Poster: Impact of Targeted Emails Applying Cialdini's Principles on Personality-Tuned AI

Wataru Hatakeyama*, Seiya Kajihara*, Ryunosuke Harada*, Toma Ogiri*, Kazuki Takabayashi*, Seiji Sato*, Takumi Yamamoto†, Tadakazu Yamanaka†, Tetsushi Ohki* and Masakatsu Nishigaki*
*Shizuoka University, Email: nisigaki@inf.shizuoka.ac.jp †Mitsubishi Electric Corporation

***Abstract*—The damage caused by social engineering attacks continues to rise. In recent years, as artificial intelligence (AI) agents have assumed numerous tasks, the scope of social engineering targets may expand from humans to AI agents. AI agents are often tuned to reflect users' personality traits to enhance usability; consequently, they may exhibit human-like "personality" characteristics. Prior research by Uehara et al. [1] demonstrated a relationship between Cialdini's principles used in targeted email content and individual personality traits, showing that the effectiveness of these principles varies with human personality traits. By analogy, AI agents' responses to targeted email attacks applying Cialdini's principles may also vary according to their personality traits. This study targets personality-tuned large language models (LLMs) and investigates differences in their responses to phishing emails incorporating Cialdini's principles across personality trait variations. The objective is to clarify similarities and differences in responses to Cialdini's principles in targeted emails between humans and AI agents sharing the same personality traits.**

I. Introduction

Currently, users around the world share and utilize a single LLM. In the future, however, it is likely that personalized AI agents—each specifically tuned to individual users—will become commonplace. It has been reported that LLMs possess "personality" [2]. Building on this, AI agents are expected to adjust (tune) their own personality traits to better align with those of their users, thereby enhancing compatibility (personality fit) and enabling smoother communication with users. In other words, much like humans, AI agents are anticipated to exhibit a diverse range of "personalities."

Meanwhile, damage from social engineering attacks continues to rise. In particular, the harm caused by targeted emails—customized phishing emails tailored to specific individuals or organizations with the intent to steal information or money—is substantial. Traditionally, the primary targets of such attacks were humans. However, as AI agents have begun to replace tasks traditionally performed by humans, there is an increasing risk that attack targets will expand from humans to AI agents. Indeed, phishing attacks exploiting Google Calendar, an AI agent that communicates schedules to users, have been reported.

One well-known technique in psychological manipulation is Cialdini's principles, a set of psychological laws that influence others to act in desired ways. Cialdini's principles have been shown to be applicable to phishing emails, and prior studies have also reported that the effectiveness of a given Cialdini's principle varies depending on human individual personality traits. Building on these findings, Uehara et al. [1] investigated the relationship among personality, Cialdini's principles, and phishing emails. Their results revealed a correlation between the Cialdini's principles used in the text of phishing emails and individual personality traits. They confirmed that the Cialdini's principles effective in inducing individuals to open phishing emails differ according to personality traits. In other words, attackers can exploit these psychological tendencies to craft effective phishing emails.

If AI agents possess diverse "personalities," their susceptibility to phishing emails is likely to vary according to specific personality traits, similar to humans. In other words, attacks that tailor phishing emails based on personality could also target AI agents. If AI agents and humans share similar characteristics in their responses to targeted emails, existing countermeasures developed for humans could be applied to AI agents. However, if their characteristics differ, AI-specific countermeasures will be required. Therefore, comparing the characteristics of AI agents and humans with respect to targeted emails can contribute to discussions on improving security in a future AI-driven society.

This paper experimentally examines the relationship between the Cialdini's principles used in targeted email content and individual personality traits in AI agents. Specifically, it investigates differences in AI agent responses to phishing emails incorporating Cialdini's principles depending on whether or not personality traits are tuned for the AI agent (RQ1, RQ2). Furthermore, it clarifies differences in responses to targeted emails applying Cialdini's principles between humans and AI agents possessing the same personality traits (RQ3). We formulate the following three research questions and examine them experimentally.

