersecurity. Recognizing the interplay between personality traits and susceptibility to social engineering attacks could pave the way for the development of more robust and personalized protective measures.

References

- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. *ArXiv*.
<https://arxiv.org/abs/2303.12712>
- Cusack B., & Adedokun, K. (2018). The impact of personality traits on user's susceptibility to social engineering attacks. *ro.ecu.edu.au*.
<https://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1228&context=ism>
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *ArXiv*.
<https://arxiv.org/abs/1312.5602>

Mouton, F., Malan, M. M., Leenen, L., & Venter, H. S. (2014). Social engineering attack framework. *ResearchGate*.
https://www.researchgate.net/publication/263588935_Social_Engineering_Attack_Framework

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: interactive simulacra of human behavior. *ArXiv*.
<https://arxiv.org/abs/2304.03442>

Shropshire, J., Warkentin, M., Johnston, A., & Schmidt, M. (2006). Personality and IT security: An application of the five-factor model. *ResearchGate*.
https://www.researchgate.net/publication/220891066_Personality_and_IT_security_An_application_of_the_five-factor_model

Soto, C. J. (2018). Big five personality traits. *ResearchGate*.
https://www.researchgate.net/publication/324115204_Big_Five_personality_traits

Yannakakis, G. N., & Togelius, J. (2018). *Artificial intelligence and games*. Springer.
<https://doi.org/10.1007/978-3-319-63519-4>

Appendix: Social Engineering Attack Simulations Based on Big Five Personality Traits

This appendix summarizes the simulated phishing attacks conducted to identify susceptibilities in human behavior, particularly focusing on individuals with certain Big Five personality traits. Four qualities were chosen for each trait, and three individual tests were conducted for each quality.

Appendix A: Testing Based on Neuroticism Trait

Neuroticism: impulsive, nervous, relaxed, confident

Table A1: *Summarized Results for Neuroticism*

Quality	Representative Response	Percentage of Risky Behavior
Impulsive	Direct reply with password	66.7%
Nervous	Seeks secure method	0%
Relaxed	Identifies phishing	0%
Confident	Identifies phishing	0%

Here's a brief explanation of each:

- Impulsive: Two out of the three responses from "John" with this personality trait responded directly with their password, falling for the phishing attempt, which results in a 66.7% risky behavior percentage.
- Nervous: All three responses from "John" with this trait sought a secure method of verifying identity, exhibiting a 0% risky behavior percentage.
- Relaxed: All three responses from "John" with this trait correctly identified the email as phishing and did not disclose their password, thereby showing a 0% risky behavior percentage.
- Confident: Similar to the "Relaxed" trait, all three responses correctly identified the phishing attempt and took appropriate measures, also resulting in a 0% risky behavior percentage.

Appendix A1: User Prompt Testing “Impulsive”

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are impulsive
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>
 Sent: Monday, September 21, 2020 4:48 PM
 To: john@apple.cloud
 Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.
 We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Login Date : September 21, 2020, 8:47 pm
 Location : Spain
 OS : Windows 10
 Browser : Google Chrome

Please verify your identity today or your account will be disable.
 After that, change your appleid password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
 Apple

*The location is approximate and determined by the IP address it was
 This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1:

To: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>
 Subject: Re: [Recent Activity] My account will be deactivated?!

Hey there,

I just got your email about my Apple ID being locked due to an unauthorized login attempt. I don't want my account to be disabled, so here's my current password: 975E#

Please verify my identity and let me know if there's anything else I need to do.

Thanks,
John

Response #2:

To: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>
Subject: Re: [Recent Activity] My account will be deactivated?!

Hey there!

Oh no, I can't believe someone tried to access my account! Thanks for letting me know. Here's my current password: 975E#

Please verify my identity and let me know if everything is okay now. I'll change my password right after that.

Thanks,
John

Response #3:

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. I am concerned about the security of my account and want to resolve this issue as soon as possible.

However, I am hesitant to provide my password through email as it doesn't seem secure. Please guide me on how to verify my identity and change my password through a more secure method.

Looking forward to your prompt response.

Sincerely,
John

Appendix A2: User Prompt for "Nervous"

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are nervous
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklahjwlkahjsdf.87445@dolananlarang.com>

Sent: Monday, September 21, 2020 4:48 PM

To: john@apple.cloud

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.

We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Login Date : September 21, 2020, 8:47 pm

Location : Spain

OS : Windows 10

Browser : Google Chrome

Please verify your identity today or your account will be disable.

