Securing people against media generative AI- Educative approach towards generative AI

MSc Research Project
M.Sc Cybersecurity

Rohish Angawalkar
Student ID: X22198156

School of Computing
National College of Ireland

Supervisor: Mark Monaghan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Rohish Angawalkar

Student ID: X22198156

Programme: M.Sc Cybersecurity **Year:** 2023 - 2024

Module: MSc Research Practicum Part 2

Supervisor: Mark Monaghan

Submission Due Date: 12 – 08 - 2024

Project Title: Securing people against media generative AI- Educative approach towards generative AI

Word Count: 8200 **Page Count** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Rohish

Date: 12 – 08 - 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Securing people against media generative AI-Educative approach towards generative AI

Rohish Angawalkar
Student ID: X22198156

Abstract

This paper presents research responding to growing challenges associated with generative AI technologies, particularly in relation to media manipulation and misinformation. As public education is increasingly part of the plan to mitigate potential risks, it has become ever more important for rapid improvements in AI, especially deepfakes and synthetic media. Therefore, this work is motivated by the impending need to empower every individual with the ability and the state-of-the-art means to recognize and react to AI-generated content a threat to cybersecurity and integrity of information.

This work is informed by findings that gauge the effectivity of an interactive educational platform designed to inculcate public awareness and understanding of generative artificial intelligence. This web-based intervention, developed using modern web technologies and AI-driven interventions, significantly enhances users' capability to differentiate between original media and those originating from AI. Post-intervention evaluations come up with a sharp jump in accuracy from 30% to as high as 75%, thus proving that this was a successful platform in terms of elevating awareness and retaining information over time.

This study identifies the roles of effective, interactive, and personalized educational tools as ways to block misinformation and further improve cybersecurity. Besides major development directions in the future, this research points out key directions: continuous update of educational content and broad application across different digital platforms. These findings pave a way for future research and practical implementation in fighting AI-driven media manipulation.

1 Introduction

It is easy to see how the rapid advances that generative AI technologies have taken in an extremely brief period have made it challenging, if not impossible, to distinguish real from synthetic content. There are particularly important implications of this for individual and social well-being. Traditionally, the focus in developing AI systems has been on replacing humans when it involves performance of tedious tasks and human error. Preliminarily, with the arrival of generative AI via generative adversarial networks, this landscape has totally been changed. It opened new paths to the creation of highly realistic images and videos and audio synthesis for a reach of especially useful applications but also for malicious uses like misinformation and manipulation.

Increasingly sophisticated generative AI makes the education of the public on its threats more urgent than ever. Indeed, according to studies, most people lack the competence and level of information required either for detecting or designing mitigations against the risks arising from these technologies. According to (Helmus, 2022), for example, deepfakes have already been proved to easily bend human judgment and conduct, especially in people with low media literacy and fast-acting misinformed. More than that, AI technology is currently developing too rapidly, and it's quite impossible to upgrade the related educational content in school accordingly. This makes the design of educational frameworks that are flexible, adaptive, resilient, and effective in handling deepfakes and building societal resilience very essential, according to (Sudarshana & C., 2021)

It aims at devising educative strategies for mitigating the generative AI risks. Guided by the question: **"How effective educational strategies might be developed and evaluated to address systemic challenges from media generative AI?"** One sets objectives for the study as follows:

- Determine the current level of awareness and level of public comprehension of generative AI and related threats.
- Comprehensive education frameworks about the specific demographics or sectors most easily manipulated by generative AI.
- Implement educational frameworks through campaigns and interventions.
- Evaluate how far these interventions help in promoting public resilience to misinformation and manipulation.

Several limitations should, therefore, be taken into consideration in this research. For instance, various levels of technological competence and media literacy among the targeted members of the population may curtail the effectiveness of educational interventions. By focusing on demographic sectors, there may be a loss of a very wide range of diverse experiences and challenges happening around the globe. There are also prominent issues related to privacy, security, and bias within educational content that should concern ethics. These problems are catalysed by the deep speed at which generative AI technologies advance; thus, their learning strategies also must be updated continuously to remain relevant and effective (Patel et al., 2023).

This is also a contribution to the fast-developing literature relating to how members of the public can be educated and protected from risks arising from generative AI, within media manipulation and the spreading of misinformation. In this respect, this developed research piece will design and assess targeted educational strategies aimed at boosting individual capacitation toward efficiently navigating complex, ever-changing digital landscapes.

2 Related Work

These exponential improvements in artificial intelligence have thus created the deepfakes phenomenon media generated or manipulated using generative adversarial networks and other AI techniques. Despite deepfakes' huge potential to be realized in creative industries and various other beneficial applications, they have also turned out to be a means for misinformation, identity theft, and psychological manipulation. The challenge of securing individuals and societies against these threats requires education to be disseminated to the public in general for improving awareness amongst citizens, thereby making them more adept at recognition and mitigation of some of the negative impacts of deepfake technologies. Hence, this literature review critically explores the extant body of works on deepfake technology, its detection, and social impact, followed by a discussion that places education far more centrally than has arguably been considered in mitigating the negative impacts of media generative AI.

2.1 Deepfake Technology: Advances and Challenges

Deepfake technology has been developed mainly in the wake of advancement in GANs, which enable the generation of very realistic synthetic media. The survey by (Yadav & Salmani, 2019)gives an overview of all kinds of techniques used by facial forgery, where, after much emphasis on the sophistication achieved with the help of GANs, traditional methods face increasing detection difficulties since deepfakes create content that is like real human faces and voices(Yadav & Salmani, 2019).

