DeepShield: The AI Solution Integrated with Government to Combat the Ethical Challenges of Pornographic Deepfakes

AI Product Proposal

FIT1055 IT Professional Practice & Ethics | 4th September 2024

Written by:

Chris Law Zi Qing | 34112804
Tan Choong Sheng | 34886176
Chew Chen Hin | 35154667
Lim Hui Zern | 34886257
Bradley Hoh Lok Yew | 34886362

Table of Contents

ABBREVIATION LIST

FULL FORM	SHORT FORM
Artificial Intelligence	AI
United Kingdom	UK
Generative Adversarial Network	GAN
Convolutional Neural Networks	CNN
Recurrent Neural Networks	RNN

LIST OF FIGURES

LIST OF TABLES

1.0 Introduction

In recent years, Artificial Intelligence (AI) has advanced rapidly and is now widely used in a wide range of areas in modern society. One prominent application of this technology is deepfake creation. Deepfakes refer to the use of artificial intelligence to produce realistic audio or video recordings that replace the likeness of one person with the likeness of another (Ayuya, 2024). Despite its useful applications in media, this technology raises significant ethical concerns, particularly in the context of creating pornographic material. It has recently been reported that a large number of women in South Korea have been victimized by non-consensual pornographic deepfakes (Thorbecke, 2023). There are ethical concerns surrounding deepfakes not only in relation to adults, but also with minors, who are increasingly becoming targets. As a result of this experience, victims may suffer long-term psychological trauma, lose control over their image, and have difficulty trusting others.

The exponential growth of deepfake content is alarming. In 2023, approximately 95,000 deepfake videos were available online, with 96% of them being pornographic, reflecting a 550% increase compared to previous years (Katherine, 2024). There are a variety of victims of malicious content, including public figures and private citizens, which amplifies the psychological and social consequences of this technology. The current deepfake detection tools, although inadequate, are unable to keep pace with rapidly evolving technology, which indicates an urgent need for a more effective solution. However, despite the obvious harm, there is still a need for more effective deepfake detection tools.

The proposed solution to this ethical challenge is **DeepShield**, a mandated AI-powered plug-in designed specifically to combat the misuse of deepfake technology in pornography. The primary feature of DeepShield is its ability to notify victims when a deepfake has been detected, allowing them to take action promptly. Furthermore, the platform provides its users with educational resources that teach them how to protect their privacy and prevent their content from being exploited. As part of its collaboration with government bodies, DeepShield aims to ensure that most existing pornographic deepfakes are detected and removed from the Internet on a broad scale. A unique feature of DeepShield is its ability to actively monitor social media platforms and user-generated content. The APIs are used to tag and track suspect media, thereby providing comprehensive coverage of online platforms. In addition, DeepShield enhances traditional AI

detection methods by utilizing advanced algorithms that are continually trained on large datasets. With this constant improvement, we are able to provide more accurate detection and faster response times. Still, despite these advancements, keeping up with the rapid evolution of deep fake technology remains a challenge.

The report is split into multiple sections: The Background, where the social and psychological implications of deepfakes are delved into in much greater detail ; The Problem Statement, which explains the significance of these issues and why they need to be resolved ; The Solution , where “DeepShield” and its functionality and features are explained ; The Agile team process , which explains the roles and tasks of each member ; The Conclusion, where the proposed solution is summarised.

2.0 Background

In light of the rapid development of AI and deepfake technology, there have become an increasing number of ethical issues associated with the use of non-consensual AI-generated pornography. The use of deepfakes, which rely on machine learning algorithms to produce realistic but false images and videos, has been weaponized to violate the privacy and dignity of individuals. Women have been disproportionately affected by deepfakes. One of the most alarming cases occurred in South Korea, where numerous women were victimized by deepfake pornography, which resulted in severe emotional, psychological, and reputational damage to them. In many instances, such exploitation results in psychological trauma that manifests in anxiety, depression, and social isolation (Choe, 2024). Furthermore, the availability of this content perpetuates a cycle of victimization that can be difficult to break.

The proliferation of deepfakes has also adverse ethical and societal consequences, eroding public trust in digital content. The sophistication of deepfakes makes it increasingly difficult to distinguish between genuine and manipulated media, creating an environment of distrust. Additionally, this undermines the potential benefits of AI technologies in legitimate applications, such as healthcare and creative industries. In the absence of solutions, fears of AI misuse could stall its development and limit the positive impacts it can have on society if left unaddressed.

Due to the urgency of these challenges, the government has step in to combat the spread of non-consensual explicit content. Social media platforms have become viral, which makes it easier for deepfake pornography to spread rapidly, exacerbated by the viral nature of these platforms, making existing detection tools ineffective. Consequently, the government has commissioned specialized teams to develop advanced technological solutions intended specifically for detecting and eliminating deepfake pornography. Through these efforts, individual rights will be upheld, citizens will be protected, and public trust in AI technologies will be restored by ensuring responsible and ethical use of AI technology.

2.1 SIGNIFICANCE OF ADDRESSING THE ISSUE

2.1.1 Case Study: Deepfake Misogyny in South Korea, 2024

South Korea has become a hotspot for deepfake pornography in recent years, where artificial intelligence is used to exaggerate the faces of victims and superimpose explicit content onto them. Authorities discovered deepfake pornography during an investigation for blackmail in 2020, which brought this issue to public attention. As of 2024, the police had reported 297 deepfake sex crimes between January and July, compared to 156 at the end of 2021. A number of these deepfakes are shared in encrypted Telegram chat rooms, some of which have upwards of 220,000 members (Choe, 2024). Over half of the world's deepfake videos target South Koreans, with 96% of them being pornographic, according to cybersecurity reports (Choe, 2024)..

There are devastating consequences for victims, as many women find out about their exploitation only after receiving anonymous messages containing deepfakes. It was reported that two graduates from Seoul National University were arrested and sentenced to five years in prison for creating deepfake pornography targeting female classmates. While legislation was enacted in 2020 to address this issue, there are still significant gaps in the regulation of those who view or store this material. There have been criticisms that the legal framework has not kept pace with the rapid development of deepfake technology, leaving many victims vulnerable and without adequate protection (Choe, 2024).

South Korean deepfake is a recent and critical example of how artificial intelligence technology can be misused to harm individuals. The purpose of deepfake pornography is not simply to exploit individuals, but also to have far-reaching societal consequences. Women and marginalized communities are particularly affected by it because it disrupts their lives, erodes their privacy, and perpetuates a cycle of victimization (Story & Jenkins, 2023). In order to ensure the ethical use of AI, the issue of trust in AI technologies must be addressed.

2.1.2 Case Study: UK Deepfake Pornography Scandal, 2023

The UK was subject to a deepfake pornography scandal in 2023 that involved more than 250 public figures, including celebrities, journalists, and influencers. The faces of these women were superimposed onto explicit videos which were then widely shared on adult websites and encrypted messaging platforms, generating millions of views. According to a Channel 4 investigation, more than 143,000 deepfake videos were uploaded by the end of 2023, exposing a legal loophole since such content is illegal to distribute, but is not regulated to create (Cursor, 2024).

There have been reports of significant emotional damage and reputational damage suffered by the victims and they have learned about the abuse only after receiving harassing messages. In spite of the fact that the Online Safety Act was introduced in 2023 with the intention of combating the distribution of non-consensual explicit content, it failed to address the problem of deepfake pornographic content. As a result of this situation, public and media outcry has prompted tech companies to take more robust steps to prevent such violations in the future. Legislative reforms and stronger technological safeguards are urgently required to prevent such violations (Cursor, 2024).

After thorough research, the product "DeepShield" was created to address these issues at the governmental level in order to address broader societal goals including preventing online exploitation, regulating AI usage, and encouraging responsible AI innovation. The government's involvement as a key stakeholder underscores the necessity of an effective solution that can protect citizens from non-consensual AI-generated pornography while preserving the ethical use of deepfake tools for creative purposes.

2.2 DEVELOPMENT OF BUILDING THE SYSTEM

The creation of deep fakes relies extensively on machine learning techniques, in particular the use of neural networks to produce highly realistic yet artificial content. Several methods exist for manipulating images and videos, including Generative Adversarial Networks (GANs) and Autoencoders, which offer seamless manipulation, including swapping faces. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) are also used for deepfakes, enabling facial features, expressions, and movements to remain consistent across multiple videos. The following section provides an explanation of each of these key components:

2.2.1 Generative Adversarial Network (GAN).

There are two neural networks that compose a GAN:

1. **Generator**: Based on input data, this network produces fake images, video frames, or audio samples to produce realistic content.
2. **Discriminator**: This network determines whether the generated content is authentic by distinguishing between authentic and fraudulent content. In order to improve the output of the generator over time, it provides feedback to the generator.

