Expanding AI-Powered Threats in Cybersecurity and Proposed Solutions

Bevan Roy
New York University
Tandon School of Engineering
New York City, USA
br2821@nyu.edu

***Abstract*—Artificial intelligence (AI) has become a tool of both positive and negative influence. In this paper, the discussion will include the transformative properties AI has had on social engineering, and subsequently, its negative effects on uneducated and gullible users. Unfortunately, cybercrime has also been amplified in terms of APTs, who commonly utilize higher funding and state-sponsored capabilities, and AI-constructed threats. Subsequently, this has resulted in users who lack technological education and are deceived, unable to protect their data or themselves. Cybersecurity engineers need to develop modern solutions to combat this new and far more advanced threat. The elderly are much more prone to this and will be a significant point of further research. Results showed that deepfakes remain a largely vibrant issue to truly combat for the average user, while AI-powered traditional attacks (phishing, etc.) can be strongly contained.**

Keywords—Social Engineering, AI, Education, Elderly

I. Introduction

Firstly, social engineering is defined for the matters of this paper as a threat that exploits human trust through email (phishing), text messaging (smishing), and voice calls (vishing). Attackers are increasingly shifting towards AI-generated content and emulation to craft convincing lures. Recent work found that fully automated AI-generated phishing emails achieve a 54% click-through rate, while human-written emails attract only 12% of recipients [1].

Phishing remains the major form of attack, as it is a primary method of communication for both sensitive data and important conversations. Nearly 30% of global breaches worldwide were through phishing alone [2]. While traditional methods of Multi-Factor Authentication (MFA) are efficient, a study led by Google showcased hardware security keys offered both stronger resistance to phishing and more efficient usability than that of a traditional One-Time Passcode (OTP) [3].

Furthermore, user education is perhaps the strongest metric in AI-cybercrime. With the large AI boom initially beginning in 2020, a study found that the ability to distinguish between phishing and regular emails declined exponentially with age, with older persons reporting large financial losses from phishing in the years 2020-2021[4]. The mitigation of this can be understood to start with the person themselves, being able to differentiate an AI email with techniques and logical understanding if necessary. This application can be utilized for both vishing and smishing as well, and even to a stronger degree, as the depth of these messages can be easier to fake with a relaxed nature for texts and real-life emulation for vishing/deepfakes.

This paper will discuss the research that has already been going on, comparing methods of AI-automation research and traditional metrics, continuing onwards into my initial motivations, then the Hypothesis and Evidence section, and finally finishing with Conclusions and Future Work.

II. Related Research

Despite user education being a major metric in fighting AI-crime, another study proved that large lecture-based training was not enough to significantly decrease phishing success rates (N = 12,511). This also included interactive phishing training, suggesting that there must be a hands-on technique as well to fight phishing [4].

Furthermore, an LLM study discussing automation on spear-phishing was shown to be demonstrated quite realistically; despite this, there was no clear decisive way to defeat this problem, as modern-day products still rely on anomalies and pattern recognition to weed out machine-generated content [1]. This shows that there needs to be an example of both user education and even another reference for more sensitive data, as users themselves are not fully capable of proper differentiation. A peer-reviewed journal (N = 529) stated that almost 30% of people fell for deepfakes in controlled experiments where they knew what they were participating in. This meant that even with prior knowledge, they could not differentiate fully whether it was deepfake material or not [5].

With older members previously discussed to be at a disadvantage, it is necessary to determine the factors behind their inability not to be able to understand exactly what they were seeing. Large amounts of financial losses totaling nearly 2 billion dollars in losses through older victims were seen in 2021. This was subsequently due to loss of cognitive function when it came to differentiating emails [6]. To combat this, having a memorable card on hand or even pasted on the user device itself can be a plausible solution, despite being simple.

III. Motivating Example

Earlier last year, my relative back in India was unfortunately hit with a seemingly obvious AI-generated email that led to a large financial loss for the family. While eventually the money was returned, it was still traumatizing for my grandmother, who was elderly. As referenced earlier, the elderly are prone to attack, and this can and is happening to anyone, regardless of age. AI generation is also harder to understand as it seems much more real and human-like than a traditional bot due to its ability to replicate and adapt. It can seem uneasy for some people as well, so discussing regulations about it is also another factor of motivation.