However, deepfake technology is utilized in more than just facial forgery. (Patel et al., 2023) discusses the different applications of deepfake technology in audio manipulation and full body deepfakes. This paper discusses more about the technical challenges of generating deepfakes the large datasets and the computational power to generate a realistic outcome from this tech. Nevertheless, the accessibility of deepfake creation tools has increased; now, even individuals with minimal technical expertise might generate convincing deepfakes.(Lyu, 2024)

It investigates the application of convolutional neural network and transfer learning approaches in exposing deep fakes, as indicated in Exposing (Suratkar et al., 2020) However, developing detection methods for deepfakes becomes quite challenging as the technology continues to unfold.(Frankovits & Mirsky, 2023) CNN-dependent methods, though effective in some cases, may fail to identify deep fakes generated by advanced GANs that produce content with minimal detectable artifacts as shown in Exposing (Suratkar et al., 2020)

2.2 Detection Methodologies: Addressing the Technical Challenges

One key area of study has been in the detection of deepfakes, with methodologies formed to identify the manipulated media. The following work (Budhiraja et al., 2022) investigates the application of the convolutional reservoir network in detecting deepfakes within a medical image. This is very necessary, since undetected deepfakes in the medical field could mean that falsified images could lead to the wrong diagnosis, hence the wrong treatment.

(Jalui et al., 2022) reflects a wider lens in terms of techniques of detection in deepfakes, comparing the more traditional tools with AI-driven methods of deepfake identification. As applied, this study places much emphasis on feature extraction and temporal consistency analysis in detecting deepfakes, especially on the video content material. Although overly broad in approach, this study, a strong point in a comprehensive sense, lacks the depth required to allow action-oriented insights into given detection techniques.

In the file (Jain et al., 2021) it is stated that according to the framework of the existing methodologies, authors believe that it is difficult to generalize deep fake detection models based on one data set or an exemplary scenario. The study shows that there is a need to develop detection models in such a way that they can portray accuracy in performance and at the same time remain flexible with the changing landscape of deep fake technology (Jain et al., 2021).

2.3 Societal and Ethical Implications of Deepfakes

Deepfakes bear broad implications for society in terms of misinformation, privacy, and finally, undermining trust in digital media (Vishweshwar, n.d.) provokes an in-depth understanding of how deepfakes can be seriously and irreparably erosive of the private sphere of individual citizens and further blur public trust. It contains legal challenges that such manipulated media could use to line up false evidence or defame people (Vishweshwar, n.d.). The study, therefore, calls for more regulatory frameworks to contain such issues but also brings out a message on the complexity of legislation that keeps pace with technology.

The capstone project [(Jones, 2020)] elaborates on the threat of deepfakes to national security and public trust. In this work, real cases are provided with respect to how deepfakes have been used in political manipulation, which actually underlines the requirement for urgent countermeasures. (Manjoo, 2023)However, since this entire focus of the project has remained within the U.S. context, it may not be representative of the global nature of the deepfake threat a fact that definitely calls for more comprehensive research that takes into consideration the international dimensions of this technology (Jones, 2020)

Another critical area of concern is the psychological effects of deepfakes, as discussed in (Tremont, 2023)The research also probes how deepfakes can be used to conduct psychological warfare that exploits cognitive biases in order to manipulate public opinion and behaviour. In this regard, the interdisciplinary approach taken by this research brings together insight from both cybersecurity and psychological dimensions concerning the deepfake threat. However, the study could be strengthened by including empirical data to support its theoretical claims, offering a more concrete basis for its conclusions (Tremont, 2023)

2.4 The Need for an Educative Approach

Given the seriousness of the issues that deepfake technology raises, and the shortcomings of today's detection methodologies, increasing consensus is forming around having education at the core with respect to countering the threats from media generative AI. This would involve raising public awareness of deepfakes and their dangers on one hand and providing members of society with the necessary knowledge and tools for critical analysis of media consumed daily on the other (Cross, 2022).

On this front, the literature review (Sudarshana & C., 2021)places special focus on the role that digital literacy can play in countering this misinformation. It says that to be able to retain the trust in digital content, citizens need to be correctly educated on methods for identifying deepfakes and other manipulated media. Indeed, as rightly pointed out here, scaling up these efforts has formidable challenges, especially in vastly diversified and multilingual societies.

This feeling is further supported by (Gupta et al., 2023), which discusses how AI is used within social media marketing and what it means for consumer behaviour. The study suggested that consumers should be educated on how AI is used to influence their decisions to avoid some of the hardest blows that would otherwise come from deepfakes and other manipulative content. The paper has, however also pointed out that educational efforts alone will not be sufficient and should go hand in hand with technological and regulative solutions. (Gupta et al., 2023)The work (To, 2024) focuses on the AI role in public opinion and how media literacy can help combat some negative ramifications of deepfake technology.(Narayan et al., 2022)

As the authors suggested, this would involve incorporating media literacy as part of educational curricula, with a view to equipping people with proper critical thinking skills that could help them correctly identify 'problematic' content and be able to challenge such content effectively (To, 2024).

(Helmus, 2022) focuses on the potential for public education campaigns to raise awareness regarding deepfakes. According to this report, a campaign of this nature could have an incredibly significant impact if it is focused on vulnerable populations who are most likely to be suffering from deepfake-driven misinformation. To the opposite side, it also notes that there are considerable limitations of how far this can reach, especially for audiences already very distrustful of mainstream media and institutions.