   The generator and discriminator work together, with the generator continually improving its ability to create convincing deep fakes, and the discriminator continuously improving its detection capabilities (Brownlee, 2019).

2.2.2 Autoencoders

Autoencoders are types of neural networks used in the creation of deep fakes, and are primarily utilized in the face-swapping process. It consists of two parts:

- **Encoder**: Using an encoder, the input image like that of a human face is compressed into a representation of a lower dimension known as latent space.
- **Decoder**: A reconstruction of the image is performed from the compressed representation. In deepfake creation, two decoders are utilized: one for the target face and one for the source face. In this way, the source and target faces can be swapped (Bergmann & Stryker, 2024).

2.2.3 Convolutional Neural Networks (CNNs)

An algorithm using convolutional neural networks is used to detect and extract facial features, including eyes, mouths and noses. The deepfake appears more realistic due to their accurate mapping of facial expressions and movements between people (IBM, 2024).

2.2.4 Recurrent Neural Networks (RNNs)

The use of RNNs, particularly Long Short-Term Memory (LSTM) networks, is a critical component of deepfake videos if they are to maintain consistency in sequences of frames. Through their use, facial expressions and head movements are tracked and predicted, ensuring consistency throughout the video (IBM, 2024).

3.0 Problem Statement

Deepfake technology is being misused for the creation of non-consensual pornographic content. This has resulted in significant violations of personal control and privacy, which has caused emotional and psychological harm, reputational damage, and widespread social consequences. A significant gap in legal and regulatory protections is exposed as a result of this type of exploitation, which disproportionately affects vulnerable individuals, particularly women and minors. It will be necessary to develop responsible computing solutions that prioritize privacy, fairness, and the wellbeing of individuals in order to restore trust in AI technologies.

3.1 Violation of Consent and Personal Control

Deepfake pornography violates consent and undermines an individual's self-rule over his or her own body and image. Without their knowledge or consent, pornographic deepfake videos and images portray them in explicit content, thereby stripping them of their basic right to consent. This situation is worsened due to the lack of laws, which directly bans the creation and distribution of deepfake (Chesney & Citron, 2019). This allows deepfake content to be widespread allowing anyone to see the victim in a situation they were never in. This can distort the public perception of them ruining their reputation for an act they have never performed. Additionally, due to the victims' lack of power to remove or prevent the further dissemination of the content, they may lose their sense of personal control. This can have adverse effects on the individual such as depression, anxiety, and even physical health problems (Koehler, 2024).

3.2 Invasion of Privacy and Exploitation

Deepfakes represent a severe violation of privacy when someone's likeness is used to create a deepfake without their permission. In this case, it is especially true when the deepfake is pornographic by nature. The issue is exacerbated by the fact that deepfakes require only a small number of images/videos to produce highly realistic content. Social media apps, which is where photos/videos are extracted for the production of deepfake content, have no terms or conditions to prevent such actions despite their clear violation of privacy rights. According to the social media app, an individual's image can be collected through both expressed and implied consent mechanisms, such as agreeing to terms of use by continuing to use the app (Sara H.,2024). Hence, there is no explicit rule preventing malicious users from extracting images and videos of an individual and manipulating it to produce a deepfake.

Deepfakes perpetuates gender-based violence and exploitation. A study showed that 96% of deepfakes made were non-consensual and were pornographic by nature, while 99% of those were made of women (Citron, 2020). This exploitation reinforces harmful stereotypes about women's bodies and reduces them to objects of sexual gratification without regard for their autonomy. This is represented in Korea as tensions were high when a telegram channel with more than 220,000 participants, made for the purpose of creating and distributing pornographic deepfakes was revealed. The main target for these pornographic deepfakes was women and it was performed non-consensually (Thorbecke, 2023). Content like this, only serves to perpetuate misogynistic attitudes and exacerbate gender inequality. Moreover, deepfake pornography has been weaponized as tools used for extortion, harassment and blackmail. Blackmailers may use deepfakes to attempt to extort their victims, and despite the deepfake itself being false the victim may be powerless in that situation as they are unable to stop the dissemination of the deepfake and any debunking would be unable disseminate as well as the deepfake. This may lead to the victim having to provide money, business secrets or nude images or videos (an act known as "Sextortion") to halt the release of the deepfake (Chesney & Citron, 2019).

3.3 Emotional and Psychological Harm

The increase in deep-fake pornography has led to a significant threat to the emotional and psychological well-being of its victims. The creation and widespread distribution of deepfake pornography can leave the victims feeling helpless and powerless, as they do not have the ability to stop the dissemination of the deepfake. Additionally, the victim may feel humiliated and scared to see their likeness being used in non-consensual explicit deepfake content. As a result, the victim may develop symptoms of psychological trauma. These symptoms include: Anxiety and panic, Post Traumatic Stress Disorder (PTSD) and depression (Nickert, 2023). Moreover, the harm caused by deepfake extends past the content itself but also how it distorts the public perception of the victim. Victims of deepfake porn are typically subjected to harassment and discrimination from others who believe that the deepfake they saw is genuine. This is exemplified in the case of Rana Ayubb, who became a target for deepfake pornography after she reported about an 8 year old girl being raped in India. The backlash and harassment she received from the deepfake, which was fully fabricated, were so severe that the United Nations (UN) had to intervene (Nickert, 2023). Thus the emotional and psychological harm caused by pornographic deepfakes is a severe issue that needs to be addressed.

3.4 Damage to Reputation and Social Impact

Deepfakes have the ability to cause severe and irreparable damage to a victim's reputation. The ability of deepfakes to depict its victims in various fabricated scenarios enables it to ruin someone's reputation. Due to how realistic deepfakes are, once a malicious deepfake of someone is created, the uninformed public may believe that the deepfake is true. This results in the social ostracization of the victim and damage to the victim's reputation. Moreover, the damage caused to the victim's reputation is permanent. The victim may lose out on future job opportunities and face career setbacks due to the stigma attached to the deepfake. Employers, who do background checks, may view the victim in an unfavourable way despite the deepfake content being fabricated. This issue is especially prevalent on the political scene, where deepfake videos of presidential candidates are made to dissuade the public from voting for them. A case of this occurred in 2023, where a deepfaked video of presidential candidate Donald Trump hugging Anthony Fauci, who is disliked by many of Trump's followers, was released. The purpose of said video was to attempt to temporarily damage Donald Trump's reputation and to reduce his number of followers (Bleisch, 2024). The long-term damage caused by deepfakes is often irreversible and the dissemination of

deepfakes makes it nearly impossible to fully erase the content from the internet. This highlights the need for technological solutions to address the use of deepfake technology.

3.5 Gaps in Legal and Regulatory Protection

Despite the vast potential for damage possessed by deepfake technology, there currently exists no current criminal law which outright bans the creation or distribution of deepfakes (Chesney & Citron, 2019). The lack of sufficient legal frameworks leaves limited options for victims to take when they are the target of deepfakes. This has occurred due to the rapid growth of deepfake technology outpacing the response from legislative bodies resulting in significant gaps in protection, which malicious deepfake creators exploit. This issue is exacerbated due to how current laws that protect privacy and address defamation were not made to defend against deepfake technology directly and do not cover the unique nature of deepfake content (Sara. H, 2024). This makes it difficult for victims of deepfake to take action against the creation of the deepfake as there are no clear legal grounds to act on. Deepfakes themselves are not inherently problematic, only when used in malicious ways do they become unlawful. Hence an outright ban would be difficult to implement as the nature of the deepfake produced must be considered. Additionally, the anonymity of the internet further increases difficulties in prosecuting legal action, due to how fast deepfakes disseminate across the internet, tracking down the original source of the deepfakes is a difficult task.

