

Guardian AI: Scam detection and prevention

Andres Sepulveda Morales

Doroteya Stoyanova

Patrick Huang

University of Aberdeen

Wanjie Zhong

University of Aberdeen

In the next 1-3 years, we anticipate a significant increase in sophisticated scams targeting vulnerable populations, particularly the elderly, across multiple communication channels.

We are introducing GuardianAI, an AI-powered application designed to protect users from fraud and identity theft to address this growing threat. GuardianAI uses advanced machine learning techniques to analyse text messages, emails, and voice calls in real-time, identifying potential scams and alerting users and their trusted contacts.

By leveraging multi-modal AI, adaptive learning, and explainable AI, GuardianAI provides a comprehensive solution that adapts to evolving scam tactics while maintaining user privacy and ease of use for technologically inexperienced individuals.

Demo link: <https://ai-scam-detection.streamlit.app/>

1. Problem overview

Identity theft and fraud have become alarmingly prevalent, with imposter scams emerging as the most common method. In 2023, the US Federal Trade Commission received over 856,000 reports of imposter scams, accounting for 43% of all fraud complaints (Liu and Fair) and \$2.7 billion in losses (Akin). The UK saw 45,367 impersonation scam cases in 2022, costing £177.6m (UK Finance).

Scams target vulnerable individuals, including the elderly, with four people aged 50+ scammed in England and Wales every minute (Age UK). Older people often face unique risk factors, including isolation, digital literacy skill needs, and cognitive impairments, making them prime targets for fraudsters (Kamoto). Scammers communicate with potential victims virtually, from phishing emails and deceptive text messages to fraudulent phone calls (NCSC UK). These scams exploit trust and the lack of technological literacy, leading to severe financial losses and emotional distress.

In this context, the elderly often rely on their adult children or trusted individuals to help them navigate these threats. However, manual monitoring and intervention are impractical, and existing solutions like basic spam filters or call blockers do not fully address the sophisticated nature of modern scams often now powered by AI (PWC).

The challenge lies in developing an AI system that can effectively analyse diverse communication channels, adapt to new scam patterns, and provide timely alerts while respecting user privacy and maintaining ease of use.

2. Your solution

Our proposed solution is an AI-powered mobile and desktop application called "GuardianAI" that leverages advanced machine learning techniques to detect and prevent scams across various communication channels. The app will integrate seamlessly with users' devices to

monitor communication channels, e.g. text messages, emails, and voice calls, for potential scam indicators. The key features with related problems are defined below.

- **Multi-modal:** *Scammers use various communication methods to target victims and traditional rule-based systems can't keep up with evolving scam tactics.* The core of GuardianAI will be a sophisticated AI model trained on vast databases of known scams, e.g. ScamSearch and government and financial institution databases. The solution will employ advanced neural networks (Kamungu) and natural language processing (NLP) for text analysis, speech recognition and analysis for voice calls, and computer vision for image-based scams. Given the diversity of data, the solution will implement deep learning models to identify complex scam patterns across various data types e.g. extending neural network models for communication across channels (Cheng).
- **Real-time monitoring and alerts:** *Voice-based scams are particularly challenging to detect.* The app will provide real-time analysis of incoming communications utilising edge AI and efficient model compression techniques (Singh) to instantly flag suspicious content and alert the user and their designated trusted contacts.
- **Trusted network integration:** *Vulnerable users may not always recognise scams, even when warned or notified.* The app will allow users to designate trusted contacts who can receive alerts and assist in decision-making when potential scams are detected.
- **User feedback for adaptive learning:** *Models need continuous improvement to stay effective.* Besides frequently gathering data from scam databases, GuardianAI will incorporate a user feedback system, allowing individuals to report suspected scams (Kaufmann). This data and continuous updates to scam databases will be used to continuously improve the AI model's accuracy and adapt to new techniques.
- **Explainable AI (XAI):** *Users may not understand or trust AI-based warnings.* To build trust, GuardianAI will include explainable AI that provides clear, understandable reasons as to why a particular communication was flagged as suspicious (Lin)
- **Privacy preservation:** *Users may have concerns about sharing data.* Using advanced federated learning techniques (Truong), GuardianAI will ensure user data remains on their devices while contributing to improving the global AI model.
- **Persistent behavioural analysis:** *Some scams are only detectable through patterns of behaviour over time.* GuardianAI will include AI-driven behavioural analysis (Hussain) to identify unusual patterns in user communications that may indicate ongoing scam attempts.
- **Voice clone detection:** *Advanced scams may use AI-generated voice clones to impersonate trusted individuals.* The solution will integrate voice authentication technology (Chandran) to detect synthetic or cloned voices in calls.
- **Proactive scam simulation and education:** *Users lack experience in identifying scams.* The solution will offer AI-generated, personalised scam simulations to educate users on recognising and responding to potential threats.
- **Cognitive assistance:** *Elderly users are generally less tech-savvy.* For elderly users, GuardianAI will include a conversational AI assistant with text-to-speech capabilities to explain potential risks and guide users through safe online practices.