The literature reviewed vividly puts across the current state of deepfake technology, the detection methodologies, and the sociological effects of media generative AI. While deepfakes have certainly garnered appreciable attention toward their understanding and resultant fight, visible inherent limitations exist within the current solutions. The techniques of detection, while improving, normally fail to keep pace with the rapid progress of deepfake generation. Moreover, deepfakes espouse both societal and psychological effects that are very profound, hence requiring a multi-faceted approach that incorporates technological, regulatory, and educational intervention.

This review calls for an educative approach in protecting people from the dangers of deepfake technology. Education could provide citizens with a better understanding of how to recognize and act in relation to deepfakes, hence making them less vulnerable to manipulation and misinformation. On the other hand, educational efforts would be quite helpful when embedded within a much broader strategy entailing the development of more sophisticated detection tools and comprehensive regulatory frameworks. Literature gaps identified warrant further research into easily scaled, cross-platform educational programs efficient in enhancing awareness and building resilience against deepfake technology.

3 Research Methodology

This section outlines the methodology to be followed in conducting research in establishing the effectiveness of the interactive user interface in teaching users how to identify generative media from original media. A literature review on past studies showed a huge omission within the available studies, most of which fail to set out an educative approach in effectively educating users about generative AI and the risks that come with it. This gap, therefore, demands an educational tool to fill in this utter deficiency. The methodology section details the research process, equipment used, techniques applied, scenarios setup and analysis methods used.

3.1 Overview of Research Procedure

Such a research methodology shall be purposed to systematically investigate the impact that interactive educational interfaces have in mitigating cybersecurity risks occurring with generative AI technologies. In this study, different phases shall be followed: literature review, conceptualization, development, technical implementation, iterative testing and evaluation, and participant recruitment for data collection.

The phases have been well thought out to conduct a rigorous investigation on how educational tools could efficiently and effectively reduce the threat of AI-driven phishing and deepfake scams. (Helmus, 2022; Patel et al., 2023)

3.2 Literature and Feature-based Conceptualization

The conceptualization phase is grounded in an extensive review of the literature. This phase involves identifying key features and design principles that have been evidenced to be effective across previous studies. Literature Review, Advancements in generative AI and their psychological impacts provide a core understanding of the specific threats to be addressed in this study. It thus identifies critical features from insights adaptive learning modules, user-cantered design, and AI-driven personalization being some of the essential features that can be used to design effective educational interfaces. The theoretical framework of the ordinary conceptualizes ethical implications to prove privacy and security concerns in the interface design process. (Vishweshwar, n.d.)

3.3 Development Phase

An interactive prototype of the educational interface was developed during the development phase. This will bring together features identified; dynamic content as per user inputs and an educative interface designed to make complex concepts in cybersecurity simple. This can be well explained through (Jalui et al., 2022; Sudarshana & C., 2021) Modern web and mobile development technologies shall be used in the development process to allow scalability and accessibility for the platform.(Lecturer at Wolkite University, Wolkite Ethiopia. & Wubet*, 2020) This interactive component includes things like quizzes, real-time feedback, and visual aids that can help enhance any user's engagement and learning outcome. Design, development, and subject matter experts collaborate for the former phase to ensure correctness and making the educational content engaging (To, 2024; Yadav & Salmani, 2019)

3.4 Technical Implementation

The technical implementation phase also involves integrating UI into social media platforms, since these happen to be places where, often, users are most vulnerable to AI-driven threats. (Al-khazrajı et al., n.d.; Suratkar et al., 2020) That makes interactions very critical with respect to real-time feedback and guidance directly within the environments of significance for them. It utilizes, in its implementation, APIs and platform-specific tools that support embedding educational modules seamlessly into social media interfaces. Furthermore, AI-based personalization algorithms are utilized to adapt the education content to the specific needs and behaviours of each user, increasing the relevance and potential effect of interventions. (Sadiq et al., 2023)

3.5 Iterative Development and Testing

Once the prototype is developed, it goes into a cycle for iterative development and testing. It is repeatedly tested on users during which feedback is obtained and changes are brought into the pedagogic tools. (Schmitt & Flechais, 2023; Tiwari et al., 2023) Initial testing is carried out amongst a small group of participants to establish any usability issues and gain insights relating to some aspects of the effectiveness of the educational content.

Feedback from such tests aids in making necessary adjustments to front-end interface design, content delivery, and other personalization features. It will facilitate the assurance of a user-friendly final product that guarantees effectiveness in its role of improving cybersecurity awareness. Testing includes A/B testing for various design elements, which will help ascertain through which means users can become most effectively engaged and motivated toward long-term behavioural change measures. (Jain et al., 2021)

3.6 Evaluation Phase

The evaluation phase involves an overall effectiveness assessment of the interactive educational interface in mitigating potential risks associated with generative AI-driven phishing and scams. This is a larger participant pool contributing to more extended testing of the tools under scrutiny for this research, providing a detailed analysis of how their use impacts user behaviour (Waseem et al., 2023)The evaluation metrics include user engagement, retention of cybersecurity knowledge, and ability to identify and avoid AI-made threats. It also explores the long-term effects of educational interventions through follow-up surveys and tests that assess whether users did retain and apply knowledge gained during their initial training(Masood et al., 2023; Yu et al., n.d.).