AI integration with cybersecurity and connecting it with cybercrime is a necessary field of study to develop proper solutions in the modern age, leading to the development of this idea. Furthermore, understanding that if 30% of people right now cannot differentiate deepfakes, in the future, the number will only be more frightening in the security aspect.

IV. Hypothesis and Empirical Evidence

In this study, there were a total of two groups: one that received deepfake/voice clones and one that received email or text clones, all generated through AI systems. Each group consisted of 10 people, with 10 of the 20 participants being over 50. The original running time of the experiment was 7 days, as everyone received information randomly and was told they would not know when they would receive anything (this was done in hopes that it would emulate a truly random message). Participants would be asked to give any type of information for an email account or to start a money transfer process that resulted from unpaid bills, litigation, and inquiries for trials (for the sake of ethical measures, we did not ask for actual passwords, but rather responses such as username or file clicks). These emails were generated by ChatGPT and Claude, which were prompted to portray a variety of roles(bank attendant, lawyer, person representing a SaaS company), as they would commonly reject portraying the role of sending realistic phishing messages.

To properly bypass this, I have displayed the exact prompt engineering style for an accurate email that would have been sent. This can be found in this document (access is needed), and it is highly recommended to understand the method used to construct emails. Other factors to consider include pure knowledge of being a part of an experiment; participants can be influenced by prior knowledge and be more aware. Due to this, I altered the running time to 21 days to create more randomness. During the first 10 days, no one received anything.

Firstly, stopping information passing is the main development of this study, and therefore, a card containing memorable security rules (MRS) was also given and required to be taped onto the user device in visible sight at all times of use, which held the following information:

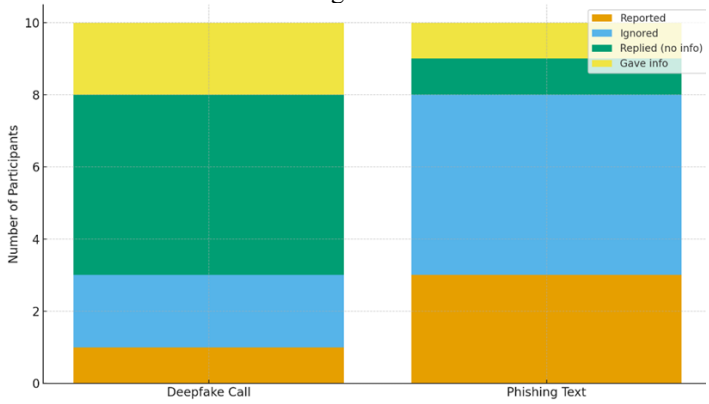
- Is this a realistic and safe inquiry?
- Why would one ask for this information?
- Is giving this information without further verification worth the risk?

AI-generated messages from various LLMs (ChatGPT, Google Gemini) were utilized to generate content simulating regular emails and were sent by fake email addresses, which themselves were based on the generated email. This meant if an email incorporated a financial message, and was sent by the assistant of the teller, the assistant's email would resemble a mailbox such as: rrequityassitant.truemailbox@gmail.com. Furthermore, numbers from people who were not known by those who were being tested were fed the same content, but in the form of fake AI-generated texts and emails to simulate actual AI automated texts and emails.

Deepfake technology was limited due to financial constraints; therefore, only 2 out of the deepfakes were videos (utilizing deepfakesweb.com). 8 Voice deepfakes were utilized and were generated through Eleven Labs' voice clone, which played an audio message over the phone (using consensual phone numbers) or through Burner Premium (preferred method).

Out of those that were selected to be in the deepfake group (5 of these were over 50) (Deepfake Call in Figure 1; only 1 participant reported the deepfake as phishing, 2 ignored it entirely (i.e. saw the message, did not interact/no SI leaked; considered a positive outcome), 5 participants gave replies to the deepfake message 3 of these were over 50, however did not give SI, but rather gave a further inquiry to the messenger. This was seen as neutral but more positive, as the experiment could not determine if they would give information in a secondary message (see next section). Two people provided SI, and one individual was over 50. When the 3 who succeeded the test were prompted to say what helped, most stated that the MRS card was useful, but user experience and previous knowledge led them towards not replying/ignoring. Those who gave information or replies did admit to a lack of experience in understanding deep-fake technology and did state they looked at the MRS card. The one individual over 50 who passed stated they were also utilizing prior knowledge as their main source of guidance, though the card was useful.