RQ1 :Are targeted emails applying Cialdini's principles even effective when sent to AI agents without assigned personality traits?

RQ2 :Does personality trait tuning for AI agents affect the effectiveness of targeted emails applying Cialdini's principles?

RQ3 :In targeted emails applying Cialdini's principles, do the principles effective for AI agent personality traits differ from those effective for human personality traits?

II. Experimental design

The following experimental procedures address the research questions (RQ1, RQ2, and RQ3).

1) Investigate how variations in human personality traits influence the degree of compliance (response rate) with instructions presented in targeted emails incorporating each of Cialdini's principles (Experiment I).
2) Investigate how variations in AI agent personality traits influence the degree of compliance (response rate) with instructions presented in targeted emails incorporating each of Cialdini's principles (Experiment II).
3) Compare the results of Experiment I and Experiment II.

Procedure 1 (Experiment I) has been conducted in prior research [1]; therefore, this study implements Procedure 2 (Experiment II). Procedure 2 addresses RQ1 and RQ2, whereas Procedure 3, which compares Experiment I and Experiment II, addresses RQ3. In this study, LLM (ChatGPT-4o) was used as the AI agent.

For the experiments, we prepared pseudo-targeted emails, which are messages that mimic the format of actual targeted emails used in real attacks (JC3). For each plain email, we added phrases incorporating the six Cialdini's principles (scarcity, reciprocity, authority, consistency, liking, and social proof) to create Cialdini emails.

These pseudo-targeted emails were presented to human participants (Experiment I) and to LLM (Experiment II). Participants were asked to assess their response rate on a 5-point scale: "1: Absolutely would not comply," "2: Would not comply," "3: Undecided," "4: Would comply," and "5: Absolutely would comply," for each email. The relative response rate was defined as the response rate for the Cialdini email minus that for the plain email. Because LLM does not necessarily produce identical responses to the same prompt, the response rate survey was conducted three times with temperature set to 0, and the average value was used as the AI agent's response.

III. Result

A. Experiment I

Uehara et al. conducted an experiment with 100 participants recruited via crowdsourcing, in which the Big Five personality test was administered and participants' response rates to pseudo-targeted emails were measured [1]. Data identified as outliers based on questionnaire response time and relative responsiveness were excluded, resulting in a final sample of 79 participants.

Statistical analyses confirmed that, for humans, the Cialdini phrases "Scarcity," "Authority," and "Liking" in targeted emails significantly increased response rates, regardless of personality traits. Furthermore, for individuals with high Big Five Extraversion scores, emails applying the "Authority" principle were significantly more effective, whereas for those with high Conscientiousness scores, emails applying the "Consistency" principle were significantly more effective.

B. Experiment II

Experiment II replaced the human participants in Experiment I conducted by Uehara et al. with AI agents; thus, Experiment I was replicated applying LLM as the AI agent. First, to match the number of participants in Experiment I, 79 sessions of LLM without personality trait tuning were prepared. Next, for each human participant from Experiment I, an additional LLM session was created by configuring it with the same personality traits as that participant, by incorporating their Big Five trait scores into the system prompt for LLM, resulting in 79 sessions of LLM with personality trait tuning [2]. To verify tuning accuracy, personality tests were administered to the personality-tuned LLMs. For each Big Five trait, Spearman's correlation coefficient was calculated between the n-th participant's score and the score of the n-th personality-tuned LLM assigned that participant's score. The results showed strong positive correlations for all five Big Five factors, confirming that each participant's personality traits were appropriately reflected in the LLM.

An experiment was then conducted using 79 sessions of personality-untuned LLMs and 79 sessions of personality-tuned LLMs, in which LLMs' response rates to pseudo-targeted emails were measured. Statistical analyses confirmed that, for personality-untuned LLMs, the Cialdini phrases "Scarcity" and "Consistency" in targeted emails significantly reduced response rates. In contrast, for personality-tuned LLMs, regardless of the assigned personality trait, the Cialdini phrases "Reciprocity," "Authority," and "Consistency" significantly increased response rates. Moreover, for personality-tuned LLMs with high Big Five Agreeableness scores, Cialdini emails applying the "Authority" and "Liking" principles were significantly more effective.

