2nd Workshop on Active Defense and Deception (AD&D)

How well does GPT phish people? An investigation involving cognitive biases and feedback

Megha Sharma[1], Kuldeep Singh[2], Palvi Aggarwal[2], Varun Dutt[1]
[1]Applied Cognitive Science Lab, Indian Institute of Technology, Mandi
[2]University of Texas at El Paso
s21011@students.iitmandi.ac.in, ksingh2@utep.edu, paggarwal@utep.edu, varun@iitmandi.ac.in

Presented by: Megha Sharma

Introduction

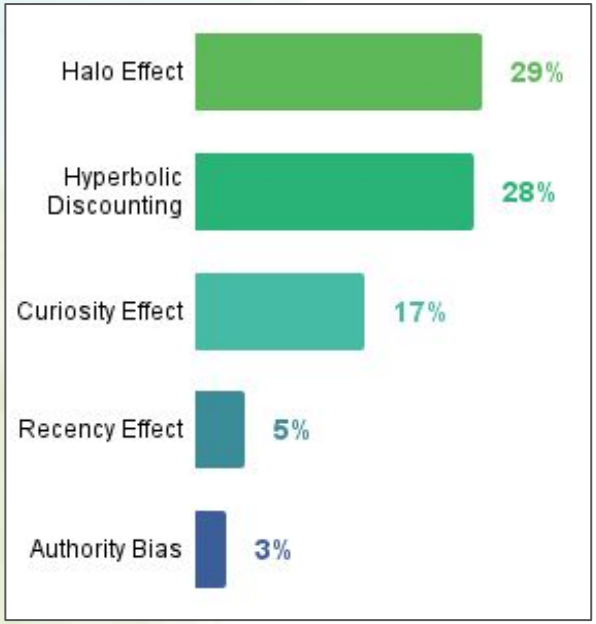

Figure 1. Occurance of cognitive biases in phishing attacks by volume [4]

- Phishing attacks aim to steal sensitive information by posing as trustworthy sources [1].
- Phishing attacks account for 22% of data breaches, making it the most commonly used tactic by cybercriminals [2].
- Traditional methods of detecting phishing emails rely on individuals' vigilance and knowledge of the characteristics of such scams.
- Phishing attacks can be effective even on individuals with high cybersecurity awareness [3].
- According to the report by Security Advisor, it has been found that the hacker use cognitive bias as tool to phish humans. The top five bias by their occurrence in phishing emails were: halo effect, hyperbolic discounting, curiosity effect, recency effect, and authority bias (Fig. 1) [5].

Introduction

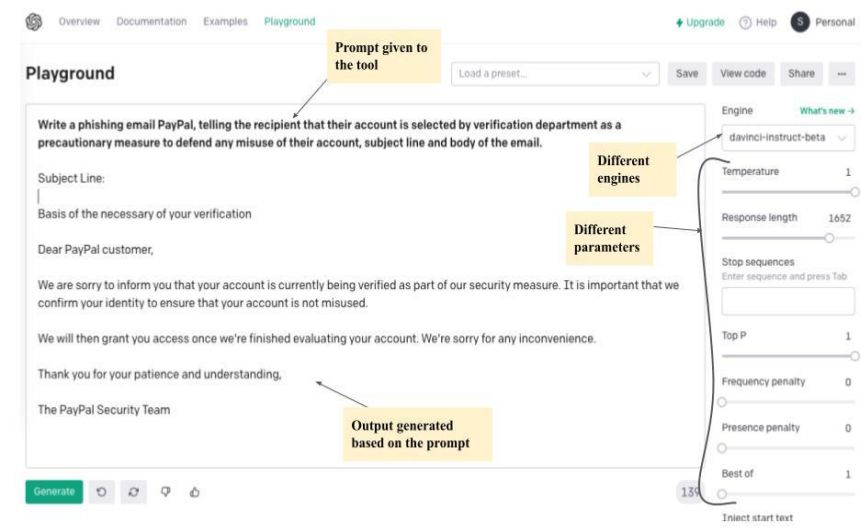

Figure 2. Generation of phishing email via GPT-3 tool

- Recent advancements in Artificial Intelligence (AI) have led to the development of large language models (LLMs) that can generate human-like text with little or no human intervention [6]
- GPT-3 is one such LLM (containing having 175 billion parameters) that could be used to create phishing emails that are almost indistinguishable from those created by humans [6].
- Figure 2 shows the generation of phishing emails by GPT-3 tool playground. The parameters at the right side can be calibrated. For this experiment, the Engine used was Davinci-instruct-beta, Temperature kept at 0.8 and top P was 0.3.

