Addressing the Threat of Deepfakes: Detection Technologies, Societal Impacts, and Future Directions

MSc Research Project
Programme Name

Michael O'Toole
Student ID: x22192131

School of Computing
National College of Ireland

Supervisor: Michael Pantridge

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Michael (Mick) O'Toole

Student ID: x22192131

Programme: MSc Cybersecurity **Year:** 2024

Module: MSc Research Project

Supervisor: Michael Pantridge

Submission Due Date: 12th August 2024 @ 14.00

Project Title: Addressing the Threat of Deepfakes: Detection Technologies, Societal Impacts, and Future Directions

Word Count: 10002 **Page Count** 27

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: Michael O'Toole

Date: 10th August 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):	

Acknowledgements

I would like to take this opportunity to thank all the people that contributed to the completion of this thesis. First I would like to thank the various lecturers in National College of Ireland that provided excellent insight and expertise into teaching the individual modules. I would also like to thank my classmates. The support I received from them throughout the course was invaluable. My company Hostelworld supported me with time and resources when required and I could not have done it with that help and support. Lucy for keeping me company as I worked toward the thesis deadline. My friends and family that constantly supported and encouraged me to keep going. Lastly I wanted to recognise Rachael, Conor & Darragh for their unwavering support and for the sacrafices that they made so that I could complete this course.

Addressing the Threat of Deepfakes: Detection Technologies, Societal Impacts, and Future Directions

Michael O'Tole
x22192131

Abstract

The invention and proliferation of deepfake technology has brought about significant advancements in digital media manipulation. These advancements have created both opportunities and challenges. This thesis explores the creation, detection of and societal impacts of deepfakes and investigates the development of countermeasures. Through a comprehensive literature review, this study discusses the ethical and legal implications and identifies areas for future work. The ultimate goal is to provide a thorough understanding of deepfake detection and explore a multifaceted approach to developing countermeasures.

1 Introduction

1.1 Background

Deepfakes can be defined as digital content that has been altered or generated to depict events or actions that did not happen. Through the use of deep learning algorithms it is now possible to create hyper-realistic content that is almost impossible to detect with the naked eye. The increase in sophistication of learning algorithms and the use of Generative Adversarial Networks (GANs) have seen a noticeable increase in the quality of the synthetic content that is being created. So much so that many deepfakes now require technological intervention to detect whether or not an image has been synthetically produced.

It is not difficult to think of the positive applications of deepfake technology. The entertainment and education sectors could benefit from the technological advancements of deepfakes to create immersive learning experiences of historically accurate depictions of events. The entertainment industry could also use deepfake technology to bring significant historical figures back to life and recreate scenes that would look so realistic that it would be akin to watching a documentary.

The darker side of deepfake technology is far more sinister. The generation of digital media depicting world leaders saying or doing things that did not happen could shape societal views and lead to governance decisions based on fictitious events. Fake content that could manipulate financial markets could cause a global recession. Experiments in how deepfake content affects the general public have shown that synthetic content can significantly shape public opinion and manipulate individuals thoughts and actions (Channel 4).

The threat of deepfake content cannot be understated. It threatens to undermine the integrity of digital media and information dissemination as we know it. The spread of misinformation and the erosion of privacy are serious concerns facilitated by the threat of deepfake content. The ability to accurately and consistently detect and report on synthetic or deepfake media is paramount to counteracting the threat of deepfakes. Indeed it can be seen in the current Presidential race that the supporters of Republican candidate Donald Trump have been caught using deepfake images to curry the vote of the African American community ("Trump supporters target black voters with faked AI images"). We have also seen instances recently where Deepfake content was used to create a profile photo that was submitted to an organisation as part of a job application. The image, which was faked, led to the organisation hiring the individual which turned out to be a North Korean hacker (Sjouwerman).

These scenarios are obvious examples of how deepfake detection capabilities are lagging behind the technology used to create them. Ensuring that the truth is not warped by deepfakes is now of societal and critical importance. The ability of the general public to create realistic deepfakes creates an almost insurmountable challenge. Of the methods that have been proposed to detect deepfakes the majority depend on deep learning algorithms.