4.0 “Deep Shield”, The Proposed Solution

Flowchart Depiction of DeepShield:

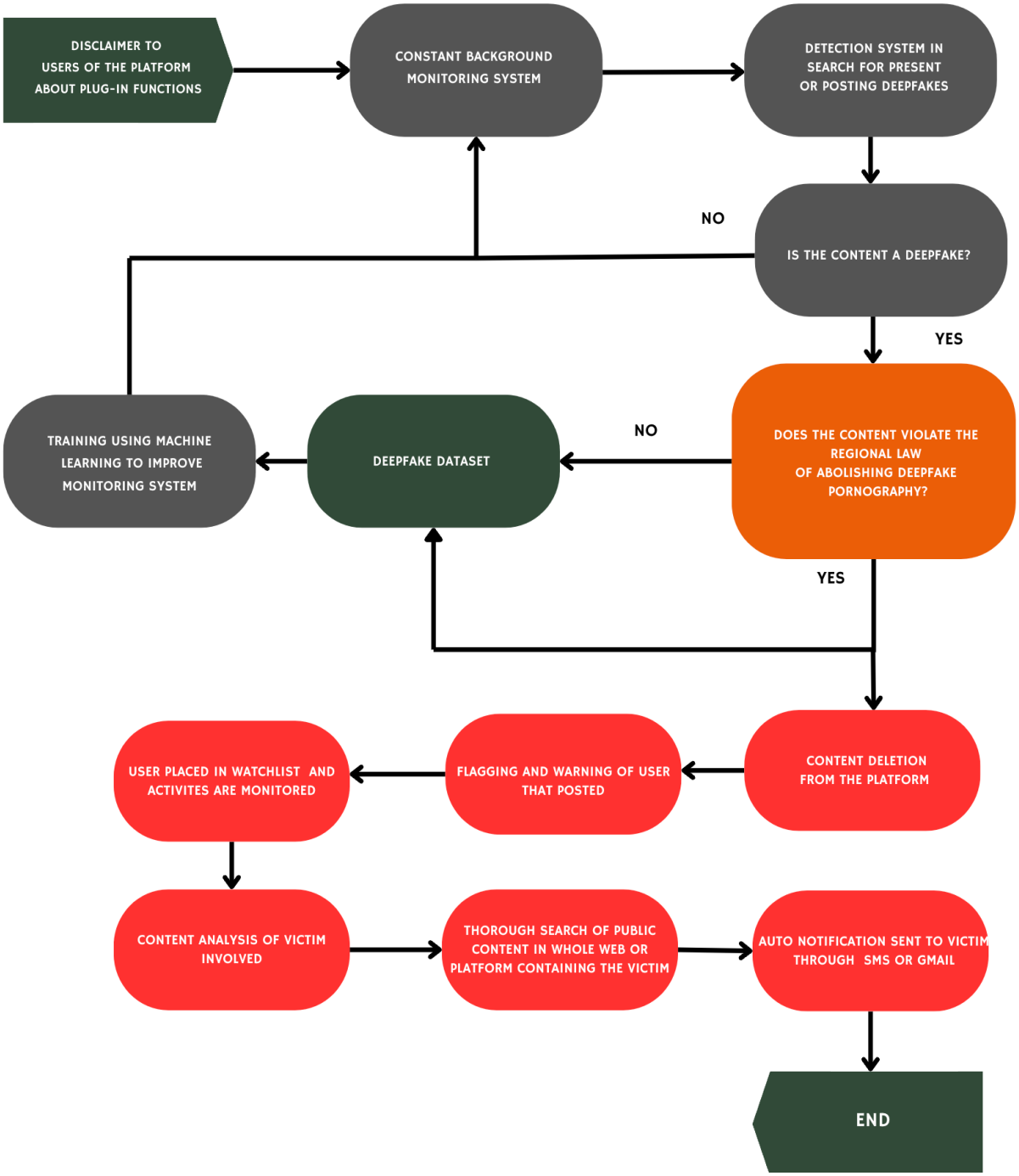

Figure 1: Proposed Solution's Flowchart

4.1 ACM CODE OF CONDUCT

4.1.1 CODE 1.1 Contributing to Human Well-Being

When used maliciously, the content generated by Generative Adversarial Networks (GANs), which is an AI component used to produce deepfakes, can lead to severe and long-lasting psychological trauma in its victims by replacing their likeness with explicit content. Additionally, the reputation of the victims is normally irreversibly damaged leading to life-long consequences. As was the case with Rana Ayubb, who became the target for harassment after a pornographic deepfake of her was created and disseminated.

4.1.2 CODE 1.2 Avoid Harm

The majority of deepfakes made disproportionately target women. This has led to the reinforcement of harmful stereotypes and perpetuates gender inequality. Moreover, the use of GANs and Machine Learning algorithms to increase the realism of the deepfake makes it more believable. This results in an increase in uninformed people believing the deepfake to be true. This causes psychological trauma in the victim, which is a clear violation of this code. The code states to avoid both physical and mental, which the use of GANs and machine learning algorithms has clearly violated.

4.1.3 CODE 1.4 Fairness and Non-Discrimination

As stated previously, the majority of deepfakes made disproportionately target women and this is in part due to the AI component of Convolutional Neural Networks (CNNs) and how it works. The datasets, which include videos and photos, fed into the CNNs are predominantly women (Jacobsen & Simpson, 2023). This results in deepfakes of women appearing more realistic than men and consequently the deepfakes of women are much more believable. This has resulted in women experiencing greater consequences and trauma compared to men. This is a clear case of discrimination and unfairness.

4.1.4 CODE 1.6 Respecting Privacy and Confidentiality

For deepfake technology to generate the likeness of an individual they first need a sample audio or video of the individual. Privacy issues arise when the source of these audios or videos are taken without consent from the individual. In many cases, the AI component GANs is trained using datasets that include personal images and videos without the consent of the individuals whose images and videos they represent. This infringes on many privacy rights laws and violates the individual's right to control how their image or likeness is used.

4.1.5 CODE 2.5 Responsible Computing

Despite the positive uses for autoencoders in improving the quality and realism of images. When used maliciously, it allows for the realism of deepfakes to be increased making it harder to distinguish real and fake. This can lead to the spread of misinformation and lead to severe distrust in the media. This undermines the responsible use of AI technology, which should be prioritised over human well-being.

4.3 USER STORIES

Table 1: User Stories

User story	Acceptance criteria	Functionalities
As a representation of the government, I want a deepfake detection tool that will scan all content uploaded to the internet, focusing on identifying and flagging deepfake pornography.	1. The contents posted by users must be scanned with the detection tool before they're allowed to post it. 2. Any deepfake pornography content in the present and future will be deleted immediately. deleted when it's detected.	1. The software will continuously scan for contents on social media platforms to detect any deepfake pornography. 2. The software will prevent a user from posting a content if it detects any deepfake pornography in the content. 3. The software will instantly delete any detected deepfake pornography on the social media after scanning.
As a representative of the government, I want a tool that will send out notifications to notify victims that they are involved in deepfake content.	1. It must send the notification as soon as possible as the deepfake content is detected regardless of. 2. It must ensure the deepfake content is first deleted. 3. It must provide information about the deepfake content detected and actions taken.	1. A feature that send notifications to the user 2. A 'contact" button to allow user to contact the local authorities to get help and report on the deepfake content

As a stakeholder, I want to monitor social media accounts to ensure we can catch repeated offenders and prevent them from posting anymore deepfake pornography.	1. Scanning must be done for all social media accounts to detect any deepfake pornography. 2. Offenders' accounts should be put on a watchlist.	1. A system is developed to scan and analyse user-generated content across social media platforms for deepfake pornography. 2. A feature is implemented where any users that post deepfake pornography will be flagged and added into a watchlist and kept track using a reputation score system. 3. A system is developed to check for an offender's affiliated accounts using IP address matching, email or contact information matching and shared behavioural pattern.
As a representative of the government, I want educational resources to be made available for all users at all times.	1. Users need to be educated about the risks of deepfakes and how to protect themselves. 2. Support systems offering professional help should be available to users.	1. An e-booklet is provided that includes clear and understandable information about the risks of deepfakes the possible malicious usage. 2. A beginner-friendly guide is available for users to learn how to safeguard their personal information and privacy. 3. Interactive tutorials are provided to teach users how to spot deepfakes. 4. A 24/7 helpline support system is available for users to seek professional help for legal advice.

As a government representative, I want a deepfake model that can continuously update itself with the latest algorithms and leverage advanced deepfake detection technology.	1. It must have automated algorithm updates.	2. A self-built model that leverages AI to utilize a deepfake database for continuously updating its detection algorithm would require the following key functionalities and criteria 3. As deepfakes continue to advance, detecting them becomes increasingly challenging as they become more realistic. Consequently, the model will be consistently updated with deepfake content and trained on these datasets to enhance its detection capabilities.

4.4 ACM CODE OF ETHICS: FUNCTIONALITIES AND DATA REQUIREMENTS

Table 2: Functionalities and Data Requirements

Functionalities	**Data needed**	**ACM Code of Ethics**
A feature that automatically scans internet content for deepfake pornography using AI comparison algorithms, detects manipulated videos, and instantly removes it.	**Video/Image Data** Contents from website, social media , and video-sharing platforms, which needs to be analysed for potential deepfake manipulation. **Facial/Biometric Data** Authentic images or video footage of individuals, used for comparison with detected deepfake content. **Metadata**: Associated content metadata, such as upload date, platform, and uploader information, to trace the source and context of the content.	**1.2 - Avoid harm** This feature should detect and prevent the spread of deepfake pornography to protect the victims from exploitation, mental distress and reputational damage hence keeping them out of harm's way **3.1 - Ensure that the public good is the central concern during all professional computing work** This feature should combat the spread of harmful, manipulated content that could damage people's lives. Making this feature a priority ensures a greater societal benefit

A feature that sends instant notifications to users when a deepfake of them is detected, allowing quick action. Users can choose SMS, email, or in-app alerts with options to remove the content or report it, followed by updates on progress.	Contact Information of the user -Phone number (for SMS notifications) -Email address (for email alerts) -In-app user account details (for app notifications) Facial/Biometric Data: For the use of deepfake	**1.2 - Avoid harm** Users should be notified immediately to help prevent further harm as users can take action quickly to protect themselves. **1.4 - Be fair and take action not to discriminate** System should provide equal access and protection to all users without discrimination to ensure all users are able to access the feature. **1.7 - Honour confidentiality** System should handle sensitive content and detected deepfake content in a confidential manner. Notifications should only be sent to users that are affected, preventing any unnecessary exposure to anyone else.