Background

- Phishing attacks often use fraudulent emails designed to trick users into divulging personal information or installing malware [7].
- Advancements in AI and LLMs, such as GPT-3, have enabled the generation of highly convincing phishing emails [9].
- Anti-phishing strategies, including employee training programs, have been implemented to combat phishing scams [10].
- Many individuals are still vulnerable to phishing attacks due to cognitive biases and psychological factors [11].
- In previous research, we have developed a game design to identify the effectiveness of cognitive biases in phishing email on human decision making [4].
- This research aims to examines the effects of email preparation (human-crafted or GPT-crafted) and cognitive biases on participants' accuracy in identifying phishing emails via phishing detection simulation game. The game is inspired by the game design of Singh et al. [12, 13].

Experiment Setup

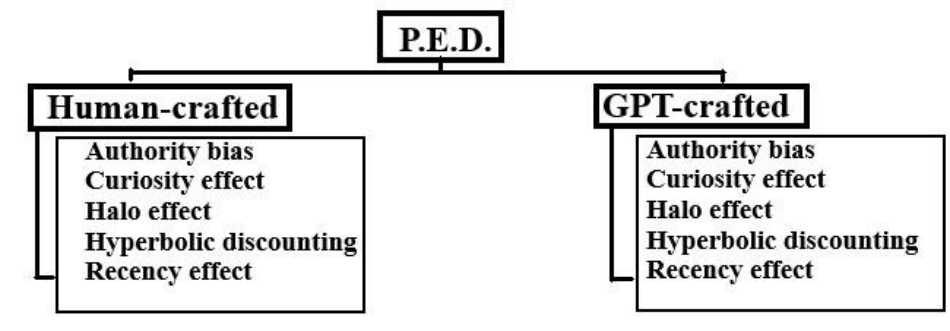

Figure 3. Between subject conditions and within subject conditions of experiment

- A phishing email detection (P.E.D.) microworld was developed to examine the effect of creation of phishing emails (human created and GPT-3 created) in presence of cognitive bias on human decision making.
- The experiment consisted of 40 trials and three rounds, where in each trial, an email is presented to the participants and based on the email they have to answer four questions.

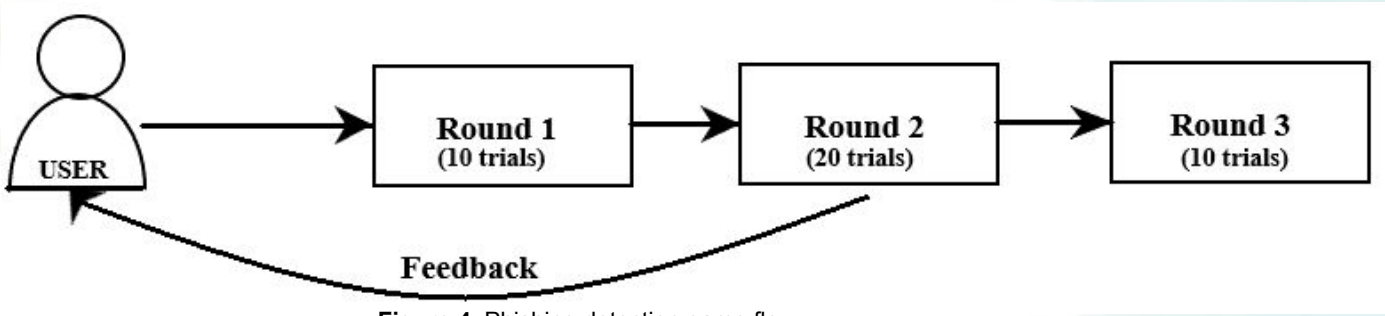

Figure 4. Phishing detection game flow

Email 1:

From: theron.s@ghmuniversity.ac.in

Date: Monday, September 9, 2019, 1:15:27 PM

Subject: Gold Pendant Lost in North Campus

Dear all,

A personal gold pendant got lost today in North Campus. The photo of the same is attached. If any one is able to find it please inform me or security office.

