Experiencing Deceptive AI: A Qualitative Study of Deepfake Fraud Victimization

Yichen Zhang
University of Michigan

Lu Xian
University of Michigan

Florian Schaub
University of Michigan

1 Introduction

Deepfake fraud—the malicious use of AI-generated media to impersonate individuals and fabricate events—has emerged as a serious threat to digital security and interpersonal trust [2]. Unlike traditional scams such as phishing [1] or identity theft [6,7], deepfake-enabled fraud exploits synthetic audio or video to create persuasive deceptions, often mimicking trusted figures. Notable cases, such as the $25 million fraud involving a deepfaked executive in Hong Kong [5], demonstrate the growing sophistication and stakes of these attacks.

As fraudulent practices become more sophisticated [8,9], there is a pressing need to understand not only the technical aspects of deepfake detection but also the victim experience, associated harms, the psychological toll, and the effectiveness of existing countermeasures. To address these concerns, we conducted a semi-structured interview study with victims of deepfake fraud exploring the psychological and practical implications of such encounters. Our study addressed the following research questions:

RQ1: What are the types of deepfake or AI-based fraud and scams that people encounter and what are their mental models of these scams?

RQ2: How do individuals recognize or detect deepfake fraud?

RQ3: What are the emotional, psychological, and financial impacts of being a victim of deepfake fraud?

RQ4: What are the participants' experiences with and attitudes towards reporting deepfake fraud?

Drawing on 7 semi-structured interviews, this study reveals that individuals often associate deepfakes with novelty rather than threat, leading to low perceived risk and reliance on intuitive trust cues like familiar voices or emojis. Social familiarity amplified credibility, delaying fraud detection. Participants described lasting emotional and relational impacts.

Grounded in Protection Motivation Theory [3,10] and Expectancy Violation Theory [4], we discuss how perceived authenticity, low threat awareness, and social context shape responses to deepfake fraud. We conclude with recommendations for public education, detection tools, and institutional reporting mechanisms.

2 Methodology

This study is based on a semi-structured interview design, supported by a brief screening survey used to identify and recruit participants with relevant experiences. Participants were recruited through multiple channels, including personal networks, referrals from those individuals, and outreach via social media platforms, a university mailing list, and Prolific. The screening survey collected basic demographic information and asked whether participants had encountered deepfake-related scams, either personally or through someone they knew. Individuals who indicated relevant experiences were asked to describe them. Those who expressed interest in further participation were invited to take part in interviews.

The interviews explored participants' understanding of deepfakes, their direct or indirect encounters with synthetic media fraud, and the emotional, cognitive, and behavioral impacts of those experiences.

All study procedures, including the recruitment process and interview protocol, were reviewed and approved by the institution's internal ethics review board (IRB).

3 Recruitment

7 participants ultimately completed interviews from a pool of 336 initial survey respondents. It is challenging to identify

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

individuals with direct experience of deepfake fraud, despite extensive recruitment efforts. The sample skewed younger and more educated, as most participants were recruited via Prolific, social media, and university mailing lists.

Participants provided personal accounts and descriptive details but they were self-reported and their statements were not technically verified. Despite this limitation, the interview data revealed consistent themes across participants, suggesting a level of conceptual saturation and enabling theory-informed analysis. The qualitative richness of the narratives, including detailed descriptions of emotional and social consequences, provides valuable early insight into how individuals interpret and are affected by deepfake-related deception.

4 Key Findings

The experiences of the participants revealed several patterns in perception, interpretation, and response to deepfake fraud.

Deviated Mental Model of Deepfake The majority of the participants had minimal or skewed mental models of deepfakes. While most were generally familiar with the concept, their prior exposure was mainly to clearly artificial or humorous content such as celebrity face swaps or political parodies often encountered on social media. As a result, they tended to associate deepfakes with entertainment or novelty, rather than serious threats. This was due to underestimating them, and thus they had a low sense of urgency and unreadiness when handling actual deceptive content which was far less overt and contextually plausible.

Vulnerability Peak Social familiarity played a crucial role in shaping susceptibility. Participants were most vulnerable when the impersonated individual was someone they were moderately familiar with such as a former classmate, distant relative, or coworker. People with close relation can easily recognize the person impersonated and protect themselves from being scammed. One participant, for example, described recognizing that a deepfake video of her sister.

> "...I've known my sister my whole life, so I am very familiar with her and her mannerisms... And I could immediately clock and say, like, that's not her..."

People with far relation have low potential to transfer money to the person impersonated even though they cannot recognize the other. Distant or online-only relationships often provided some protection, as low emotional closeness encouraged greater skepticism. As one participant explained, her familiarity with the social patterns of an acquaintance helped her evaluate the implausibility of a deepfake video

> "...I think that maybe people who are friends with her, but don't hang out with her as often would probably think that that was the real video..." (P3)

Mid-distance relationships created sufficient familiarity to prompt trust but not enough intimacy to detect small deviations in speech, actions, or communication style. This "gray zone" of social proximity emerged as a key risk factor across multiple cases.

One participant recounted an incident involving a message from someone posing as a former classmate.

> "It was his face and name and everything... and he called me by my nickname, which made it feel personal."

Plausibility Over Proof Participants evaluated authenticity primarily based on contextual plausibility. If the message seemed emotionally appropriate or aligned with previous interactions, they were less likely to question it even when there were minor red flags. Visual or auditory imperfections were often overlooked in favor of content that "felt right."

Psychological Distress and Trust Erosion The impact of deepfake fraud extended beyond the moment of deception. Participants reported ongoing emotional distress, including anxiety, insomnia, and lingering self-blame. Some participants experienced fractured relationships—not with the scammer, but with the person who was impersonated. In one case, a participant rejected her friend's claim that a suspicious video was a deepfake, leading to mistrust and the end of their friendship. Others described a broader erosion of trust in digital communication, expressing reluctance to rely on voice or video as proof of identity after the incident. Many also expressed deep frustration with the institutional response, citing slow investigations, lack of resolution, or inadequate support from both police and platforms.

5 Implications

Deepfake fraud reveals how trust can be manipulated not only through realistic media but also through the social context in which it is embedded. The effectiveness of these scams often depends less on technical sophistication and more on how well they align with familiar relationships and everyday communication patterns.

The study suggests that responses to deepfake fraud must account for these social and emotional dynamics. Technical detection tools are important, but they are not sufficient on their own. Public education is important not only in raising awareness of deepfakes as a general concept, but in helping people recognize the subtle ways deception can unfold within trusted networks. At the same time, institutions such as social media platforms and law enforcement must take active roles in supporting users who report such incidents.

References

[1] Anti-Phishing Working Group (APWG). Phishing activity trends report, 2022. Retrieved from https://apwg.org/trendsreports/.

[2] Jon Bateman. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace., 2022.

[3] Hendrik Boer and Erwin R Seydel. Protection motivation theory. In *Predicting health behaviour: Research and practice with social cognition models. eds. Mark Conner, Paul Norman*, pages 95–120. Open University Press, 1996.

[4] Judee K Burgoon. Expectancy violations theory. *The international encyclopedia of interpersonal communication*, pages 1–9, 2015.

[5] CNN. Finance worker pays out $25 million after video call with deepfake 'chief financial officer'. 2024. Retrieved from https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

[6] Federal Trade Commission. Consumer Sentinel Network Data Book, 2022.

[7] Javelin Strategy & Research. Identity Fraud Study, 2022.

[8] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026, 2023.

[9] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.

[10] Ronald W Rogers. Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation. *Social psychology: A source book*, pages 153–176, 1983.