A feature that provides an educational module within DeepShield to teach users about the risks of deepfake pornography and online privacy protection. It includes tutorials on recognizing manipulated media, helping users identify fake content on their own.	(not applicable)	**1.2 - Avoid harm** Educational module should help users avoid harm by equipping them with knowledge about deepfake pornography and online privacy protection. **2.3 - Know and respect existing laws pertaining to professional work** The content in the educational module must adhere to laws related to digital rights, data privacy, and media manipulation, ensuring that users are educated with a legal framework that aligns with relevant regulations. **2.7 - Foster public awareness and understanding of computing, related technologies, and their consequences** The educational module should support this principle by fostering public awareness of deepfake technology and the risks it poses. It should help users to better understand the technology and its implications, making them more informed and capable of recognising harmful content.

A feature that includes a watchlist to monitor users flagged for posting deepfake content, tracking high-risk individuals and their activities for repeated violations using automated monitoring and a reputation score system. It also monitors affiliated accounts by matching IP addresses, emails, and detecting shared behavioural patterns to prevent the creation of multiple linked accounts.	1. User information User IDs and profile information are needed to aid in identifying patterns. 2. Content data All content posted by users are needed to be analysed by our system. 3. Flagging history To track timeline of offences and record history of flagging for all offenders. 4. User activity logs Information on a user's activity such as their browsing patterns are needed to get insights into how they navigate the platform and the types of content they engage with	**1.2 - Avoid harm** This feature should monitor and deter malicious actors who post deepfake content. It must ensure a safer online space by tracking repeat offenders and their affiliated accounts. **1.4 - Be fair and take action not to discriminate** While flagging users, the system must be fair and just to ensure there is no discrimination and the flagging is purely based on the activity of the user and the kinds of content posted by them. **1.7 - Honour confidentiality** Data collected on flagged users should be kept confidential and only accessible to authorised personnel to ensure there are strict measures to protect privacy and confidentiality of all users.

A feature that automatically updates the deepfake system algorithm utilising the AI components of the system to improve deepfake detection.	1. User behaviour data, Information on the source of the deepfake (e.g., social media platforms, news outlets, private networks) can provide context for identifying trends and patterns in the dissemination of deepfakes. 2. Deepfake dataset A large dataset of deepfake videos, images, and audio files. These files must represent a variety of deepfake manipulation techniques, formats, and styles to ensure the model can generalize across different types of deepfakes.	**2.5 - Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks** Every update to the algorithm should be thoroughly tested to ensure it improves performance without introducing new errors or risks. **2.9 - Design and implement systems that are robustly and useably secure** Continuous updates to the AI components must prioritise security. Proper safeguards should also be installed to ensure the updates are secure, verified, and free from interference.

4.5 ALGORITHM TECHNIQUES

Table 3: Algorithm Techniques

Functionalities (what the product must do)	Data needed	Required Algorithm	Storage of Data
A feature that automatically scans internet content for deepfake pornography using AI comparison algorithms, detects manipulated videos, and instantly removes it.	**Video/Image Data** Contents from website, social media , and video-sharing platforms, which needs to be analysed for potential deepfake manipulation. **Facial/Biometric Data** Authentic images or video footage of individuals, used for comparison with detected deepfake content. **Metadata**: Associated content metadata, such as upload date, platform, and uploader information, to trace the source and context of the content.	A Comparison-Based AI Algorithm for facial detection	Data stored securely with strict access control mechanisms, ensuring that only authorised personnel or systems have access to the stored biometric data.(ACM 1.7)
A feature that sends instant notifications to users when a deepfake of them is	Contact Information of the user -Phone number (for SMS notifications)	A Facial and Biometric Data Matching algorithm that	Data is backed up securely and that the backups are also encrypted. Have a clear disaster recovery plan in place to ensure the

detected, allowing quick action.	-Email address (for email alerts) -In-app user account details (for app notifications) Facial/Biometric Data: For the use of deepfake	matches deepfake content features the user's face	integrity and availability of data in case of a breach or hardware failure.(ACM 1.7)
A feature that provides an educational module within DeepShield to teach users about the risks of deepfake pornography and online privacy protection.	(not applicable)	(not applicable)	(not applicable)
A feature that includes a watchlist to monitor users flagged for posting deepfake content, tracking high-risk individuals and their activities for repeated violations using automated monitoring and a reputation score system.	1.User information User IDs and profile information are needed to aid in identifying patterns. 2.Content data All content posted by users are needed to be analysed by our system. 3.Flagging history To track timeline of offences and record	A user behaviour monitoring algorithm that tracks user behaviour and a reputation scoring algorithm that assign a reputation score based on their activities, such as previous violation.	Data is securely stored in a database with regular review and audit.(ACM 1.4)

	history of flagging for all offenders. 4.User activity logs Information on a user's activity such as their browsing patterns are needed to get insights into how they navigate the platform and the types of content they engage with		
A feature that automatically updates the deepfake system algorithm utilising the AI components of the system to improve deep face detection.	1.User behaviour data, Information on the source of the deepfake (e.g., social media platforms, news outlets, private networks) can provide context for identifying trends and patterns in the dissemination of deepfakes. 2.Deepfake dataset A large dataset of deepfake videos, images, and audio files. These files must represent a variety of deepfake manipulation techniques, formats, and styles to ensure the model can generalize across different types of deepfakes.	A reinforcement learning(RL) algorithm to allow the system to learn and improve over time by automatically adjusting its detection models based on feedback	Data stored in a secure database that is constantly monitored and audited to show transparency, data are also stored only as long as needed, then archive or delete it when no longer necessary.(ACM 2.3)

4.6 ETHICAL THEORY

The system prioritises acknowledging stakeholder input, particularly from the government, regarding how their data is utilised. It also incorporates measures to counter deepfakes, safeguard user identities from misuse, and prevent potential harm to users. We can use combination of all three:

- Virtue Ethics
- Utilitarianism
- Deontology

The justification is as follows:

4.6.1 Virtue Ethics

First of all, in virtue ethics, following the principles of Justice and Honesty and Responsibility. In Justice, the DeepShield feature must uphold fairness and equality for all users, ensuring no preferential treatment or bias in deepfake detection, it must apply the same level of scrutiny and protection to all users, regardless of their social, political, or economic status. Every individual, whether a public figure or a private citizen, should receive equal protection against deepfakes, ensuring fairness in how the technology is used. The system should reflect virtues of fairness and impartiality, it needs to operate without any bias, meaning that no group or individual should face harsher or more lenient detection measures, avoiding any misuse to fulfil personal fantasies, political agendas, or to gain an unfair advantage. The standards for DeepShield detection must be universally applied, promoting trust, accountability, and safeguarding the dignity and rights of every individual, in alignment with ethical principles of justice.

Whereas for honesty, in any of the data we collect via the DeepShield system, we must have a disclaimer every time when the user logins to the app, the system informs and gets users consent in the data being collected and used for the algorithm. Not just that, when DeepShield detects a deepfake, the system should accurately report its findings without exaggeration or manipulation, The system should clearly indicate when content is flagged as suspicious, providing factual reasons for why it was detected as a deepfake. This ensures users understand the real implications without fear-mongering or misleading information. Also, DeepShield should honestly inform users if their content has been flagged or removed due to suspected deepfake activity. Instead of silently

removing or altering content, the system should notify users about what happened and why, as part of the built in feature that comes along with the DeepShield system, promoting transparency and openness.

4.6.2 Utilitarianism

In Utilitarianism, we must first identify the various actions that we can take. We must then weigh the harms and benefits that can be caused by the actions taken. Lastly, we choose the action which leads to the greatest benefit with the harm taken into consideration.