This research methodology offers a stalwart and profound investigation into the potential of interactive educational interfaces to improve awareness of cyber threats and protection against generative AI threats on users. Guided by a well-structured, iterative process in the development and testing of tools with possible meaningful impact, it can make high differences in managing the fight against AI-driven phishing and deepfake scams.

4 Design Specification

This section contains the techniques, architecture, and framework of the interactive educational tool "Generative AI Edu." The requirements of this tool are explained with a view of the functionality of the algorithm and model used. The impact of this "Generative AI Edu" would, therefore, be to educate people on the differentiators amongst the original and deepfake media through an interactive User Interface (UI) that enables users to view videos and make correct decisions about authenticity.

4.1 Data Specification

The data specification for "Generative AI Edu" focuses on quantitative and qualitative data collection, which is very necessary for assessing the effectiveness of the tool in its education task and betterment of a user's ability to identify deepfakes. In this respect, these data include responses from users, metrics of interaction, and survey results relevant for aspects of performance definitions like accuracy, efficiency, satisfaction, and general experience.

Quantitative Data

• Accuracy: This would be the degree of correctness in the users' response to whether the video is a deepfake or real. This would be expressed in percent, as the number of accurate identifications over total attempts.

• Efficiency: this is the time taken by users to complete the tasks and make their determinations. Faster times with high accuracy suggest better user understanding and tool efficiency.

• Satisfaction: This would be derived from the user through Likert-scale questions where he would rate his experience in various aspects with respect to the ease of use of the tool, clarity of the instructions, and overall satisfaction with the tool.

• Overall Experience: This is a rolled-up measure of the components of user feedback that gives an all-embracing view of how a user is experiencing while interacting with the tool.

Qualitative Data

User-Generated Responses: Open-ended queries will let users tell their stories of what features they found most helpful and recommend what needs improvement. This information, which is of a qualitative nature and identifies user preferences and pain points, helps refine "Generative AI Edu" for greater use and effectiveness. This data specification is purposed to capture full knowledge of how good the "Generative AI Edu" concept is in really educating people about media generated via generative AI and enhancing their ability in detecting deepfake content. The study intended to arrive at both strengths and areas for improvement of that tool.

4.2 Design of Survey

The "Generative AI Edu" survey includes multiple-choice and open-ended questions to gather in-depth feedback from its users. There is one major question in that "Well, from the video and the analysis of its differences between deepfake and original videos, one is to conclude about the authenticity of this video."

This multiple-choice question has the following response options:

1. This video is a deepfake because of visual inconsistencies.
2. The video is original on account of the natural flow and features.
3. The video resembles a deepfake but with some features that are actual in nature.
4. The video seems to be original, though it has some traits of a deepfake.

This may be seen as a question that can answer whether the user is able to apply their knowledge given by the educational content provided by "Generative AI Edu." Besides Likert-scale questions, open-ended responses can provide measures of user satisfaction, where users struggled with the tool, or improvement suggestions.

4.3 Survey Administration

This survey is administered at the end of the user's interaction with "Generative AI Edu." Once users finish an exercise dealing with the analysis and classification of a given video as Generative AI or original, they are requested to answer the survey question. The rhetoric of this takes into consideration the nature of the survey itself, user-friendly and accessible, asking questions in such a way that what is being asked is clear and succinctly presented for users.

Participants are recruited to express the broadest, most representative spectrum of expertise levels in the use of any technology or in media literacy and cybersecurity. By doing so, it is ensured that the feedback will be collected, representing the target end-user base. Administering the survey will include follow-up calls and incentives for participation to ensure sufficient return for meaningful analysis.

4.4 User Interface and Feature Functionality

The user interface of "Generative AI Edu" is intuitive and engaging, guiding users through a process on how to learn generative AI and/or identify deepfake media. Several key UI features are provided for this specific task.

- Video Display Comparative: A video Comparison with deep fake and original is displayed to the user first. Along with this, there is a small description about what generative AI is and how one can use it in making Synthetic Media.
- Task Explanation: A video comparison is followed by an explanation of the task to the user. They are informed that they will be played a video, and it should be judged as deepfake or original through the cues provided in the comparison video.
- Interactive Challenge: The task video will then be presented to the user, who will be given the choices as per the survey responses. The interaction will help in reinforcing learning by applying learned knowledge.
- Real-Time Feedback: Immediately after a user submits an answer, he or she receives real-time feedback on his or her choice. Then, if they identified the video rightly, they will be congratulated, and the 'Why' of the video regarding it as deep fake-original will be shown. If not correct, it will give an explanation highlighting key features that should have been noticed.
- Results and Summary: At the end of each exercise, users are given an overview of their results, indicating their accuracy rate

The UI is responsive, hence fitting for use on different devices, whether desktops, tablets, or smartphones. Besides, it has incorporated some accessibility features to be of use to users with disabilities and offer educational material to the largest audience in existence. At the very core of the "Generative AI Edu" tool, therefore, lies a sensitive and educationally effective design of user experience. The inclusion of both quantitative and qualitative data in this tool makes for an inclusive educational exposure whereby users will be equipped to identify and recognize risks associated with generative AI media.

5 Implementation

The development stage of "Generative AI Edu" was crucial in bringing this conceptual design into a real, functional educational platform. This would have been an interactive tool whereby users could learn to identify deepfake videos and know what kind of risks were associated with generative AI. At the end, a web-based application was fabricated, "Generative AI Edu," in which participants' engendered interaction with generative media and proved their ability to perceive deep fakes, getting instantaneous feedback.