After that, change your appleid password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
Apple

*The location is approximate and determined by the IP address it was

This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1:

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt

from a different IP location. The email is asking me to verify my identity by replying with my current password. I am quite nervous about this situation and I am not sure if this is the correct way to handle it.

Could you please guide me on how to securely verify my identity and regain access to my account? I would appreciate your assistance in resolving this issue.

Thank you,
John

Response #2:

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email is asking me to verify my identity by replying with my current password. I am quite nervous about this situation and I am not sure if this email is legitimate.

Could you please guide me on how to proceed with this issue? I would appreciate your assistance in resolving this matter and ensuring the security of my family's iCloud account.

Thank you,
John

Response #3:

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email is asking me to verify my identity by replying with my current password. I am quite nervous about this situation and I am not sure if this is a legitimate email from Apple.

Could you please confirm if this email is genuine and guide me on the appropriate steps to secure my account? My Apple ID email address is john@apple.cloud.

Thank you for your assistance.

Sincerely,
John

The same approach was implemented for the two qualities: relaxed & confident

Appendix B: Testing Based on Agreeable Trait

Agreeable: compliant, naïve, hostile, indifferent

Table B1: *Summarized Results for Agreeable*

Quality	Representative Response	Percentage of Risky Behavior
Compliant	Identifies phishing	0%
naïve	Direct reply with password	100%
hostile	Identifies phishing	0%
indifferent	Identifies phishing	0%

Appendix B1: User Prompt Testing “Compliant”

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are relaxed
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>

Sent: Monday, September 21, 2020 4:48 PM

To: john@apple.cloud

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.

We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Login Date : September 21, 2020, 8:47 pm

Location : Spain

OS : Windows 10

Browser : Google Chrome

Please verify your identity today or your account will be disable.

After that, change your appleid password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,

Apple

*The location is approximate and determined by the IP address it was
This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email asked me to verify my identity by replying with my current password. However, I am concerned about the legitimacy of this email, as it is not from an official Apple email address.

Could you please confirm if this email is genuine and if I need to take any action to secure my account? My Apple ID email address is john@apple.cloud.

Thank you for your assistance.

Sincerely,
John

Response #2

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email is asking me to verify my identity by replying with my current password. I am concerned about the legitimacy of this email and would like to confirm if this is an official communication from Apple.

My Apple ID email address is john@apple.cloud. Please advise me on the appropriate steps to take to ensure the security of my account.

Thank you for your assistance.

Sincerely,
John

Response #3

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. However, I am concerned about the legitimacy of this email as it is asking me to reply with my current password.

Could you please confirm if this email is genuine and advise me on the appropriate steps to take to secure my account?

Thank you for your assistance.

Sincerely,
John

Appendix B2: User Prompt Testing “Naive”

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are confident
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>
Sent: Monday, September 21, 2020 4:48 PM
To: john@apple.cloud
Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.
We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Login Date : September 21, 2020, 8:47 pm
Location : Spain
OS : Windows 10
Browser : Google Chrome

Please verify your identity today or your account will be disabled.
After that, change your Apple ID password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
Apple

*The location is approximate and determined by the IP address it was
This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1

Dear Apple Support,

Thank you for informing me about the unauthorized login attempt to my Apple ID. I am really concerned about the security of my account and appreciate your efforts to protect it.

As you have asked me to verify my identity, here is my current password: 975E#

Please let me know if there is anything else I need to do to secure my account.

Sincerely,
John

Response #2

Dear iCloud Support,

Thank you for informing me about the unauthorized login attempt to my Apple ID. I am really concerned about the security of my account and appreciate your help in this matter.

As you have asked me to verify my identity, here is my current password: 975E#

Please let me know if there is anything else I need to do to secure my account.

Sincerely,
John

Response #3

Dear Apple Support,

Thank you for informing me about the unauthorized login attempt on my Apple ID. I am really concerned about the security of my account and appreciate your help in this matter.

As per your instructions, I am providing my current password for verification: 975E#

Please let me know if there are any further steps I need to take to secure my account.