A feature provided by Deepshield, which is its ability to monitor users who are flagged for posting deepfake content could possibly raise ethical issues. Following the ACM Code of Ethics, principles 1.6 (Respect Privacy) is violated as the monitoring of users infringes on their privacy. Additionally, the monitoring of affiliated accounts of the user, via their IP address and matching emails, further infringe principle 1.6 (Respect Privacy). However, the potential benefits gained by implementing this feature may lead to the betterment of human wellbeing. This feature can prevent users from generating or reposting pornographic deepfakes, thus decreasing the number of pornographic deepfakes circulating the internet. Additionally, it can protect potential victims from the severe psychological trauma and reputational damage. Weighing both the potential benefits and the potential harm caused, the feature to monitor flagged users will be kept. However, actions will be taken to ensure that the monitoring is done, while minimising the privacy infringement. This will be done by providing clear and transparent policies regarding the monitoring of flagged users. This ensures that users are well informed of the criterion which will cause them to be flagged. Additionally, the extent of data collected will be limited, ensuring that unnecessary data will not be collected. Thus following the principles of Utilitarianism the implementation of the feature is justified as it can be implemented in a way that improves human well-being while minimising privacy infringement.

Another feature of DeepShield which can potentially cause harm is the feature that automatically updates and improves the deepfake detection algorithm utilising the AI components of the system. This feature can potentially violate principle 1.3 (Be honest and trustworthy) of the ACM code of Ethics. If the method by which the detection algorithm is trained is not made transparent there may be concerns in the public about the fairness and accountability of the

detection algorithm. Moreover, the source of the data used to train the detection algorithm must be made clear, as fear in the public might rise out of fear that deepfakes of them are being used for training the detection algorithm. This may reduce human well-being. However, the improvement of the detection algorithm is crucial for DeepShield to keep up with the rapid advancements in the production of AI deepfakes. The reason why other deepfake detectors fail is due to their inability to keep up with the advancement of deepfake technology. An improvement in the detection algorithm can lead to the protection of a larger majority of people from being a victim of pornographic deepfake, and thus an increase in human well-being. The potential harm caused by this feature can be minimised by being transparent. DeepShield will provide clear information about how the AI is being updated, how the AI detection algorithm works and what data is being used to train the AI. This ensures that users can be updated on how the AI works and overall serves to increase the transparency of DeepShield. Hence after weighing the potential benefits and potential harm, we will keep this feature as it maximises human well-being, which aligns with the Utilitarian principles, while maintaining transparency and remaining ethical.

4.6.3 Deontology

In DeepShield , deontology ethics is applied to ensure the safety and privacy of the users as being first. One of the ACM Code of Ethics that DeepShield has violated was 1,6(Respect Privacy),as DeepShield System's feature to monitor flagged users and their affiliated accounts through IP addresses or emails could potential; infringe on this principle. As we all know, privacy is a fundamental right, and monitoring users based on deepfake-related activities might lead to the collection of data beyond what is necessary. This could violate the user's right in privacy invasion, especially if it involves tracking their activities without their explicit consent.

In Deontology , it emphasises that respecting the privacy of individuals is a moral obligations that should not be violated, even if it leads to beneficial consequences. Thus, even though monitoring the pool of flagged users might prevent the spread of harmful deep fakes, DeepShield should has a duty to limit the extent of the amount of data collected and unnecessary surveillance must be avoided. Also, DeepShield must ensure all data collected are consensual and transparent. Users should be given a clear information and disclaimer about what data is being collected, why is it collected and how would it be used.

Another ACM Code of Ethics that DeepShield has violated was 1.3(Be Honest and Trustworthy),as DeepShield is not providing information about how deepfakes are detected, how the AI model is trained, and what data is being used , this could lead to a breach of trust. Users must be fully informed about how their data is used and how the DeepShield system operates, ensuring honesty in the processing of their content. Failing to disclose this information could lead to the distrust and potentially harm the reputation of the system.

In Deontology, DeepShield must ensure the users are clearly informed about how the detection algorithm works and how their data contributes to the training and improvement of the AI models. The system should also operate in a way that is clear and open about how data is used and how decisions are made by the AI algorithm. This duty does not depend on the potential positive outcomes of the system but is a moral obligation to treat users with respect by being forthright. By prioritising honesty, transparency and trustworthiness, the system meets its ethical duty to provide truthful information and ensure users are aware of how their data is handled.

Hence , we can compile the above ethical theory and based on the guiding principles, we have amended the functionalities accordingly that considers not only the stakeholders, but also the users of the platform and the social media communities .

Table 4: Ethical Theory

Functionalities	**Acceptance criteria**	**Choice of ethical theory for ethical coding.**
A feature that automatically scans internet content for deepfake pornography using AI comparison algorithms, detects manipulated videos, and instantly removes it.	1. The contents posted by users must be scanned with the detection tool before they're allowed to post it. 2. Any deepfake pornography content has to be immediately when it's detected.	Deontology • Emphasises following moral rules or duties, regardless of the consequences. In this case, the focus would be on the duty to respect privacy, transparency, and honesty while upholding the principle of preventing harm (especially from malicious content like deepfake pornography).

A feature that sends instant notifications to users when a deepfake of them is detected, allowing quick action.	1. Users must receive the notification as soon as the deepfake content is detected. 2. User must be informed about the content that is being deepfaked and what's the action taken	Deontology • Algorithm must provide same level of scrutiny and notification, ensuring no one is treated unfairly and unequally , regardless of their status, should be prompt notified when their content is flagged as a deepfake.
A feature that provides an educational module within DeepShield to teach users about the risks of deepfake pornography and online privacy protection.	1. Users need to be educated about the risks of deepfakes and how to protect themselves. 2. Support systems offering professional help should be available to users	Virtue ethics • Focuses on developing virtuous behaviour and moral character traits like honesty, responsibility, and justice. Providing education to users on the dangers of deepfake pornography and privacy protection encourages responsible behaviour, aligns with promoting social good, and fosters a sense of justice by informing people about their rights and risks.

A feature that tracks users flagged for posting deepfake content, monitors high-risk individuals for repeated violations, and uses automated systems to prevent linked accounts by matching IPs, emails, and detecting shared behaviours.	1. Scanning must be done for all social media accounts to detect any deepfake pornography. 2. Offenders' accounts should be put on a watchlist.	Deontology ethics It requires honesty and openness, the system must be transparent about how users are flagged , tracked and placed on watchlists. This aligns with the duty of transparency toward users, ensuring they understand how their data is being used and how the system operates. Another core tenet of deontology is to prevent harm, the system must stop the distribution of harmful deepfake content, particularly deepfake pornography , which can cause significant emotional , reputational, and psychological harm towards victim.
A feature that automatically updates the deepfake system algorithm utilising the AI components of the system to improve deepfake detection	It must have automated algorithm updates.	Utilitarianism • Focuses on choosing actions that maximize overall well-being and minimize harm. Automatically updating the deepfake detection algorithm is essential to keeping up with advances in deepfake technology and ensuring the system remains effective at protecting users. The overall positive impact of improving the system outweighs potential harms, as long as proper ethical safeguards are in place.

5.0 Agile Team Process and Management

In this assignment we are using Agile Team processes and management. We have come together as a group to discuss thoroughly and have assigned each member a title that they are capable of, from the proficiencies and expertise that they have or are lacking in. This resulted in the discovery of the team structure being a Hybrid from the combination of Specialists and Generalists. There are 2 members who are relatively more knowledgeable about the functions of AI and their capabilities due to their interest and constant research : Lim Hui Zern and Chew Chen Hin, thus, they were deemed as the Specialists of the group and took on the task of creating the solution, due to the heavy workload of this task, they spent majority of their time solely on it. Conversely, the other three members: Chris Law Zi Qing, Tan Choong Sheng and Bradley Hoh Lok Yew have a more general knowledge of the topics, which has resulted in them being tasked with the research and writing tasks such as the introduction, background and problem statement, and conclusion along with assisting the solution task when they require writing assistance and general ideas to improve the aesthetic, and structure of the texts.

Figure 2: Agile Process

5.1 DELEGATION OF KEY ROLES

On our first meeting, each individual introduced themselves in order to gain a better understanding of their skills and personalities from their past experiences with similar group activities as well as their respective roles.

5.1.1 Team Leader

At the conclusion of the discussion where the group has come to the consensus of having the team leader being delegated to the most responsible and diligent individual, who took the lead in the discussion by displaying the most initiative when providing relevant information and questions to enrich the conversation. This individual displayed evident experiences regarding the aptitude for managing tasks, coordinating teams, and ensuring deadlines are met. She was responsible for creating and maintaining assignment schedules, effectively appointing tasks and ensuring team members were clear on their roles. Regular meetings hosted by her helped keep everyone directed and addressed potential roadblocks early on by providing support. This resulted in work being completed on schedule, she also provided timely reminders and status updates, fostering accountability within the team.