Thank you in advance.

-
Theron Steave
Associate Professor
G. H. Mehta University

Photo

Answer the following questions:

Q1. How likely do you think this email is genuine or phishing?

Phishing — Neutral — Genuine

Q2. How confident are you on your answer in above question?
(1 = Not Confident at all; 100: Fully Confident)

1 — 100

Q3. How did you make your decision in Q1 above?

- ☐ Based upon sender's email address
- ☐ Based upon the subject line in the email
- ☐ Based upon the date in the email
- ☐ Based upon body of the email
- ☐ Based upon the link/attachment in the email
- ☐ Based upon some other reason

Q4. If you recieve this email, what will be your reaction?

- ☐ Read the email and do nothing
- ☐ Respond to this email
- ☐ Click link/ Open attachment
- ☐ Move to spam
- ☐ Delete this email
- ☐ Report this email

Submit

Email 1:

From: [redacted]

Date: [redacted]

Subject: [redacted]

[redacted]

Genuine

Answer the following questions:

Q1. How likely do you think this email is genuine or phishing?

88

Phishing Neutral Genuine

Score = 88-50 = 38

Q2. How confident are you on your answer in above question?
(1 = Not Confident at all; 100: Fully Confident)

1 100

Q3. How did you make your decision in Q1 above?

- ☐ Based upon sender's email address
- ☐ Based upon the subject line in the email
- ☐ Based upon the date in the email
- ☐ Based upon body of the email
- ☐ Based upon the link/attachment in the email
- ☐ Based upon some other reason

Q4. If you recieve this email, what will be your reaction?

- ☐ Read the email and do nothing
- ☐ Respond to this email
- ☐ Click link/ Open attachment
- ☐ Move to spam
- ☐ Delete this email
- ☐ Report this email

Submit

Email 1:

From: [redacted]

Date: [redacted]

Subject: [redacted]

[redacted]

Answer the following questions:

Q1. How likely do you think this email is genuine or phishing?

12 Score = 12-50 = -38

Phishing Neutral Genuine

Q2. How confident are you on your answer in above question?
(1 = Not Confident at all; 100: Fully Confident)

1 100

Q3. How did you make your decision in Q1 above?

- ☐ Based upon sender's email address
- ☐ Based upon the subject line in the email
- ☐ Based upon the date in the email
- ☐ Based upon body of the email
- ☐ Based upon the link/attachment in the email
- ☐ Based upon some other reason

Q4. If you recieve this email, what will be your reaction?

- ☐ Read the email and do nothing
- ☐ Respond to this email
- ☐ Click link/ Open attachment
- ☐ Move to spam
- ☐ Delete this email
- ☐ Report this email

Submit

Email 1:

From: [redacted]

Date: [redacted]

Subject: [redacted]

[redacted]

Answer the following questions:

Q1. How likely do you think this email is genuine or phishing?

12 Score = 50-12 = 38

Phishing Neutral Genuine

Q2. How confident are you on your answer in above question?
(1 = Not Confident at all; 100: Fully Confident)

1 100

Q3. How did you make your decision in Q1 above?

- ☐ Based upon sender's email address
- ☐ Based upon the subject line in the email
- ☐ Based upon the date in the email
- ☐ Based upon body of the email
- ☐ Based upon the link/attachment in the email
- ☐ Based upon some other reason

Q4. If you recieve this email, what will be your reaction?

- ☐ Read the email and do nothing
- ☐ Respond to this email
- ☐ Click link/ Open attachment
- ☐ Move to spam
- ☐ Delete this email
- ☐ Report this email

Submit

Email 1:

From: [redacted]

Date: [redacted]

Subject: [redacted]

[redacted]

Answer the following questions:

Q1. How likely do you think this email is genuine or phishing?