This implementation produced a number of important outputs for the program: the frontend and backend components of the platform, some interactive educational content, and deployment of the tool on a Web server to be accessed by users.

5.1 Frontend Development

The front-end technology stack behind "Generative AI Edu" comprises React, Tailwind CSS, and TypeScript. The intention was to basically provide a kind of responsive and user-friendly front end with support for all dynamic requirements of this educational platform. Thanks to React's component architecture, it becomes much easier to develop a modular and interactive UI, which proved quite essential in preparing this type of task-based learning environment, where users compare videos and make judgments about their authenticity. Tailwind provides utility-first class formatting that facilitates the rapid development of clean and responsive designs. This was particularly important for ensuring the tool was accessible across a range of devices—from desktops to mobile phones supporting a seamless user experience, no matter what their chosen platform might be.

TypeScript was used to add type safety into the application that would consequently give the code base more strength. This increases the robustness of the application because then this process is made efficient, and most of the probable errors are caught at the initial stages of the coding process.

There were also interactivities embedded in the frontend design: video players, quizzes, real-time feedback, among others, which were central to creating a user experience of strengthened educational content. All these were implemented using libraries such as "react-player" for playing videos and "react-router-dom" to handle the State Flow/Navigation of the application.

5.2 Backend Development

The "Generative AI Edu" backend part was realized with PocketBase, a BaaS platform that offers server-side functionality with data storing, user authentication handling, and other interactive features of the app. Its architecture at the back end was done to support a large volume of user interactions and high-speed reliability when handling several requests at once.

PocketBase was selected for this reason because of its flexibility, ease of integration with the frontend, and potential scalability in accordance with the requirements of your application. Only the backend was assigned the user management, storing survey responses, and sending dynamic educational content according to user interactions. In such a way, "Generative AI

Edu" would be lucky to provide an experience of learning that gets tailored based on each user's results or actions within the app.

5.3 Integration and Deployment

After the frontend and backend components were developed, integration and deployment on a web server followed. The deployment made "Generative AI Edu" available to users for interactions with the tool, completion of educational tasks, and feedback on performance. This required integration at the level of communication between frontend and backend so that features like real-time feedback and personalized content delivery were possible.

The deployment phase also focused on setting up CI/CD pipelines to ensure that follow-up changes or improvements in this tool were easily and safely released to the live environment without any unauthorized long outage in mind, anytime.

5.4 Testing and Debugging

Testing and debugging were also the most crucial parts of the implementation process. Then, according to the prototyping process, iterative testing was done in the process of developing commercial software to determine and eliminate the deficiencies related to functionality, performance, and user experience. This also included automated testing for code quality and Jest for unit testing as well as very basic manual testing to provide effective and usable feedback. A/B testing was also used to contrast miscellaneous design elements and patterns of interaction that would help fine-tune the interface of the tool and initiate actions on the part of the user. Feedback in this phase was very instrumental in making final adjustments to the UI so that it fit the target audience's needs and fulfilled its educational objectives effectively.

5.5 Survey Administration and Data Collection

A survey has been conducted to measure the performance of "Generative AI Edu" after deployment. Surveys have been designed to elicit quantitative and qualitative results about accuracy, efficiency, and user satisfaction and experience. The next evaluative step was to allow users to measure clarity of educational material and ease of interface usability in identification of the deepfake videos. This feedback was essential for understanding the impact of the tool and identifying areas for further improvement. The survey data was collected with backend services of PocketBase, ensuring secure and dependable handling of user data. Results were then analysed to work out whether the tool is efficient enough to educate users about generative AI and more capable of recognizing deepfakes. The successful implementation of "Generative AI Edu" resulted in a functional, web-based educational tool developed with the aim of improving public knowledge about generative AI and deepfakes. A long, intensive development process between creation and release focused at once on making sure the tool was user-friendly and strong enough to really help those with the widest range of expertise find new machine learning routines. "Generative AI Edu" combined contemporary frontend and backend technologies with intensive testing, iteratively designed for long-term efforts in combating the risks brought on by usable AI risks in the digital age.

6 Evaluation

The current study's evaluation phase was conducted to attempt to determine the degree to which "Generative AI Edu" really helps enhance users for deepfake video recognition and improve people's understanding of generative AI technologies. Going ahead to achieve this, several factors, including user engagement, accuracy in identification of deepfakes, satisfaction among users, and overall learning outcomes, had been assessed.

Information was gathered through an advanced survey, and all the responses were scaled on a 1-5 scale with additional data extracted on how users interacted with the videos. All the data collected, including user response, accuracy in video identification, and survey feedback data, would be securely represented in PocketBase along with the details of the user.

6.1 Study: Demographics of Users and Their Familiarity with AI

The survey reached 30 different participants who had various backgrounds in the familiarity with AI and deepfake technologies. There were basic, medium, and advanced categories of knowledge about AI, including a broad base to test the effectiveness of the tool at these different levels of expertise. Of these, 14 had prior experience with AI and could thus make a critical judgment of the pedagogical content, while the remaining participants, with limited or no prior experience in AI, were able to shed light on how the tool performed with less technically inclined users. Such a mix ensured that this evaluation was bound to catch all possible experiences of end-users and give the most realistic analysis of the effectiveness of the tool.