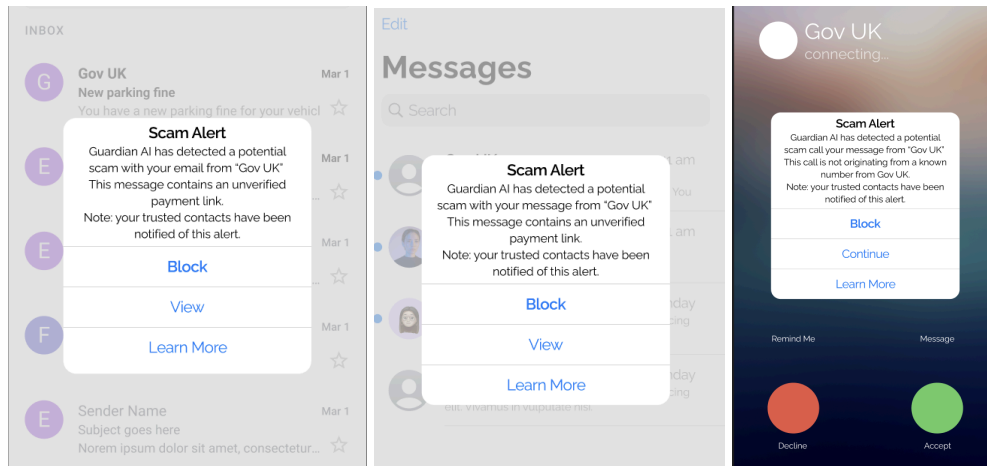
3. Pilot experiment or demo

We utilised the OpenAI API to handle text and audio processing for this demo. Our website was built using Python and the Streamlit platform to facilitate rapid deployment and focus on the demo content. This demo was split into different attack vectors - text message, email, and audio - to showcase various applications of a model towards GuardianAI. We also included a

rudimentary "scam education" section to provide insight on what kind of training we could provide users to guide against scams as new methods arise proactively.

Building a total solution for GuardianAI would require extensive research into which models can facilitate specific workflows, fine-tuning both input and output, and leveraging other API solutions to provide a deeper analysis of the likelihood of a scam with the attack vector provided. Solutions like Hive.ai and Resemble.ai are likely candidates to handle situations where a transcript alone cannot prove whether the communication is a scam or AI-generated.

We have also included mockups to showcase how Guardian AI *might* help with user scam prevention in real-time.



Figures: potential alerts from Guardian AI on a user's device across all communication channels: email, messages/SMS, and voice call

4. Process

With unlimited resources and one year, our commercialisation steps would include:

- Expand data collection and model training to cover a wider range of scams
- Develop iOS and Android versions of the mobile app and create a mobile app or a web interface for trusted contacts
- Implement email, messaging, and voice call analysis features
- Conduct extensive user testing with the target demographic to refine the UI/UX
- Partner with companies for pre-installation
- Develop an API for integration with other communication platforms and apps

Timeframe	What will you do?
Next 3 months	MVP/Validation • Refine and expand the scam detection model • Develop MVP Android and iOS apps with email, messaging, and voice call analysis • Begin beta testing with a small group of users in one English-speaking country
2025	Product-Market Fit • Deploy real-time voice analysis during calls

	● Implement AI-powered chatbot for user education and support ● Implement advanced user behaviour analysis to detect social engineering attempts ● Experiment with native integrations with communications platforms and devices e.g. major email providers, messaging apps, device manufacturers
2026	Scale ● Scale to other English-speaking countries ● Start building the solution for other languages ● Identify and experiment with different customer and user segments, e.g. children/families, Gen Z, small-to-medium businesses, and potential partnerships with larger corporates who likely have similar cybersecurity solutions already in place ● Identify and experiment with related solutions beyond scam detection to provide holistic digital safety and well-being for vulnerable users, e.g. identity theft/fraud insurance, bank/credit card integration ● Develop predictive models to anticipate and prevent potential scam attempts before they occur
2027	Expansion ● Expand to non-English-speaking countries, other customer/user segments, and related solutions ● Integrate with other devices, e.g. smart home devices, for comprehensive protection ● Develop AI-powered virtual assistant

5. Impact on AI safety & key risks

We expect that this solution has a positive impact as follows:

- Protects vulnerable populations from financial and emotional harm
- Contributes to the broader fight against cybercrime and online fraud
- Advances the development of explainable AI in consumer applications

We anticipate these key AI-related risks and defined mitigation strategies as follows:

- **Privacy Concerns:** Users may be concerned about sharing their communication data. We intend to use federated learning and data encryption protocols to address privacy.
- **False Positives/Negatives:** Over-flagging legitimate communications as scams or missing genuine threats could erode trust. The app will explain flagged items clearly and allow easy reversals.
- **Adaptation by Scammers:** As scammers evolve, continuous model updates and community-driven feedback will be crucial to stay ahead of new tactics.
- **AI Safety Challenges:** Robust testing against adversarial attacks and ongoing research into AI safety will ensure the models remain secure and reliable.