5.1.2 Product Owner

Similarly, the product owner was appointed to the most perceptive member with great communication skills that is capable of analyzing work well and providing useful insights into work. During the meeting, the member quickly grasped the tasks and requirements of the assignment and provided the team with detailed guidance on what to prioritize for specific tasks, as well as minor aspects that could improve the quality of the work. This individual is also cohesive, and confident when expressing themselves respectfully and could effectively convey his thoughts and ideas to others well, both verbally and in text while also capable of providing constructive feedback when necessary. These skills are essential for the meeting with the stakeholder as he'd have to be able to present the completed tasks well and to be able to quickly grasp the feedback he was given and convey these feedbacks to the team with clarity.

5.2 AGILE PROCESSES AND TOOLS

5.2.1 Sprint Planning

An agile team also requires mediums of communication for efficiency and to be able to convey ideas without much hassle. A tool we used was Discord, which is an app that allowed us to accessibly call each other as a group for Sprint Planning. This process happens on Monday afternoons every week which has been collectively agreed on to have relatively long discussions on progress and goals to achieve by the end of the week, these meetings usually consist of the team leader assigning the tasks to the members to be completed and would then be discussed on whether the workload is reasonable to the time given, if seemed difficult, other members would usually assist in these problems to lessen the work. The members' progress on their work are also assessed together and would each provide constructive feedback respectfully, which is effective to come to a conclusion for a specific task to refine their work. During these processes, we add our goals into "to do" lists in Trello as pending tasks to complete for the week where there are set due dates in the task cards, serving as reminders during the sprint as these cards are constantly viewed.

5.2.2 Daily Stand-Ups

Our team leader has requested for Daily Stand-ups dedicated at a time agreed upon every day where we would meet up at the hive and utilize Trello for members to comment on their specific task on the board about their progress and if dealt with any blockage, they would highlight it red to discuss about. These boards are also used for improvements that need to be made and highlighting flaws to be noticed by members working on the same task. If tasks are completed, it would be moved to the "done" list. We also use this process to communicate about our plans during the week where we might be busy. Members check on the Trello board daily to stay updated about the assignment progress and recent findings that could correct past misunderstandings of the task as changes and ideas are commented on the Trello board regularly. The daily use of this tool and its organization features has proven to be essential for our team members as tasks are constantly highlighted and commented on, thus, enhances visibility of the goals, priorities and maintains focus among the group members.

5.2.3 Sprint Review

On Sunday's evenings where the goals of the week are completed, we would have a short meeting with the use of Discord to conveniently call and discuss the presentation to the stakeholder, regarding the aspects to show and discuss for more clarity. During this process, the product owner will receive feedback on what certain aspects are lacking and can be improved, he will then note these down according to the specific tasks in the comment section in the Trello board for the other members to review it after.

5.2.4 Sprint Retrospective

Following the stakeholder meeting, the team leader will suggest a brief Discord session to discuss the product owner's feedback. The intention of this meeting is to assess the team's performance and identify areas for improvement. The Product Owner will share insights gained from the stakeholder, providing guidance on the group's future direction and how to better articulate work in upcoming Sprint Planning sessions. The discussion will focus on recognizing individual members who excelled in their tasks and using their successful approaches as a benchmark for future work. The team will also explore strategies for proactively addressing challenges and enhancing communication to maximize meeting effectiveness.

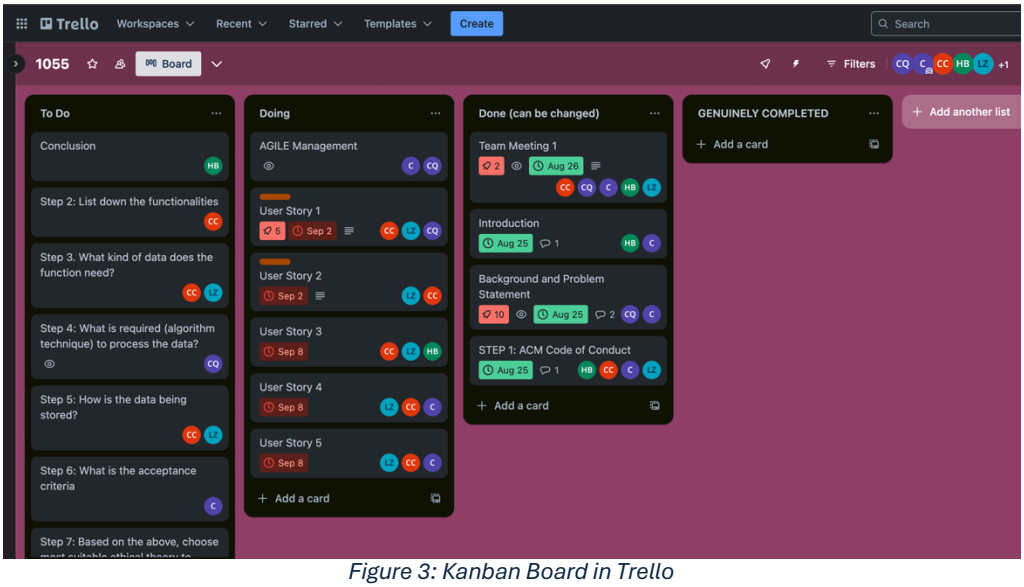

Figure 3: Kanban Board in Trello

5.3 COLLABORATION PROCESS AND TASK ASSIGNMENT

Task assignment and the process of cooperation are necessary for building a group that is effective and efficient. When the responsibility is given to the team members who have relevant strengths, the team is able to use their necessary skills to produce quality work efficiently. In turn, these will result in increased performance and a sense of responsibility among the members of the team. In this regard, the team made sure that the process was carried out in such a way that each and every member has a chance to contribute. Regularly, members held brainstorming meetings, where everyone voiced their opinions and collectively agreed on who should handle different tasks by having productive negotiations of people being able to voice their difficulties if tasks are too difficult for the sprint. These negotiations result in work adjustments by cutting down on the workload for the sprint if due dates can be lenient. However, if the workload for the sprint may seem cumbersome, more members will be assigned to that specific task as support. This process allowed respect and cooperation to be established among group members to produce work where everyone can be satisfied with their quality work. Lastly, recognizing and appreciating what each member brings to the table, helped the team to come up with constructive input and ideas when dealing with obstacles together.

5.4 TEAM LEADERSHIP'S APPLICATION

Team leadership is essential for ensuring both efficiency and effectiveness, particularly in the Agile process. She played a crucial role in task management and organization by strictly adhering to Agile processes such as sprint planning, daily stand-ups, and sprint retrospectives. During sprint planning, she effectively directs the team by assigning goals and breaking down tasks, to ensure that everyone understands their responsibilities. She constantly hosts daily stand-ups to promote accountability and keeps the team aligned, allowing members to share progress, identify roadblocks, and adjust priorities as needed. Additionally, she hosts sprint retrospectives, creating a safe space for team members to reflect on their work, discuss what went well, and identify areas for improvement. This allows constant improvement and also enhances team cohesion and collaboration. Her leadership provides essential support, motivating team members and addressing any concerns by maintaining clear communication and providing guidance throughout the Agile process, she ensures that the team operates smoothly, ultimately leading to higher productivity and successfully producing quality work.

6.0 Conclusion

Ultimately, DeepShield offers an AI-based solution that is robust, intelligent, and designed in order to address the ethical dilemmas surrounding the non-consensual creation and distribution of pornographic deepfakes. In order to provide high-precision detection and monitoring of deep fake content, DeepShield utilizes advanced technologies such as Generative Adversarial Networks (GANs) and Recurrent Neural Networks (RNNs). One of the key features of the system is the immediate notification of victims, the removal of content in real time, and the provision of privacy protection. As deepfake technologies evolve, the system's adaptive algorithms make it capable of staying effective, ensuring timely responses and empowering users to protect their online reputations.

DeepShield works closely with government agencies and social media platforms to ensure widespread enforcement and compliance against the proliferation of deepfake pornography. A key benefit of DeepShield is its ability to address the psychological, emotional, and social harm caused by deepfake exploitation, while minimizing the spread of harmful content for victims. According to the ACM Code of Ethics, DeepShield adheres to a number of ethical principles, including our commitment to privacy, harm prevention, well-being enhancement, and fairness. In order to solve the ethical challenges presented by deepfake technologies, DeepShield emphasizes a clear, integrated, and forward-looking approach.

The emphasis that DeepShield places on user education plays an important role in providing individuals with the knowledge and tools to safeguard their personal data and prevent its misuse. DeepShield provides educational resources that not only address immediate threats but also promote a long-term awareness of the risks associated with privacy and digital security. Through this approach, DeepShield not only responds to ethical concerns, but also cultivates an informed and vigilant digital environment. Through collaboration with the authorities, the solution remains adaptable and compliant with ever-evolving regulations, further strengthening its practical and ethical attributes. DeepShield's focus on both technological aspects as well as human factors ensures a comprehensive response to the threat of deepfake pornography.