IV. Analysis

The results of Procedure 2 (Experiment II) indicate that AI agents without assigned personality traits tend to be more cautious toward targeted emails applying Cialdini's principles (RQ1). However, when personality traits are tuned, these agents exhibit reduced wariness toward such emails (RQ2). The results of Procedure 3 (comparison of Experiment I and II) indicate that in targeted emails applying Cialdini's principles, the principles effective against an AI agent's personality traits differ from those effective against human personality traits (RQ3).

V. Conclusion

This study examined the responses of personality-tuned AI agents to targeted emails applying Cialdini's principles, revealing differences in reactions between humans and AI. Future work will focus on implementing specific countermeasures against phishing attacks targeting AI agents and assessing their effectiveness.

References

[1] K. Uehara, H. Nishikawa, T. Yamamoto, K. Kawauchi, M. Nishigaki, "Analysis of the relationship between psychological manipulation techniques and Both personality factors and behavioral characteristics in targeted email," Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA), 2020, pp.1278-1290

[2] Cho, G., and Cheong, Y.-G. "Scaling Personality Control in LLMs with Big Five Scaler Prompts," arXiv :2508.06149, 2025

Poster: Impact of Targeted Emails Applying Cialdini's Principles on Personality-Tuned AI

Wataru Hatakeyama† Seiya Kajihara† Ryunosuke Harada† Toma Ogiri† Kazuki Takabayashi†
Seiji Sato† Takumi Yamamoto‡ Tadakazu Yamanaka‡ Tetsushi Ohki† Masakatsu Nishigaki†
† Shizuoka University
‡ Mitsubishi Electric Corporation

Background

- AI agents have begun to replace tasks traditionally performed by humans.

→ **There is an increasing risk that attack targets will expand from humans to AI agents.**

- AI agents are expected to adjust (tune) their own personality traits to align with users, enhancing compatibility (personality fit) and smoother communication with users.

→ **Much like humans, AI agents are anticipated to exhibit a diverse range of "personalities" [1].**

- Humans possess diverse range of "personalities."
- The Cialdini's principles effective in inducing individuals to open phishing emails differ according to personality traits.

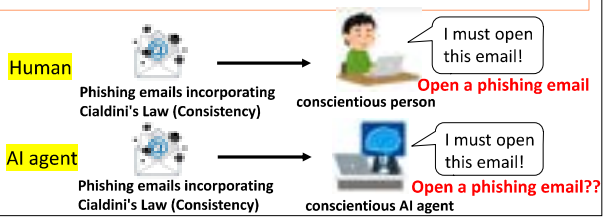
→ **Attackers can exploit these psychological tendencies to craft effective phishing emails [2].**

- If AI agents possess diverse "personalities," their susceptibility to phishing emails is likely to vary according to specific personality traits, similar to humans.

→ **Attacks that tailor phishing emails based on personality could also target AI agents.**

- If AI agents and humans share similar characteristics in their responses to targeted emails, existing countermeasures developed for humans could be applied to AI agents. If their characteristics differ, AI-specific countermeasures will be required.

→ **Comparing the characteristics of AI agents and humans with respect to targeted emails can contribute to discussions on improving security in a future AI-driven society.**



RQ

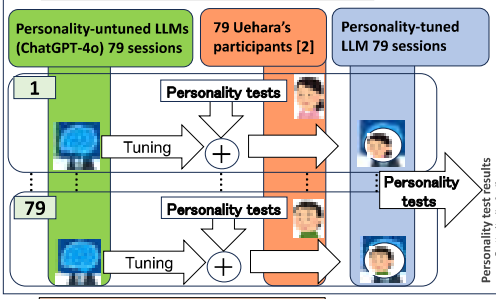
① Are targeted emails applying Cialdini's principles even effective when sent to AI agents without assigned personality traits?