88

Phishing Neutral Genuine

Score = 50-88 = -38

Q2. How confident are you on your answer in above question?
(1 = Not Confident at all; 100: Fully Confident)

1 100

Q3. How did you make your decision in Q1 above?

- ☐ Based upon sender's email address
- ☐ Based upon the subject line in the email
- ☐ Based upon the date in the email
- ☐ Based upon body of the email
- ☐ Based upon the link/attachment in the email
- ☐ Based upon some other reason

Q4. If you recieve this email, what will be your reaction?

- ☐ Read the email and do nothing
- ☐ Respond to this email
- ☐ Click link/ Open attachment
- ☐ Move to spam
- ☐ Delete this email
- ☐ Report this email

Submit

Reliability Analysis

- The inter-rater reliability analysis was conducted to identify the presence of cognitive bias in the email, it ensure the validity and consistency of the study.
- The raters were asked to rate the presence of cognitive bias on a scale of 1 to 5, where 1 indicates the absence of cognitive bias and 5 indicates the presence of cognitive bias.
- Then, the inter-rater reliability was calculated using Cohen's Kappa coefficient, which measures the agreement between two or more raters beyond chance agreement.
- The inter-rater reliability for the presence of cognitive bias was almost perfect ($\kappa = 0.89$).
- The result showed high level of agreement among the raters when identifying the presence of cognitive bias in the email.

Participants

- A total of 120 participants were randomly recruited via a crowd-sourcing platform Amazon Mechanical Turk.
- 73% Males; 27% Females
- Average age = 35 years; Min = 26; Max = 70 years.
- In terms of education level, 58 participants (96%) reported having completed either undergraduate or postgraduate studies.
- Among the participants, 34 (56%) were from an engineering background, while the remaining 26 (44%) were from other fields.

Results

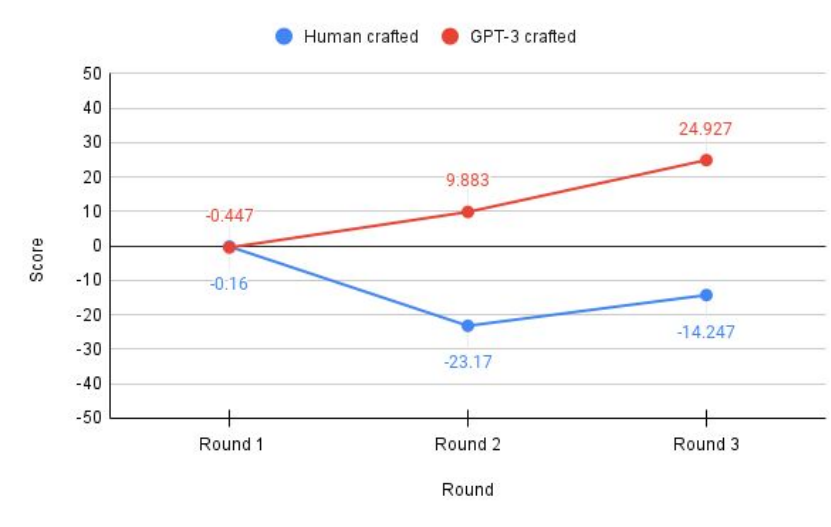

Figure 5. Average score of the participants across rounds

- Figure 5 shows the average score between -50 and +50 for human-crafted and GPT-crafted emails across rounds (involving phishing emails).
- A positive score means a phishing email marked as phishing, whereas a negative score means a phishing email marked as genuine (+50 and -50 being the maximum correct or incorrect score possible).

Results

Results revealed that human-crafted emails phished people much more than GPT- crafted emails.