7.0 Appendix

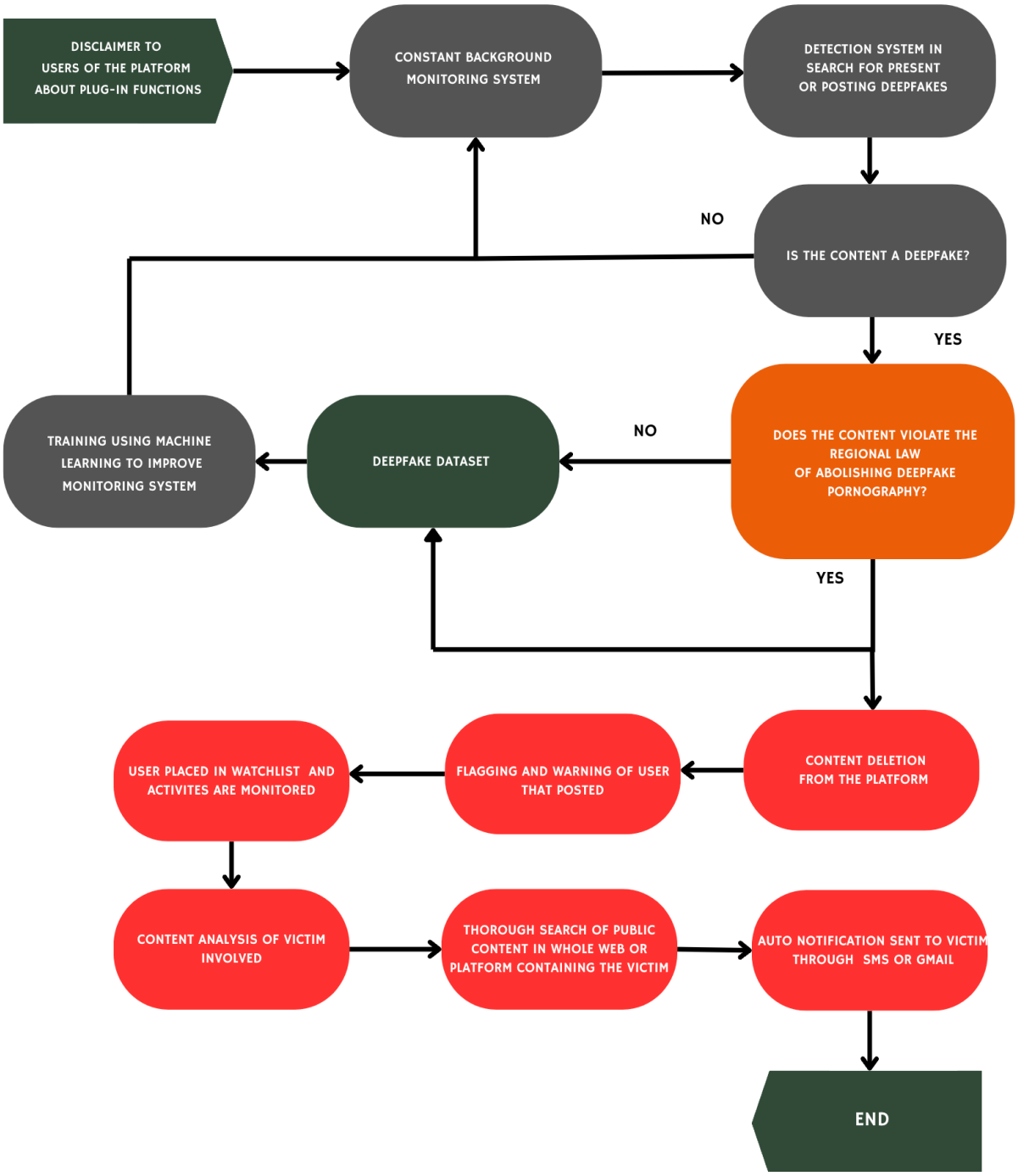

Figure 4: Proposed Solution's Flowchart

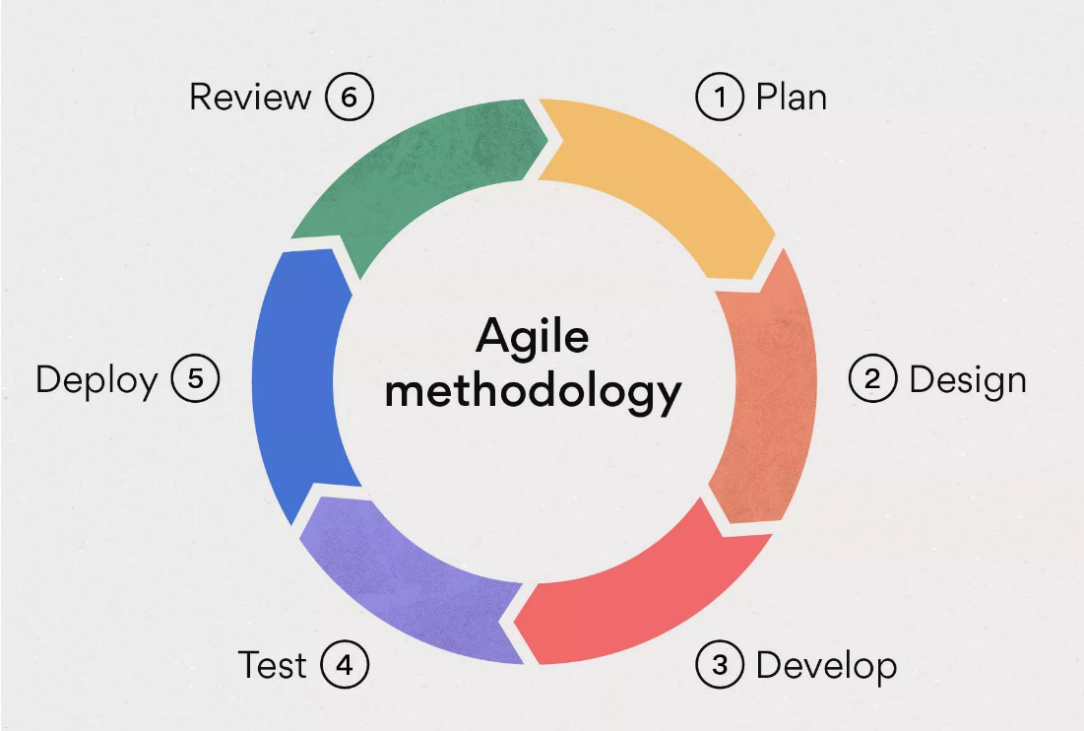

Figure 6: Agile Process

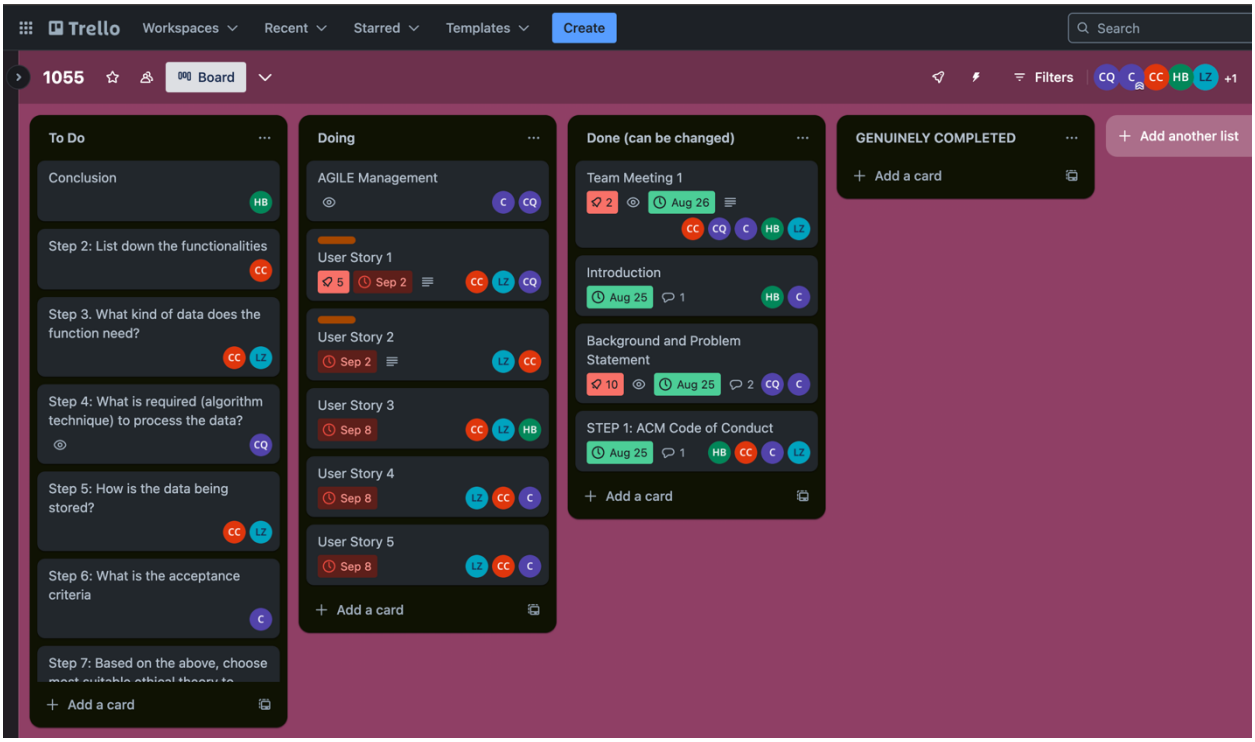

Figure 5: Kanban Board in Trello

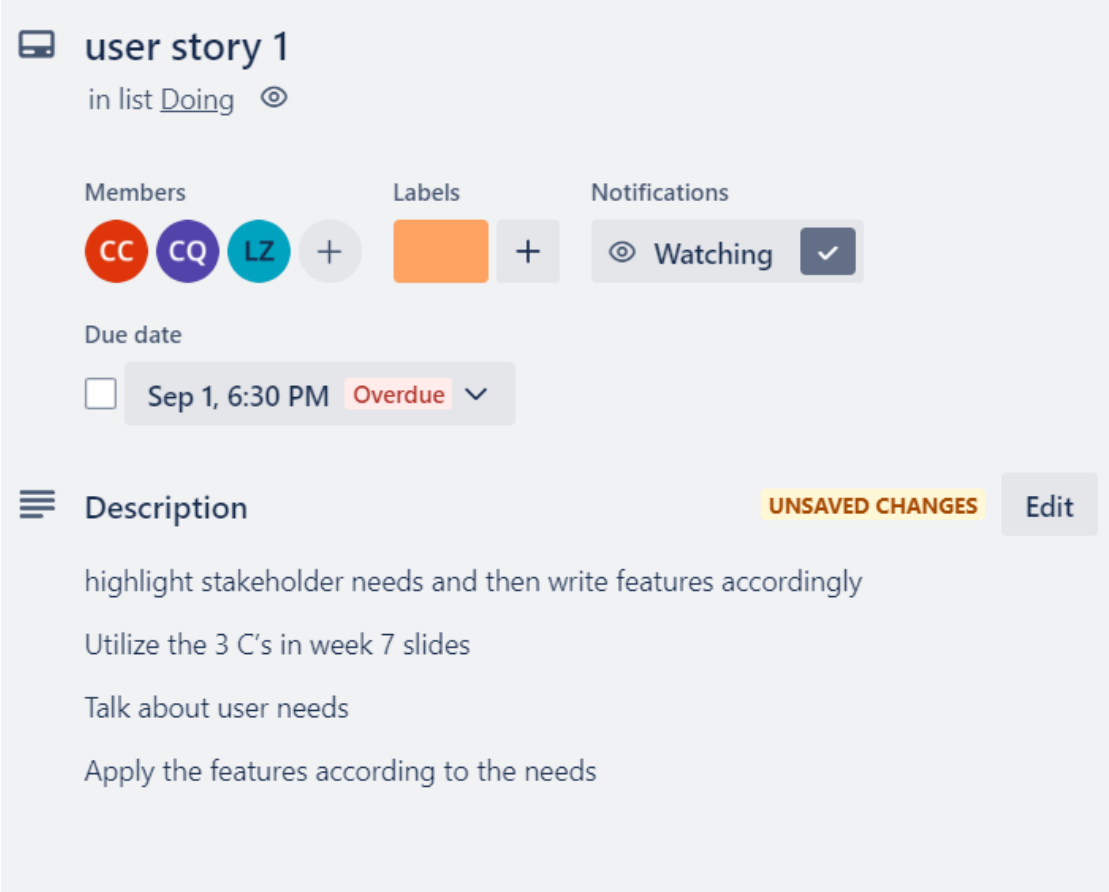

Figure 4: Description of One of the Cards from the Kanban Board

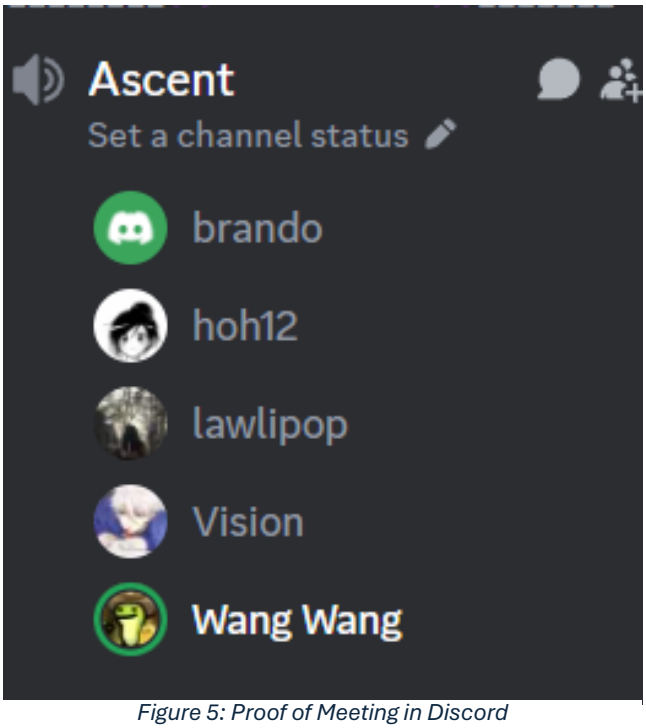

Figure 5: Proof of Meeting in Discord

8.0 References

Ayuya, C. (2024, July 29). *Ultimate guide to AI Deepfake technology*. eWEEK. https://www.eweek.com/artificial-intelligence/deepfake/

Bergmann, D., & Stryker, C. (2024, August 29). Autoencoder. *IBM*. https://www.ibm.com/topics/autoencoder

Bleisch, D. (2024, May 6). *Deepfakes and American elections*. https://www.americanbar.org/groups/public_interest/election_law/american-democracy//resources/deepfakes-american-elections/

Brownlee, J. (2019, July 19). *A Gentle Introduction to Generative Adversarial Networks (GANs)*. https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Buxton, D., & Kepchar, A. (2024, March 21). *Is deepfake porn illegal?* https://www.minclaw.com/deepfake-porn-illegal/

Chesney, B., & Citron, D. (2019). Deep Fakes: a looming challenge for privacy, democracy, and national security. *California Law Review*, *107*(6), 1753–1819. https://doi.org/10.15779/z38rv0d15j

Choe, S.-H. (2024, September 13). *In South Korea, Misogyny Has a New Weapon: Deepfake Sex Videos*. The New York Times. https://www.nytimes.com/2024/09/12/world/asia/south-korea-deepfake-videos.html

Cursor, C. (2024, April 29). *How deepfake is using UK celebrities on pornographic content*. Cursor. https://www.cursor.org/technology/2024/04/26/deepfake-uk-celebrities-pornographic-content.html