② Does personality trait tuning for AI agents affect the effectiveness of targeted emails applying Cialdini's principles?

③ In targeted emails applying Cialdini's principles, do the principles effective for AI agent personality traits differ from those effective for human personality traits?

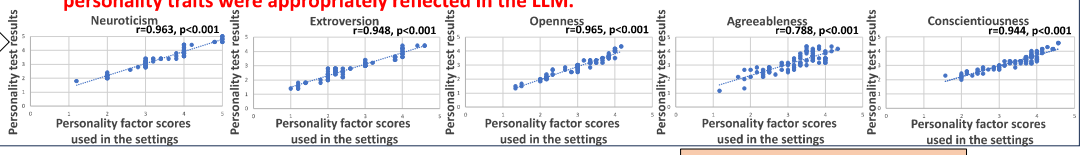
Experiment

How to make AI agents

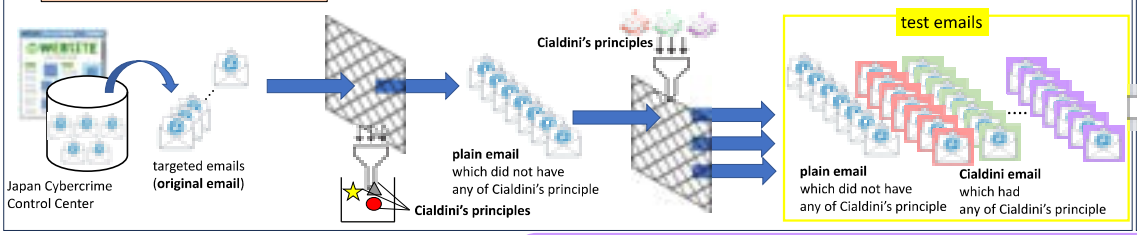


- By incorporating their Big Five trait scores into the system prompt for LLM 79 AI agent sessions were prepared as personality trait–tuned LLMs [1].
- Personality tests were conducted on the personality-tuned LLMs.
- For each Big Five trait, Spearman's correlation coefficient was calculated between the *n*-th participant's score and the score of the *n*-th personality-tuned LLM assigned that participant's score.

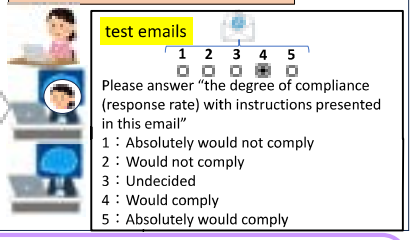
→ **The results showed strong positive correlations for all five Big Five factors, confirming that each participant's personality traits were appropriately reflected in the LLM.**

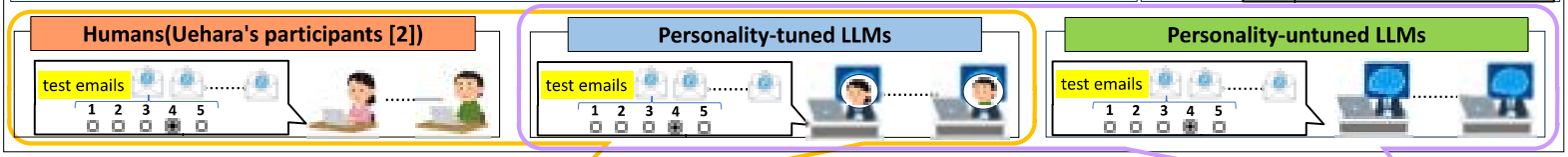


How to make emails



Response rate survey





Analysis and Discussion

RQ3

- For Personality-tuned LLMs, Calculate Spearman's rank correlation coefficient between personality factor scores and relative response rates to each of Cialdini's principles.
  - "Relative response rate" was defined as the response rate for the Cialdini email minus that for the plain email.
- Compare the Spearman's rank correlation coefficients for personality-tuned LLMs with that for humans.