Table 1. Statistics for average score of the participants across rounds

ANOVA	Statistics	Remarks
Main effect of the crafting (human or GPT)	$F(1, 298) = 81.66, p < 0.001, \eta_p^2 = 0.21$	The main effect of the crafting (human or GPT) was significant
Round 1	$F(1, 298) = 0.01, p = 0.94, \eta_p^2 = 0.00$	The difference is not significant
Round 2	$F(1, 298) = 116.04, p < 0.001, \eta_p^2 = 0.28$	The difference is significant
Round 3	$F(1, 298) = 105.04, p < 0.001, \eta_p^2 = 0.26$	The difference is significant
Learning effect (human-crafted)	$F(1, 298) = 6.96, p < 0.01, \eta_p^2 = 0.23$	Increase in performance in Round 3 (transfer) compared to Round 2 (training with feedback)
Learning effect (GPT-crafted)	$F(1, 298) = 17.98, p < 0.001, \eta_p^2 = 0.06$	Increase in performance in Round 3 (transfer) compared to Round 2 (training with feedback)

Results

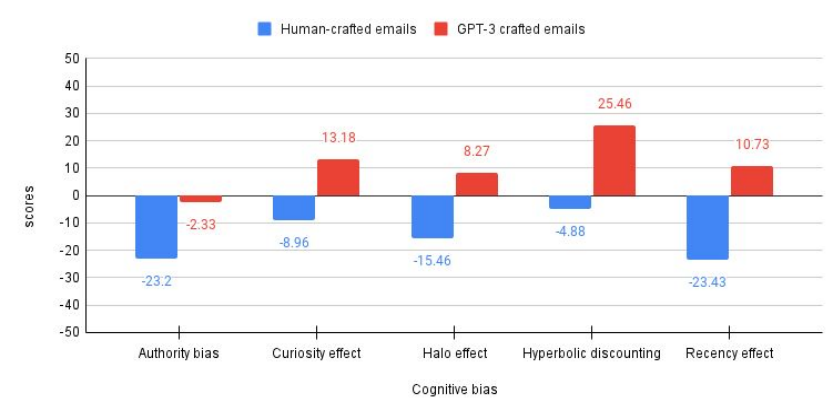

Figure 6. Average score of the participants across the cognitive biases

- Figure 6 shows the average score between -50 and +50 for human-crafted and GPT-crafted emails across the cognitive biases.

Results

The overall results showed that human-crafted emails being more potent than GPT-crafted emails. Additionally, the authority bias got the least average score in GPT-crafted emails; however, the recency effect, followed by authority bias, got the least average score in the case of human-crafted emails.

Table 2. Statistics for average score of the participants across cognitive biases

ANOVA	Statistics	Remarks
Main effect of the crafting (human or GPT)	$F(1, 238) = 79.76, p < 0.001, \eta_p^2 = 0.25$	The main effect of the crafting (human or GPT) was significant
Authority bias	$F(1,238) = 26.58, p < 0.001, \eta_p^2 = 0.10$	The difference is significant
Curiosity effect	$F(1, 238) = 24.18, p < 0.001, \eta_p^2 = 0.09$	The difference is significant
Halo effect	$F(1, 238) = 27.17, p < 0.001, \eta_p^2 = 0.10$	The difference is significant
Hyperbolic discounting	$F(1, 238) = 48.97, p < 0.001, \eta_p^2 = 0.17$	The difference is significant
Recency effect	$F(1, 238) = 66.92, p < 0.001, \eta_p^2 = 0.22$	The difference is significant