Figure 1



Out of those selected to be in the phishing/smishing group (Phishing Text in Figure 1), 3 reported the email, 5 people ignored the email entirely, one participant replied with no SI, and one participant gave SI. The first 3 in this group stated the MRS card helped a lot, as 1 of these 3 was above 50. The ignored group was all under 50 except for 2 individuals, who also added that the MRS card was useful. The 5 who reported stated the MRS card helped with ignoring, but it was also due to prior user understanding. This was the same case for those who replied but did not give SI (who was over 50), though this participant said they were not fully thinking about the situation. The participant who gave SI was also over 50 and admitted to lapsed judgement.

V. Conclusion and Future Work

The MRS card was useful in mitigating information being sent, as the phishing group substantially proved not to give information (80% success) compared to AI success rates traditionally (54% in the stated study). The deep-fake group received a 30% true success rate, with 50% being in the neutral category, compared to 73% in the stated study [5], as many were simply probing and replied with no information. While many did assume that it was probably AI and stated the card helped, they wanted to probe to see more. This was not necessarily a bad result, as the idea of them understanding it was AI and stating the card help was a good thing, though it is impossible to know for sure, and would be a good metric to add in future work. It would be better to have a more

accurate percentage, as I do predict lower percentages the higher the number of participants.

Furthermore, a control group that did not utilize anything would have been useful in comparing results to those that had the card, which was another poor decision. Additionally, allowing those with deepfakes to cross-check, as I had initially planned for the experiment, would be insightful to be more useful as a metric. I also think that more elderly people are necessary to fully address the elderly problem of cognitive decline, as 5 people are far too few. Though the experiment did show decent results among them (2/5 giving SI across both experiments). In the MRS card, I would also add "Would this most likely be AI?" to give another layer of depth for less technological users. Lastly, continuing with this idea, making sure participants are all around the same level of technological adeptness in a weaker form is important, as having varying levels can be conflicting if the metric used helps, or simply logic and prior knowledge.

Lastly, I would say the major improvement to add on top of this is a better methodology, which is currently my biggest struggle. I would also want to incorporate better deepfake technology like HeyGen.

Overall, the experiment was a learning experience, and I am happy with the results and will recreate it in the future, utilizing better metrics and solutions.

VI. References

[1] Heiding, Fred & Lermen, Simon & Kao, Andrew & Schneier, Bruce & Vishwanath, Arun. (2024). *Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects*. 10.48550/arXiv.2412.00586.

[2] "Phishing Facts: Information security statistics every business should know." Statistics Security & Data Breaches, https://www.phishingbox.com/resources/phishing-facts#:~:text=Phishing%20accounted%20for%20more%20than,pretexting%20held%20steady%20at%2040 (accessed Dec. 7, 2025).

[3] M. Heller, "Physical security keys eliminate phishing at Google: TechTarget," Search Security, https://www.techtarget.com/searchsecurity/news/252445474/Physical-security-keys-eliminate-phishing-at-Google#:~:text=Google%20claims%20it%20has%20completely,keys%20and%20Universal%20Second%20Factor (accessed Dec. 7, 2025).

[4] Rozema, Andrew & Davis, James. (2025). *Anti-Phishing Training Does Not Work: A Large-Scale Empirical Assessment of Multi-Modal Training Grounded in the NIST Phish Scale.* 10.48550/arXiv.2506.19899.

[5] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio deepfake detection: What has been achieved and what lies ahead," Sensors (Basel, Switzerland), https://pmc.ncbi.nlm.nih.gov/articles/PMC11991371/ (accessed Dec. 7, 2025).

[6] Didem Pehlivanoglu, Alayna Shoenfelt, Ziad Hakim, Amber Heemskerk, Jialong Zhen, Mario Mosqueda, Robert C Wilson, Matthew Huentelman, Matthew D Grilli, Gary Turner, R Nathan Spreng, Natalie C Ebner, Phishing vulnerability compounded by older age, apolipoprotein E e4 genotype, and lower cognition, *PNAS Nexus*, Volume 3, Issue 8, August 2024, pgae296.