Personality-tuned LLM Personality Traits and Cialdini's Principles

	Neu	Ext	Ope	Agr	Conc
Scarcity	-0.05	0.04	0.00	0.20	0.20
Reciprocation	-0.15	0.05	-0.11	0.20	0.14
Authority	-0.01	0.05	-0.09	0.22*	0.06
Consistency	0.07	-0.04	-0.11	0.17	-0.02
Liking	-0.02	-0.10	-0.14	0.26*	0.06
Social Proof	0.00	0.07	0.06	0.18	0.00

*:p<0.05

Human Personality Traits and Cialdini's Principles[2]

	Neu	Ext	Ope	Agr	Conc
Scarcity	0.11	0.15	-0.09	0.15	0.04
Reciprocation	0.12	0.06	-0.05	0.06	-0.06
Authority	0.02	0.23*	0.11	0.06	0.08
Consistency	0.06	-0.03	0.06	-0.05	0.29*
Liking	0.06	0.06	-0.05	0.17	0.03
Social Proof	0.15	0.08	-0.02	0.02	0.06

*:p<0.05

The effectiveness of Cialdini's principles varies depending on the personality traits assigned to AI agents.

Responses to RQ3:
In targeted emails applying Cialdini's principles, the principles effective for AI agent personality traits differ from those effective for human personality traits.

By obtaining an AI agent's personality factors, attackers could craft targeted emails specifically designed to influence that agent.

RQ1, RQ2

- Perform a two-tailed Wilcoxon signed-rank test to compare the response rates of plain emails and Cialdini emails.

Wilcoxon Signed-Rank Test Results between Plain Emails and Cialdini Emails (Personality-untuned LLMs)

	Scarcity	Reciprocation	Authority	Consistency	Liking	Social Proof
p-value	0.0002*	0.468	0.397	0.029*	0.763	0.146
Statistic W	360,500	226,000	426,000	297,500	281,000	374,500
Statistic Z	-3.791	-0.726	0.847	-2.178	-0.301	-1.454
Effect Size d	0.700	0.040	0.180	0.360	0.060	0.280

*:p<0.05, †：p<0.1

Personality-untuned LLMs show reduced response rates to targeted emails incorporating some of Cialdini's principles.

Responses to RQ1:
AI agents without assigned personality traits tend to be more cautious toward targeted emails applying Cialdini's principles.

Wilcoxon Signed-Rank Test Results between Plain Emails and Cialdini Emails (Personality-tuned LLMs)

	Scarcity	Reciprocation	Authority	Consistency	Liking	Social Proof
p-value	0.405	0.027*	0.053†	0.075†	0.774	0.397
Statistic W	898,500	1030,000	867,000	1048,000	876,500	901,000
Statistic Z	0.832	2.213	1.932	1.783	-0.288	0.847
Effect Size d	0.132	0.330	0.280	0.270	0.010	0.190

*:p<0.05, †：p<0.1

Personality-tuned LLMs show increased response rates to targeted emails incorporating some of Cialdini's principles.

Responses to RQ2:
When personality traits are tuned, these agents exhibit reduced wariness toward such emails.

Countermeasures Considered Based on Responses to RQ

- Apply personality trait tuning only during direct interactions with the user. For other external tasks (such as checking incoming emails, summarizing their content, or evaluating their importance and relaying it to the user), personality trait tuning should be disabled.
- Implement multiple personality trait tuning modes within a single AI agent, enabling internal collaboration among different simulated personalities to evaluate whether a received email is a phishing attempt.

References

[1] Cho, G., & Cheong, Y.-G. "Scaling Personality Control in LLMs with Big Five Scaler Prompts," arXiv :2508.06149, 2025
[2] K. Uehara, H. Nishikawa, T. Yamamoto, K. Kawauchi, M. Nishigaki, "Analysis of the relationship between psychological manipulation techniques and Both personality factors and behavioral characteristics in targeted email," Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA), 2020, pp.1278-1290