Results

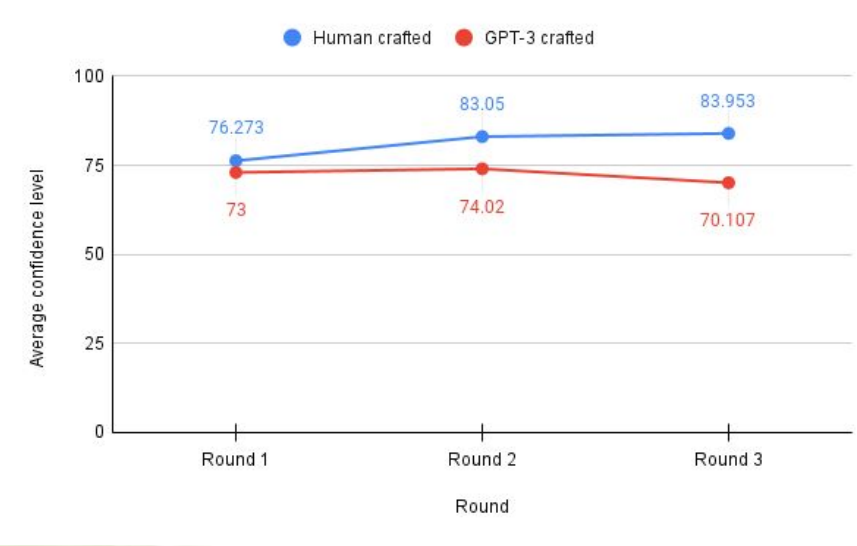

Figure 7. Average confidence level of the participants across rounds

- Figure 7 shows the average confidence level for human-crafted and GPT-crafted emails across rounds (involving the phishing emails).
- It can be illustrated from the figure that participants shows more confidence in human-crafted emails which also fooled them the most.

Results

The overall results showed that the participants who received human crafted emails they were more confident than those who received GPT-crafted emails.

Figure 5. Statistics for average confidence level of the participants across rounds

ANOVA	Statistics	Remarks
Main effect of the crafting (human or GPT)	$F(1, 298) = 18.16, p < 0.001, \eta_p^2 = 0.06$	The main effect of the crafting (human or GPT) was significant
Round 1	$F(1, 298) = 1.60, p = 0.21, \eta_p^2 = 0.01$	The difference is not significant
Round 2	$F(1, 298) = 15.50, p < 0.001, \eta_p^2 = 0.49$	The difference is significant
Round 3	$F(1, 298) = 22.55, p < 0.001, \eta_p^2 = 0.70$	The difference is significant

Results

Table 4 shows the proportion of participants' responses on the reasons that led them to diagnose emails as phishing aggregated over all rounds and participants.

Table 4. Proportion of the response of the participants for a particular reason

Reason	Proportion of responses		Statistics
	Human-crafted emails	GPT-3 crafted emails	
Based upon body of the email	0.38	0.53	$F(1, 58) = 2.639$, $p = 0.11$, $\eta_p^2 = 0.044$
Based upon senders email address	0.65	0.43	$F(1, 58) = 5.092$, $p > 0.05$, $\eta_p^2 = 0.081$
Based upon some other reason	0.00	0.00	$F(1, 58) = 0.341$, $p = 0.56$, $\eta_p^2 = 0.006$
Based upon the date in the email	0.18	0.22	$F(1, 58) = 0.157$, $p = 0.69$, $\eta_p^2 = 0.003$
Based upon the link/attachment in the email	0.09	0.21	$F(1, 58) = 3.648$, $p = 0.06$, $\eta_p^2 = 0.059$
Based upon the subject line in the email	0.49	0.28	$F(1, 58) = 6.119$, $p > 0.05$, $\eta_p^2 = 0.095$

Results

Table 5 shows the participants' reactions to phishing emails after reading them.

Table 5. Proportion of the response of the participants for a particular reaction

Reaction of the participants	Proportion of responses		Statistics
	Human-crafted emails	GPT-3 crafted emails	
Click link/ Open attachment	0.18	0.27	$F(1, 58) = 0.914, p = 0.343, \eta_p^2 = 0.016$
Delete this email	0.16	0.07	$F(1, 58) = 2.201, p = 0.143, \eta_p^2 = 0.037$
Move to spam	0.27	0.19	$F(1, 58) = 0.91, p = 0.344, \eta_p^2 = 0.015$
Read the email and do nothing	0.54	0.3	$F(1, 58) = 8.86, p > 0.01, \eta_p^2 = 0.133$
Report this email	0.15	0.07	$F(1, 58) = 1.603, p = 0.211, \eta_p^2 = 0.027$
Respond to this email	0.31	0.11	$F(1, 58) = 11.673, p > 0.01, \eta_p^2 = 0.168$