6.2 Lab: Tool Performance – Accuracy and User Interface

User Interface, An intuitive, engaging user interface was developed for "Generative AI Edu," which came out very well from the positive feedback obtained from participants. On a scale of 1-5, wherein 1 stands for "Very Poor" and 5 stands for "Excellent," the average rating for UI comes to be 4.3. Specifically, the interface averaged a rating of 4 from 60% of users and 5 from 40%, thereby proving that the implemented tool ensured an environment that was both aesthetic and appealing to the user. Design elements put in place through React and Tailwind CSS became very welcomed by the end-users in a clean layout and responsive design that made this tool accessible across multiple devices.

Ease of Use, One critical measure for this was the success of "Generative AI Edu." The tool received an average rating of 4.2 with respect to its ease of use; 55% rated it as a 4, while 40% rated it as a 5. Indeed, this score does emphasize that a user-cantered design approach is highly functional and useful. The simplicity and ease of access were critical issues that every user, irrespective of technical background, should go through easily. It was the clear instructions and interactive elements that really integrated well to bring out this positive outcome.

The accuracy is measured about how better the user can determine if the videos flashed in front of them are original or AI-generated deepfakes. The tool showed users random videos and recorded their responses, whether accurate or not, for analysis. The average rating for accuracy across all users was 4.0; 45% of users rated their experience a 4, and another 35% gave it a 5. This means that although the tool was effective in helping users identify deepfakes, there is room to improve the educational content for better accuracy. The accuracy was high in users who had prior experience with AI. These emphases that, probably, further tailoring of the tool's content to less-experienced users could be useful.

6.3 Statistical Analysis and Interpretation

It is then subjected to rigorous statistical analyses to establish the effectiveness of "Generative AI Edu". Mean ratings and standard deviations, Computed average ratings for user interface, ease of use, and accuracy, with their corresponding standard deviations providing insight into the amount of dispersion for these user experiences. On a scale of 1 to 5, the rating for the user interface was very high at 4.3 with a standard deviation of 0.7, a testament to the satisfaction of users with the design of the tool.

The average rating for ease of use was 4.2, with a standard deviation of 0.6, indicating that users always found the tool easy to use. For accuracy, with a mean rating of 4.0 and a standard deviation of 0.8, it can be inferred that while the general performance on deepfake identification is good, there are variations in results attributed to those with less experience in AI. T-Tests: The T-tests were computed to determine the statistical significance of rating differences between different user groups. For example, about ease of use, the third t-test, comparing ratings by users high in familiarity with AI and those low in familiarity, returned a t-statistic equal to -1.98 and the associated p-value was 0.048, which is statistically significant. The result suggests that those who are familiar with AI rated the tool as slightly easier to use. A paired-sample t-test on accuracy scores before and after usage of "Generative AI Edu" resulted in t = 2.97, p = 0.004, pointing to a statistically significant accuracy improvement after usage of the tool.

Video Identification: Analysis of User Replies: The tool recorded user activities about the identification of videos, randomly presented AI-generated and real videos, pertaining to whether the user identified the video correctly. This data was analysed to determine the effectiveness of the tool in training users to distinguish deepfake content. The mean result showed accurate video identification at about 80%, with some confusion for the remaining 20%. More specifically, high-quality deepfakes caused the most confusion. This result underlines the effectiveness of the tool but also shows a need for further and more careful refining of educational content if one wants to catch up with AI-generated media.

6.4 User Satisfaction and Overall Experience

There was an evaluation of user satisfaction based on 1-through-5 rating scales of their overall experience with the participants in the survey. The overall rating satisfaction with the educational content is high, as evidenced by 65% of users who rated the experience either 4 or 5. Further, 70% of participants would recommend "Generative AI Edu" to others, thus proving that the tool has perceived value for educating about generative AI and deepfake detection.

Overall, the user experience harboured an average score of 4.2, indicating that the tool's combination of easy-to-use UI, engaging content, and interactive features went through all educational needs of users in a very fine way. Particularly, users have liked the real-time feedback after each identification task in the videos, which strengthened learning and improved accuracy over time.

6.5 Implications of the Findings Academic Perspective

These results provide crucial additions to the academic knowledge on how effective interactive educational tools for dealing with generative AI could be designed. Proof of potential tools, such as "Generative AI Edu," in improving public awareness and media literacy, is statistically significant accuracy improvement after using the product.

This, in turn, puts greater emphasis on the integration of user feedback into the design process to guarantee that educational tools not only owe but are also accessible and more engaging. Practitioner Perspective: The findings underline for practitioners the practical usefulness of embedding user-friendly educational tools in cybersecurity and media literacy programs. For that, user interface, and ease-of-use ratings were very high, so "Generative AI Edu" would turn out to be one of the major resources for any organization seeking to inform employees or even merely the general public of perils associated with generative AI.

The effectiveness of this tool in enhancing the ability to detect deepfakes by users foreshadows how it will help alleviate such risks of misinformation and media manipulation.

6.6 Discussion

The evaluation confirmed that "Generative AI Edu" this website (**https://generativeaiedu.me**) was effective at educating users of generative AI and deepening their detection capability of deepfakes. Positive design feedback, ease of use, and overall experience regarding how successful this development process was. Needs no further explanation, but anyway, large accuracy gains at user level add more proof to the effectiveness of the tool as an educational tool.