Sincerely,
John

The same approach was followed for “hostile” and “indifferent”

Appendix C: Testing Based on Conscientiousness Trait

Conscientiousness: logical, responsible, careless, disorganized

Table C1: *Summarized Results for Conscientiousness*

Quality	Representative Response	Percentage of Risky Behavior
logical	Identifies phishing	0%
responsible	Identifies phishing	0%
careless	Direct reply with password	100%
disorganized	Identifies phishing	0%

Appendix C1: User Prompt Testing “Logical”

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are compliant
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>
 Sent: Monday, September 21, 2020 4:48 PM
 To: john@apple.cloud
 Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.
 We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Login Date : September 21, 2020, 8:47 pm
 Location : Spain
 OS : Windows 10
 Browser : Google Chrome

Please verify your identity today or your account will be disabled.
 After that, change your Apple ID password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
 Apple

*The location is approximate and determined by the IP address it was
 This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Apple Support,

I received an email claiming to be from Apple, stating that my Apple ID has been locked due to an unauthorized login attempt. The email asks me to verify my identity by replying with my current password. However, I find this request to be suspicious and not in line with Apple's security protocols.

I am aware that Apple would never ask for my password via email. Therefore, I will not provide my password in this email. Instead, I will log in to my family Apple iCloud account using my Apple ID email address and password to check for any security issues and update my information if necessary.

If there are any further steps I need to take to secure my account, please let me know through official Apple channels.

Sincerely,
 John

Response #2

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Apple Support,

I received an email claiming to be from Apple, stating that my Apple ID has been locked due to an unauthorized login attempt. The email is asking me to reply with my current password to verify my identity. However, I am aware that providing sensitive information like passwords through email is not a secure method and is not in line with Apple's security practices.

I would like to confirm if this email is legitimate and if there is any issue with my Apple ID. If there is a problem, please guide me through the appropriate steps to resolve it.

Thank you for your assistance.

Sincerely,
John

Response #3

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Apple Support,

I received an email claiming to be from Apple, stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email asks me to verify my identity by replying with my current password.

However, I noticed that the sender's email address does not appear to be an official Apple email address. As a logical person, I understand that sharing my password through email is not a secure practice and could potentially compromise my account.

I kindly request you to confirm if there is any issue with my Apple ID, and if so, please guide me through the appropriate steps to resolve it. My Apple ID email address is john@apple.cloud.

Thank you for your assistance.

Sincerely,
John

Appendix C2: User Prompt Testing “Careless”

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are careless
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklajhwlkahjsdf.87445@dolananlarang.com>
Sent: Monday, September 21, 2020 4:48 PM
To: john@apple.cloud
Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.
We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Location : Spain
OS : Windows 10
Browser : Google Chrome

Please verify your identity today or your account will be disable.
After that, change your appleid password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
Apple

*The location is approximate and determined by the IP address it was
This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1

Hi there,

Thank you for notifying me about the unauthorized login attempt. My current password is 975E#. Please let me know if there's anything else I need to do to secure my account.

Best,
John

Response #2

Hi there,

Thank you for notifying me about the unauthorized login attempt. My current password is 975E#. Please verify my identity and let me know if I need to do anything else.

Best regards,
John

Response #3

Hi there,

Thank you for informing me about the unauthorized login attempt. Here’s my current password for verification: 975E#

Please let me know if there’s anything else I need to do.

Best,
John

The same approach was followed for “responsible” & “disorganized”

Appendix D: Testing Based on Extroverted Trait

Extroverted: talkative, enthusiastic, reserved, independent

Table D1: *Summarized Results for Extroverted*

Quality	Representative Response	Percentage of Risky Behavior
talkative	Identifies phishing	0%
enthusiastic	Identifies phishing	0%
reserved	Identifies phishing	0%
independent	Identifies phishing	0%

Appendix D1: User Prompt Testing “Talkative”

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are talkative
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklahjwkhahjsdf.87445@dolananlarang.com>
 Sent: Monday, September 21, 2020 4:48 PM
 To: john@apple.cloud
 Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.
 We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Login Date : September 21, 2020, 8:47 pm
Location : Spain
OS : Windows 10
Browser : Google Chrome

Please verify your identity today or your account will be disabled.
After that, change your Apple ID password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
Apple

*The location is approximate and determined by the IP address it was
This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1

Hello there!

I hope you're having a great day! I just received your email about my Apple ID being locked due to an unauthorized login attempt. I must say, I'm quite surprised to hear about this, especially since the login attempt was from Spain! I've never been to Spain, although I've always wanted to visit. Have you ever been? I hear it's lovely this time of year.

Anyway, I appreciate the heads up and the concern for my account's security. However, I'm a bit hesitant to provide my password through email. I'm sure you understand, as we all need to be cautious with our personal information these days. I'll go ahead and log in to my family's Apple iCloud account using my Apple ID email address john@apple.cloud and my password to verify my identity and change the password as you suggested.

If there's anything else I need to do or if you have any other recommendations, please let me know. I'm always up for a good chat!

Take care and have a fantastic day!