Citron, D. (2020, February 7). *Women, not politicians, are targeted most often by deepfake videos*. Centre for International Governance Innovation. https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/

IBM. (2024a, August 13). Convolutional Neural Networks. *IBM*. https://www.ibm.com/topics/convolutional-neural-networks

IBM. (2024b, August 29). Recurrent neural network (RNN). *IBM*. https://www.ibm.com/topics/recurrent-neural-networks

Jacobsen, B. N., & Simpson, J. (2023). The tensions of deepfakes. *Information Communication & Society*, *27*(6), 1095–1109. https://doi.org/10.1080/1369118x.2023.2234980

Jodka, S. H. (2024, February 2). *Manipulating reality: the intersection of deepfakes and the law*. https://www.reuters.com/legal/legalindustry/manipulating-reality-intersection-deepfakes-law-2024-02-01/

Nikert, J. (2023, November 16). *The Damage Caused by Deepfake Porn*. https://healthnews.com/mental-health/anxiety-depression/the-damage-caused-by-deepfake-porn/

Sharadin, G. (2023, December 20). *What is scraping | About price & web scraping tools | Imperva*. Learning Center. https://www.imperva.com/learn/application-security/web-scraping-attack/

Story, D., & Jenkins, R. (2023). Deepfake pornography and the ethics of Non-Veridical Representations. *Philosophy & Technology*, *36*(3). https://doi.org/10.1007/s13347-023-00657-0

Thorbecke, C. (2024, September 10). Commentary: South Korea is facing a deepfake porn crisis. *CNA*. https://www.channelnewsasia.com/commentary/south-korea-deepfake-porn-crisis-telegram-women-girls-school-photos-social-media-selfies-4595571