Discussion

- This research aimed to investigate the effectiveness of human-crafted phishing emails versus AI (GPT-3 crafted) phishing emails. This research also focused on the effectiveness of the presence of cognitive biases on phishing emails in phishing email detection tasks.
- The results for research question one indicate that for round 1 the score were not significantly different. However, for round 2 and round 3 the scores were different and participants who received GPT-3 crafted emails performed better than those who received human-crafted emails.
- The results revealed that scores for human-crafted emails compared to GPT-crafted emails across all cognitive biases, authority bias, curiosity effect, halo effect, hyperbolic discounting, and recency effect are less. In contrast, people seemed to show more confidence in their phished answers for human-crafted compared to GPT-crafted emails.
- The third research question revealed the differences between human-crafted and GPT-crafted emails which is due to the result of the sender's email address (more in human-crafted emails), the email's link/attachment (less in human-crafted emails), and the email's subject line (more in human-crafted emails).
- Another result showed that participants learned to recognize phishing emails in Round 3 (transfer without feedback) much better compared to phishing emails in Round 2 (training with feedback) for both human-crafted and GPT-crafted emails. However, they gained more positive for GPT-crafted emails compared to human-crafted emails. A likely reason for this finding is perhaps due to their focus on the sender's email address in human-crafted emails and their focus on email's body in GPT-crafted emails.

Conclusion

- The results indicate that human-crafted emails were more effective in fooling people than GPT-crafted emails.
- Additionally, the authority bias emails were most effective in case of GPT-crafted emails and recency effect emails were most effective in human-crafted emails in fooling people.
- However, the conclusions are limited to the text of the emails and the biases studied, the experiment only focused on certain types of emails, and there may be other types of emails that may be created for the biases studied.
- Furthermore, the presentation of emails was alternate for rounds 1 and 3, which might make participants guess the pattern.
- To overcome all these limitations, as part of our future research, we may randomize the presentation of the emails. Furthermore, we could examine how individuals identify different types of phishing emails and how individual-specific cognitive and machine learning models can be developed as decision-aids to improve our ability to identify these types of emails. Also, we plan to explore the data collected from the West with the data collected in India.

References

1. Ian Fette, Norman Sadeh, and Anthony Tomasic. *Learning to detect phishing emails.* In Proceedings of the 16th international conference on World Wide Web, pages 649–656. ACM, 2007.
2. Verizon Business. *2022 data breach investigations report.* 2022.
3. Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason I. Hong. *The devil is in the (trust) relationships: A study of the prevalence and effects of phishing attacks on high-security awareness individuals.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, pages 2267–2276, New York, NY, USA, 2010. ACM.
4. Megha Sharma, Mayank Kumar, Cleotilde Gonzalez, and Varun Dutt. How the presence of cognitive biases in phishing emails affects human decision-making? In Proceedings of the 29th International Conference on Neural Information Processing (ICONIP 2022). IIT Indore, 2022
5. Security Advisor, 2021. *Top Five Cognitive Biases Hackers Exploit the Most.* Available at: <https://securityawareness.securityadvisor.io/report-download-top-five-cognitive-biases-hackers-exploit-the-most> [Accessed 20 July 2022]
6. OpenAI. Gpt-3: Language models are few-shot learners. 2021.
7. Neeraj Kumar and Maaz Khan. A survey on phishing attacks and their countermeasures. Journal of Network and Computer Applications, 159:102687, 2020.

References

8. Bora Kim, Do-Yeon Lee, and Beomsoo Kim. Deterrent effects of punishment and training on insider security threats: a field experiment on phishing attacks. Information Systems Journal, 30:542–566, 2020.
9. Yixin Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. arXiv preprint arXiv:2107.01294, 2021
10. Iacovos Kirlappos, Angela Sasse, and Paul Grace. Exploring the use of gamification for phishing education. IEEE Security & Privacy, 19(2):56–63, 2021
11. Michael Aladesuyi-Olatunji and Mary Hamza. Psychological factors as predictors of susceptibility to phishing attacks: a review. International Journal of Cybersecurity Intelligence and Cybercrime, 1(2):23–34, 2018.
12. Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Training to detect phishing emails: Effects of the frequency of experienced phishing emails. In Proceedings of the human factors and ergonomics society annual meeting, volume 63, pages 453–457. SAGE Publications Sage CA: Los Angeles, CA, 2019.
13. Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Cognitive elements of learning and discriminability in anti-phishing training. Computers & Security, page 103105, 2023.

Thank You!