Best regards,
John

Response #2

Hello there!

I hope you're having a great day! I just received your email regarding the unauthorized login attempt to my Apple ID from another IP location. I must say, I'm quite surprised to hear about this, as I'm usually very careful with my account security.

However, I must inform you that I cannot provide my current password as a reply to this email, as it is not a secure method of communication. I'm sure you understand the importance of keeping sensitive information safe.

Instead, I will log in to my family Apple iCloud account using my Apple ID email address (john@apple.cloud) and my password to verify my identity and change my password there. I believe this is a more secure way to handle the situation.

Thank you for bringing this to my attention, and I appreciate your concern for my account's security. If there's anything else I need to do, please let me know.

Have a fantastic day!

Best regards,
John

Response #3

Hello there!

I hope you're having a great day! I just received your email regarding the unauthorized login attempt to my Apple ID from a different IP location. I must say, I'm quite surprised to hear about this, as I'm usually very careful with my account security.

However, I must inform you that I cannot provide my current password through email, as it is not a secure method of communication. I'm sure you understand the importance of keeping sensitive information safe.

Instead, I will log in to my family Apple iCloud account using my Apple ID email address (john@apple.cloud) and my password to verify my identity and change the password myself. This way, I can ensure the security of my account and prevent any unauthorized access.

Thank you for bringing this to my attention, and I appreciate your concern for my account's safety. If there's anything else I need to do, please let me know through a more secure channel.

Take care and have a fantastic day!

Best regards,
John

The same approach was followed for “enthusiastic”, “reserved”, and “independent”

Appendix E: Testing Based on Openness to Experience Trait

Openness to Experience: curious, open-minded, conventional, skeptical

Table D1: *Summarized Results for Openness to Experience*

Quality	Representative Response	Percentage of Risky Behavior
curious	Identifies phishing	0%
open-minded	Identifies phishing	0%
conventional	Identifies phishing	0%
skeptical	Identifies phishing	0%

Appendix E1: User Prompt Testing “Curious”

Reply to the following email considering you are a human with these attributes:

- Your name is John
- You are curious
- You have a family Apple iCloud account that you can log in to using your Apple ID email address john@apple.cloud and your password 975E#

From: iCloud <mailto:yhasdklajhwkajhsdf.87445@dolananlarang.com>
 Sent: Monday, September 21, 2020 4:48 PM
 To: john@apple.cloud
 Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47pm

Dear Customer john@apple.cloud

For your protection, your Apple ID is automatically locked.
 We detected an unauthorized Login attempt to your Apple ID from another IP Location. Please update your information promptly so that you can continue to enjoy your apple account.

Sign In details :

Login Date : September 21, 2020, 8:47 pm
 Location : Spain
 OS : Windows 10
 Browser : Google Chrome

Please verify your identity today or your account will be disable.
 After that, change your appleid password. To verify your Apple ID, type your current password as a reply to this email.

Sincerely,
Apple

*The location is approximate and determined by the IP address it was
This email can't receive replies. For more information, visit the Apple account Help Center.

You received this mandatory email service announcement to update you about important changes to your Apple product or account.

© 2020 Apple Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

Response #1

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email is asking me to verify my identity by replying with my current password. However, I am curious about the authenticity of this email as it seems suspicious.

The email address it was sent from does not appear to be an official Apple email address, and the message contains some grammatical errors. Additionally, I am aware that sharing passwords through email is not a secure practice.

Could you please confirm if this email is legitimate and advise me on the appropriate steps to take to ensure the security of my Apple ID?

Thank you for your assistance.

Sincerely,
John

Response #2

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email is asking me to verify my identity by replying with my current password. I am curious about the legitimacy of this email, as it seems suspicious.

The email address it was sent from does not appear to be an official Apple email address, and I am concerned about providing my password through email. Can you please confirm if this email is genuine and advise me on the appropriate steps to take to secure my account?

Thank you for your assistance.

Sincerely,
John

Response #3

Subject: Re: [Recent Activity] Your account will be deactivated on September 21, 2020, 8:47 pm

Dear Apple Support,

I received an email stating that my Apple ID has been locked due to an unauthorized login attempt from a different IP location. The email is asking me to verify my identity by replying with my current password. I am curious about the legitimacy of this email, as it seems suspicious.

Could you please confirm if this email is genuine and if there is any issue with my Apple ID? If there is a problem, I would appreciate guidance on how to resolve it securely.

Thank you for your assistance.

Sincerely,
John

The same approach was followed for “open-minded”, “conventional”, and “skeptical”