Detecting Social Engineering Scams While Preserving User Privacy in the Digital Era (Proposal Position Paper)

Atul Prakash
University of Michigan
aprakash@umich.edu

Shivani Kumar
University of Michigan
shivanii@umich.edu

Elisa Tsai
University of Michigan
eltsai@umich.edu

***Abstract*—The escalation of social engineering (SE) scams via messaging tools and phone calls on mobile devices presents a critical threat to individual privacy and financial security. The less tech-savvy segments of the population are particularly vulnerable to such scams. Detecting such scams is hard because of their multi-modal nature and privacy concerns with any solution – the interactions can include phone calls and messages in a certain context where the interaction is expected (e.g., user selling an item online, user needing tech support, etc.) and monitoring such interactions for the analysis itself can introduce security and privacy risks. This proposal proposes this area as one in urgent need of investigation. We discuss a simple version of the problem to motivate the research challenges and then discuss potential research areas that may contribute to designing an infrastructure to detect and prevent such scams.**

1. Introduction

In the digital age, fraud has emerged as a pervasive challenge, often disproportionately affecting the most vulnerable among us, such as seniors. Yet it is a problem that spares nobody, transcending age and education. In 2023, a former White House science advisor lost almost all her life savings. Misled into believing she was speaking to the fraud department of her bank, she transferred her funds into a "safe" place, which turned out to be controlled by the attackers [7]. This incident is not an isolated one; similar cases of social engineering, where individuals are manipulated into sending money to fraudsters, occur with alarming regularity [2], [9].

The research question we propose to the community is the following: *Is it possible to build frameworks that help potential victims in detecting such scams and being alerted to them as they are happening?*

Developing such frameworks poses significant challenges due to the need for advancements across operating systems, user interfaces, and multi-modal analysis [3], [4]. To illustrate, consider a scenario in India where bank transactions require a one-time password (OTP) for completion. Apps like Google Pay, PayTM, and PhonePe utilize a system called UPI (Universal Payment Interface), mandating OTP entry for transfers. Similar infrastructure for mobile money transfer, now proposed by the Federal Reserve for the U.S. [1], is gaining global adoption.

Despite appearing secure [6], many users still fall victim to scams, losing money in transactions where they expect to receive funds. For instance, a victim selling an item online may receive a call from a buyer claiming to be a remotely stationed soldier, offering to purchase at full price via UPI to "reserve" the item, and establish trust by sending a fake Army ID. The buyer then sends a QR code representing their UPI credentials. When scanned, it prompts the victim to send money instead of receiving it. The victim, convinced of the buyer's legitimacy, enters an OTP as instructed, inadvertently transferring money to the scammer. The above is unfortunately not a hypothetical scenario. This is a common scam in India and is often reported in newspapers [5]. Despite awareness, it continues to dupe many, leading to significant financial losses.

Even in this simple scenario, building a fraud detection tool is very challenging, even though all the communication and financial transaction (typically) takes place on a single device – the victim's mobile phone. We propose the following key research challenges (among others):

- Detecting such events automatically, requires addressing the following research question: **How can we specify dark patterns that can be automated?** An example simple dark pattern [8] could be receiving an OTP while being on a phone call, suggesting a possible prelude to an attack.
- The detection code will need multi-modal access in order to be effective (such as phone calls, SMS, and email privileges). Receiving all these privileges raises another question: **How can we ensure the privacy of monitored data by a potential fraud detection app?** Support from the operating system is crucial as the detection code requires extensive privileges to monitor user interactions. Furthermore, any monitoring apps must be guaranteed to be sandboxed to prevent them from becoming a potential source of privacy leaks.
- Given the details of scams published by newspapers or authorities, **can Large Language Models (LLMs) automatically generate specifications of the dark patterns?**
- Finally, **what advances are required in operating systems and applications to permit safe and**

private multi-modal analysis across different apps used for communication (e.g., phone calls, SMS, WhatsApp)?

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments on the paper. This work is supported by a gift from the OpenAI Cybersecurity Grant program. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of OpenAI.

References

[1] "The FedNow Service," 2023, https://www.federalreserve.gov/paymentsystems/fednow_about.htm.

[2] AARP, "Americans lost record-breaking $8.8 billion to scams in 2022," 2023, https://www.aarp.org/money/scams-fraud/info-2023/ftc-consumer-losses.html.

[3] R. Chaganti, B. Bhushan, A. Nayyar, and A. Mourade, "Recent trends in social engineering scams and case study of gift card scam," *arXiv preprint arXiv:2110.06487*, 2021.

[4] A. Derakhshan, I. G. Harris, and M. Behzadi, "Detecting telephone-based social engineering attacks using scam signatures," in *Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics*, 2021, pp. 67–73.

[5] HindustanTimes, "Delhi Police warn OLX users on QR code frauds. Here's how you can avoid losing money," 2023, https://www.hindustantimes.com/india-news/delhi-police-warn-olx-users-on-qr-code-frauds-here-s-how-you-can-avoid-losing-money-101615771153492.html.

[6] R. Kumar, S. Kishore, H. Lu, and A. Prakash, "Security analysis of unified payments interface and payment apps in India," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1499–1516.

[7] M. Laris, "A former White House scientist was scammed out of $655,000. Then came the IRS," 2023, https://www.washingtonpost.com/dc-md-va/2023/12/14/cyber-crime-scams-irs-taxes/.

[8] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark patterns at scale: Findings from a crawl of 11K shopping websites," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–32, 2019.

[9] Politico, "€1.8B worth of fraud in EU in 2022, watchdog says"," 2023, https://www.politico.eu/article/olaf-reports-increased-eu-funds-fraud/.