6. Appendix

Works Cited

- Age UK. "Scams Prevention and Support Programme." *Age UK*, 7 August 2024, <https://www.ageuk.org.uk/our-impact/programmes/how-we-deliver-advice/scams-prevention-and-support-programme/>. Accessed 30 August 2024.
- Akin, Jim. "Identity Theft Statistics: Fraud Is on the Rise." *Experian*, 25 July 2024, <https://www.experian.com/blogs/ask-experian/identity-theft-statistics/>. Accessed 30 August 2024.
- Chandran, Deepak Ramesh. "Use of AI Voice Authentication Technology Instead of Traditional Keypads in Security Devices." *Scientific Research Publishing*, 2022, <https://www.scirp.org/journal/paperinformation?paperid=117753>. Accessed 31 8 2024.
- Cheng, Dawei. "Spatio-Temporal Attention-Based Neural Network for Credit Card Fraud Detection." *Spatio-Temporal Attention-Based Neural Network for Credit Card Fraud Detection*, <https://ojs.aaai.org/index.php/AAAI/article/view/5371/5227>. Accessed 31 August 2024.
- Hussain, Altaf. "AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems." *AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems*, 2023, <https://www.sciencedirect.com/science/article/abs/pii/S0952197623014021>. Accessed 31 8 2024.
- Kamoto, John. "Scams Prevention and Support Programme Report: Empowering Older People in a Fraud Epidemic." *Age UK*, 2024, <https://www.ageuk.org.uk/globalassets/age-uk/documents/services/reports/scams-prevention--and-support-programme-report---empowering-older-people-in-a-fraud-epidemic.pdf>. Accessed 30 August 2024.

- Kamuangu, Paulin. "(PDF) A Review on Financial Fraud Detection using AI and Machine Learning." *ResearchGate*, 14 February 2024,
https://www.researchgate.net/publication/378147101_A_Review_on_Financial_Fraud_Detection_using_AI_and_Machine_Learning. Accessed 31 August 2024.
- Kaufmann, Timo. "[2312.14925] A Survey of Reinforcement Learning from Human Feedback." *arXiv*, 22 December 2023, <https://arxiv.org/abs/2312.14925>. Accessed 31 8 2024.
- Lin, Xiexiong. "A Logic Aware Neural Generation Method for Explainable Data-to-text." *A Logic Aware Neural Generation Method for Explainable Data-to-text*, 2022,
<https://dl.acm.org/doi/10.1145/3534678.3539082>. Accessed 31 8 2024.
- Liu, Henry, and Lesley Fair. "Facts about fraud from the FTC – and what it means for your business." *Federal Trade Commission*, 9 February 2024,
<https://www.ftc.gov/business-guidance/blog/2024/02/facts-about-fraud-ftc-what-it-means-your-business>. Accessed 30 August 2024.
- NCSC UK. *How to spot a scam email, text message or call*,
<https://www.ncsc.gov.uk/collection/phishing-scams/spot-scams>. Accessed 30 August 2024.
- "New figures show £177.6m was lost to impersonation scams in 2022 as Take Five to Stop Fraud issues warning to the public." *UK Finance*, 17 April 2023,
<https://www.ukfinance.org.uk/news-and-insight/press-release/new-figures-show-ps1776m-was-lost-impersonation-scams-in-2022-take>. Accessed 30 August 2024.
- PWC. "Impact of Artificial Intelligence on Fraud and Scams." *PwC UK*, 2023,
<https://www.pwc.co.uk/forensic-services/assets/impact-of-ai-on-fraud-and-scams.pdf>. Accessed 30 August 2024.

Singh, Raghubir. “Edge AI: A survey.” *Edge AI: A survey*, 2023,

<https://www.sciencedirect.com/science/article/pii/S2667345223000196>. Accessed
31 8 2024.

Truong, Nguyen. “Privacy preservation in federated learning: An insightful survey from the
GDPR perspective.” *Privacy preservation in federated learning: An insightful survey
from the GDPR perspective*, 2021,

<https://www.sciencedirect.com/science/article/pii/S0167404821002261>. Accessed
31 8 2024.