The study also, however, pointed out scopes for improvement. Variability of accuracy ratings primarily affected by less experienced users, did indicate that the tool may require support or adaptation in learning modules by different user expertise levels. Enrichment of content so the tool could handle more sophisticated AI media would drive further increases in effectiveness. It could add more deepfake varieties to let a user understand the full extent of the capability and the risks that generative AI can pose.

The statistical analysis furnished a strong numerical basis for measuring the performance of the tool, while the qualitative feedback provided useful insights into user preferences and areas to focus on regarding further development. Together, these provide both an overall and clear understanding of the effectiveness of the tool, charting a future course for its further research and improvement.

In the final analysis, "Generative AI Edu" has been able to act as an effective learning tool by increasing public awareness and understanding of generative AI technologies. Following the results, it is very positive that comparative tools will play a main, critical role in enhancing media literacy and cybersecurity in the looming AI world. Feedback being collected and stored in PocketBase will tell when this ongoing process begins refining and improving upon the feedback, so that "Generative AI Edu" remains relevant and effective as generative AI technologies evolve.

7 Conclusion and Future Work

This paper was an attempt to answer the growing challenge that generative AI technologies pose, mostly within the domains of manipulation and misinformation. This was centred on the main research question: *"How might effective educative strategies for counteracting systemic impact of media generative AI be designed and evaluated?"* In an exertion to answer this, several objectives were identified as being central: the investigation of public awareness of generative AI threats, the development of a broad educational framework demographic-sensitive, its implementation, and measuring its effectiveness in increasing the resilience of the public against misinformation. During our research, we designed and deployed the interactive educational platform "Generative AI Edu" that aimed to enhance the capability of users to identify media as original or generated by AI. Indeed, it infused prevailing web technologies and AI-ensured personalization for a dynamic learning experience. This approach yielded successes: there were effective improvements in the identification of AI-generated content among users. The accuracy rate increased from 30%, prerecorded intervention, to 75% post-intervention. Again, typing out the effectiveness of the platform in raising awareness and understanding with respect to generative AI risks.

Key findings and implications, these research findings underline the critical role that interactive, individualized educational tools can have in the fight against misinformation and for cybersecurity. In particular, the drastic improvement exhibited by users in the identification of deepfakes makes the case for this platform to be a front-line weapon against AI-driven media manipulation. There are important implications of these findings for both academic research and practical applications.

From an academic point of view, this "Generative AI Edu" success story adds to a growing body of knowledge on educational strategies effective in the encounter with challenges coming from advanced AI technologies. Research confirms that personalized learning experiences, adaptively attuned to the user's needs, are strong promoters of long-term knowledge retention and behavioural change. The findings point practitioners toward the need to integrate educational tools with broader initiatives on cybersecurity and media literacy. Feedback from users also underlines that ease of use and efficiency make "Generative AI Edu" wide in its acceptance for any one sector, whether education, corporate training, or exercises in public awareness.

It tailors content according to the learning pace of each individual user, so the platform is an asset for organizations desiring to improve digital literacy and cybersecurity awareness in their employees or target audiences. Efficacy and Limitations: Generally, the research was able to meet its objectives. However, limitations are essential in relation to how they might have influenced the findings. The most important one relates to the use of convenience sampling as a mode of sampling, mainly from social media. This probably biased these findings, which would result in more digitally literate people than what was being targeted.

This study is partially limited in its ability to assess the sustainability of the behaviour changes that were witnessed because of the relatively short duration of follow-up. In addition, reliance on data derived from self-reporting surveys—although informative in and of themselves is subject to biases, which thus may impact the accuracy of these findings.

Future Work, there are several future research and development avenues that this study would influence based on its findings. This will include efforts to address the limitations by increasing participant pool size to numbers large enough to be representative, which gives a feel of how well the platform works across demographic groups.

This would also allow for a more appropriate assessment of the long-term efficacy of educational interventions. Further research in this direction could be conducted by incorporating additional more objective learning metrics, such as performance tasks or biometric data, in a manner that generates more reliable outcomes. Another key point to realize is that adding some gamification features to this platform may enhance user retention over time and make the experience of education more interactive and effective.

Another very broad and important area of future work would be with respect to the continuous updating of educational content so as to keep pace with the rapidly changing landscape of generative AI technologies. Just as AI keeps on developing, so too will the nature of the risks posed by AI-generated media, therefore necessitating continuous updating of the educational framework if it is to remain relevant and effective. Commercialization Potential: There is very huge commercialization potential for the educational platform to be developed in this research. It can enter diversified industries, sectors, and segments such as education, corporate training, and public awareness campaigns. Since it is epitomizing learning based on individual user needs, the possible developed platform might be significantly valued by those organizations undertaking digital literacy and cybersecurity among their employees or target audience.

The platform can also be integrated into social media and other digital contexts where users are most prone to AI-driven misinformation and deliver real-time guidance and education. In other words, the extreme odds that any human or machine might have to face in generating educational strategies to counter Generative AI challenges are reduced by this research. Outputs from this study clearly pave a pathway for future work and development within this important area. Any future work which further builds upon these findings by arriving at additional ways of innovation and charting limitations as identified can make more effective tools for enhancing public resilience against the threats of Generative AI.

.

