Harnessing Citizen Science for Comprehensive Deepfake Research: Balancing Data Collection and Ethical Consideration

Thai Le - Computer Science, Luddy, Indiana University, Bloomington

Thai Le got his doctorate degree from the College of Information Science and Technology (IST) at the Pennsylvania State University. He was awarded the IST Ph.D. Student Award for Research Excellence and was a DAAD Postdoctoral Fellow. He has industry research experience at Amazon Alexa, Yahoo Research and VMWare. Dr. Le's mission is to enhance the robustness, safety, and transparency of Machine Learning and Artificial Intelligence algorithms, especially Natural Language Processing models with critical applications in cybersecurity contexts, ensuring that the society and netizen can harness their power with safety and clarity. His work has been published in venues such as ACL, EMNLP, NAACL, AAAI, AAMAS, KDD, WebConf and CHI, and many of his interdisciplinary work have been featured in ScienceDaily, DefenseOne, and Engineering and Technology Magazine.

In an era where digital deception threatens the very fabric of our online reality, imagine a world where everyday citizens become the frontline defenders against the rising tide of deepfakes. The proliferation of deepfake technology poses an unprecedented threat to online trust and information integrity. Such technology also fueled a surge in devastating cybercrimes, including sextortion schemes targeting vulnerable individuals, sophisticated financial frauds leveraging fake video evidence, and large-scale disinformation campaigns. Current deepfake research primarily relies on synthetic datasets or limited real-world examples. While valuable, these resources often fall short in capturing the full complexity of actual deepfake incidents as many important contextual information are missing, including how the victims were approached with deepfake, through what conversations and on what platforms, what are the social dynamics of the victims and the malicious actors, and the psychological experience of the victims. ***This gap between data synthesis and real-world data hinders our ability to develop truly effective countermeasures.*** To move the science forward, we urgently need comprehensive, real-world data for deepfake research. Only through holistic understanding of deepfake frauds and scams through real-world data, can we hope to create AI systems capable of detecting, mitigating, and ultimately preventing the harmful effects of deepfakes. To do this, ***we need a framework that allows for comprehensive data gathering while rigorously protecting the privacy and dignity of victims***. This requires not only technological solutions but also the development of ethical guidelines and community-driven research protocols.

To address this challenge, we propose leveraging citizen science and human-in-the-loop as a vital approach to bridge the knowledge gap in deepfake research. The most important research objective is then to create a secure, ethical framework for collecting real-world deepfake data through citizen participation. By allowing deepfake victims to report their own experience with deepfake in an ethical and responsible manner, we hope that this approach will yield a more diverse and representative dataset, enabling the development of more effective and contextually aware AI pipelines for deepfake detection and mitigation. Here we propose a three-tier framework for **citizen science based deepfake research** (Figure 1) with ***security, privacy-preserving, and ethics as the foundation***:

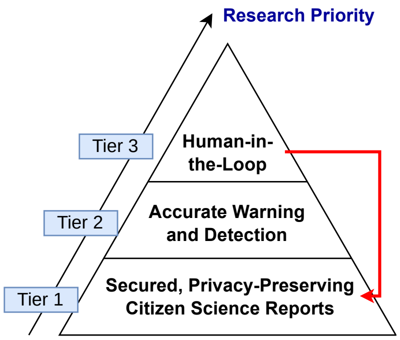
Research Priority
Tier 3
Human-in-the-Loop
Tier 2
Accurate Warning and Detection
Tier 1
Secured, Privacy-Preserving Citizen Science Reports

Figure 1: Framework of the Proposed Citizen Science-Based Deepfake Research

- **Tier 1. Ethics -** Develop a secure, privacy-preserving, and anonymized platform for citizen scientists to report and document deepfake incidents and establish a community oversight board to ensure ongoing ethical compliance and victim protection.
- **Tier 2. Accuracy -** Design and implement AI algorithms that can analyze both the technical and social aspects of reported deepfakes and effectively warn the users of potential deepfake scams and frauds.
- **Tier 3. Continual Learning -** Develop interpretability framework that can explain the developed AI algorithms and use the resulting explanation as a feedback mechanism to **guide citizen scientists as**

human-in-the-loop to provide additional or missing signals that might be useful to continually improve the AI models. This is very important as the malicious actors change their scam strategies overtime, not only with different persuasive tactics but also with new deepfake generation technologies. Such signals can then be fed to the security, privacy-preserving and ethical framework developed in Tier 1, creating a comprehensive, closed-loop pipeline.

Successful completion of this research will revolutionize our approach to combating deepfakes. By creating an ethically sourced, comprehensive dataset of real-world incidents, we will enable the development of more effective detection and mitigation strategies. Moreover, this project will set a new standard for responsible AI research, demonstrating ***how citizen engagement can be harnessed to address critical technological challenges while prioritizing the protection of vulnerable individuals***. Ultimately, this work will contribute to a safer, more trustworthy digital ecosystem, reinforcing the foundations of safety online.