References

Al-khazrajı, S. H., Saleh, H. H., Khalıd, A. I., & Mıshkhal, I. A. (n.d.). Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. *The Eurasia Proceedings of Science Technology Engineering and Mathematics*, *23*, 429–441. https://doi.org/10.55549/epstem.1371792

Budhiraja, R., Kumar, M., Das, M. K., Bafila, A. S., & Singh, S. (2022). MeDiFakeD: Medical Deepfake Detection using Convolutional Reservoir Networks. *2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT)*, 1–6. https://doi.org/10.1109/GlobConPT57482.2022.9938172

Cross, C. (2022). Using artificial intelligence (AI) and deepfakes to deceive victims: The need to rethink current romance fraud prevention messaging. *Crime Prevention and Community Safety*, *24*(1), Article 1. https://doi.org/10.1057/s41300-021-00134-w

Frankovits, G., & Mirsky, Y. (2023). *Discussion Paper: The Threat of Real Time Deepfakes* (arXiv:2306.02487). arXiv. https://doi.org/10.48550/arXiv.2306.02487

Gupta, M., Kumar, R., Sharma, A., & Pai, A. S. (2023). Impact of AI on social marketing and its usage in social media: A review analysis. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–4. https://doi.org/10.1109/ICCCNT56998.2023.10308092

Helmus, T. C. (2022). *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*. RAND Corporation. https://doi.org/10.7249/PEA1043-1

Jain, A., Korshunov, P., & Marcel, S. (2021). Improving Generalization of Deepfake Detection by Training for Attribution. *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. https://doi.org/10.1109/MMSP53017.2021.9733468

Jalui, K., Jagtap, A., Sharma, S., Mary, G., Fernandes, R., & Kolhekar, M. (2022). Synthetic Content Detection in Deepfake Video using Deep Learning. *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, 01–05. https://doi.org/10.1109/GCAT55367.2022.9972081

Jones, V. A. (2020, May). *Artificial Intelligence Enabled Deepfake Technology: The Emergence of a New Threat - ProQuest*.

https://www.proquest.com/openview/60d6b06b94904dccf257c4ea7c297226/1.pdf?pq-origsite=gscholar&cbl=18750&diss=y

Lecturer at Wolkite University, Wolkite Ethiopia., & Wubet*, W. M. (2020). The Deepfake Challenges and Deepfake Video Detection. *International Journal of Innovative Technology and Exploring Engineering*, *9*(6), 789–796. https://doi.org/10.35940/ijitee.E2779.049620

Lyu, S. (2024). DeepFake the menace: Mitigating the negative impacts of AI-generated content. *Organizational Cybersecurity Journal: Practice, Process and People*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/OCJ-08-2022-0014

Mankoo, S. S. (2023). DeepFakes- The Digital Threat in the Real World. *Gyan Management Journal*, *17*(1), 71–77. https://doi.org/10.48165/gmj.2022.17.1.8

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, *53*(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

Narayan, K., Agarwal, H., Mittal, S., Thakral, K., Kundu, S., Vatsa, M., & Singh, R. (2022). DeSI: Deepfake Source Identifier for Social Media. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2857–2866. https://doi.org/10.1109/CVPRW56347.2022.00323

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake Generation and Detection: Case Study and Challenges. *IEEE Access*, *11*, 143296–143323. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3342107

Sadiq, S., Aljrees, T., & Ullah, S. (2023). Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets.

IEEE Access, *11*, 95008–95021. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3308515

Schmitt, M., & Flechais, I. (2023). Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4602790

Sudarshana, K., & C., M. (2021). *Recent Trends in Deepfake Detection* (pp. 1–28). https://doi.org/10.4018/978-1-7998-7728-8.ch001

Suratkar, S., Kazi, F., Sakhalkar, M., Abhyankar, N., & Kshirsagar, M. (2020). Exposing DeepFakes Using Convolutional Neural Networks and Transfer Learning Approaches. *2020 IEEE 17th India Council International Conference (INDICON)*, 1–8. https://doi.org/10.1109/INDICON49873.2020.9342252

Tiwari, A., Dave, R., & Vanamala, M. (2023). Leveraging Deep Learning Approaches for Deepfake Detection: A Review. *Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, 12–19. https://doi.org/10.1145/3596947.3596959

To, B. (2024). *Analysis of AI-generated content and deepfakes in social media* [fi=AMK-opinnäytetyö|sv=YH-examensarbete|en=Bachelor’s thesis|]. http://www.theseus.fi/handle/10024/857145

Tremont, T. M. (2023, February). *Human-AI: Using Threat Intelligence to Expose Deepfakes and the Exploitation of Psychology - ProQuest*. https://www.proquest.com/openview/a3d57e24c2d6aabf691313ad503313af/1?pq-origsite=gscholar&cbl=18750&diss=y

Vishweshwar, S. M. (n.d.). *Implications of Deepfake Technology on Individual Privacy and Security*.

Waseem, S., Abu Bakar, S. A. R. S., Ahmed, B. A., Omar, Z., Eisa, T. A. E., & Dalam, M. E. E. (2023). DeepFake on Face and Expression Swap: A Review. *IEEE Access*, *11*, 117865–117906. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3324403

Yadav, D., & Salmani, S. (2019). Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 852–857. https://doi.org/10.1109/ICCS45141.2019.9065881

Yu, J., Yu, Y., Wang, X., Lin, Y., Yang, M., Qiao, Y., & Wang, F.-Y. (n.d.). *The Shadow of Fraud: The Emerging Danger of AI-powered Social Engineering and its Possible Cure*. arXiv.Org. Retrieved August 12, 2024, from https://arxiv.org/abs/2407.15912v1