The threat of deepfakes is being taken seriously however. The United States Defense Advanced Research Projects Agency (DARPA) started a research program in Media Forensics aimed at creating effective countermeasures to the threat of deepfakes. DARPA is not the only institution taking this seriously. In 2016, Google, Facebook, Amazon, IBM and Microsoft formed the Partnership of AI which was dedicated to advancing the public's understanding of Artificial Intelligence as well as creating standards for future researchers in the area to adhere to. It is clear that deepfakes create a significant threat to society and the ability to detect them is of paramount importance.

1.2 Problem Statement

The ease of use and access to powerful online tools make generating synthetic media easier than ever. This is also fueled by the availability of significant computing power brought about by cloud computing and cheaper and more powerful processors. Deepfake technology could be used to destabilise governments, undermine political processes, manipulate financial markets or ruin reputations.

The ubiquitous use of social media means that the spread of deepfakes can far outpace the detection and alerting capabilities that currently exist. To address the threat effectively, ongoing research, developments and innovation into countermeasures is absolutely necessary.

1.3 Objectives

In this thesis I will attempt to review the current state of deepfake detection technologies. I will analyse the societal and legal impact of deepfakes. I will identify the gaps in existing research and propose areas for future work and finally I will develop a comprehensive framework for mitigating the threats posed by deepfakes.

2 Related Work

2.1 Deepfake Detection Technologies

2.1.1 Machine Learning Models

The paper by Nguyen et al. (2019) titled ‘Deep Learning for Deepfakes Creation and Detection: A Survey' offers a comprehensive review of the state of the art of deepfake technology and detection capabilities. Nguyen et al. (2019) explored the methods of generating deepfakes as a possible avenue to understand opportunities of detection. The authors particularly focused on the use of GANs to create realistic digital content. GANs are very effective as they use two competing models—a generator that creates deepfakes material and a discriminator that tries to distinguish between real and fake data. This adversarial process enables the creation of highly realistic videos and audio.

Nguyen et al. (2019) explore the use of traditional image processing techniques as a means of identification as well as looking into other machine learning approaches to detect synthetically generated content.

Figure SEQ Figure * ARABIC 1: GAN Architecture Diagram

Nguyen et al. (2019) also discuss using Convolutional Neural Networks (CNNs) as a potentially effective way of analysing deepfakes by capturing subtle identifiers that distinguish between fake images and real ones. The University of Hull (Tonkin) recently presented a similar approach to the detection of deepfakes by analysing the reflective patterns in the eyes of synthetically generated content. Using CNNs as a method to detect these tiny clues in synthetic media may unlock the secret to quickly and accurately detecting deepfake content. In reality, the effective and accurate method of detecting deepfake content may only be achieved by using a multi-layered or iterative approach to detection.

Nguyen et al. (2019) discuss the challenges and future directions in deepfake detection research. They discuss how the development in the area of deepfake creation also creates opportunities in the development of deepfake detection. They identify a major concern regarding the scalability and resiliency of deepfake detection techniques. The ability of generative technology to adapt and change may make the creation of detection techniques obsolete incredibly quickly. In this case the analysis of synthetic media may not be sufficient in detecting deepfakes and the use of alternative forensic analysis and detection techniques is almost certainly going to be required to support detection capabilities. Deepfake technology is a societal problem and as such, collaboration across academia, industry and government to create rules and guidelines is of vital importance.

2.1.2 Digital Forensics and Steganalysis

Stenography has roots that can be traced back to ancient Greece. Steganography is the technique of hiding information within another image or message. Steganalysis is the practice of detecting this hidden information. As steganographic techniques become more sophisticated, the need for advanced detection methods has grown. This has led to the development of "rich models" which are capable of detecting a wide array of steganographic techniques.

Fridrich's 2012 paper, "Rich Models for Steganalysis of Digital Images," presents a framework for steganalysis using rich models. This work is extremely important in advancing the detection capabilities for various steganographic methods. To understand how steganalysis can become a key component in the detection of deepfakes it is important to understand the components of the model which are outlined below.

Feature Extraction

Fridrich’s rich model relies on the detailed and comprehensive extraction of features from digital images. Fridrich’s paper goes on to describe the necessity of capturing a diverse array of features to identify subtle manipulations introduced by steganographic techniques. He puts forward a number of advanced features that could have important uses. These features include:

Statistical Measures: Basic statistical properties such as mean, variance, skewness, and kurtosis, which provide insights into the overall distribution and behaviour of pixel values within an image. E.g. skewness analyses the asymmetry of data distribution. A skewed distribution could be one indicator of steganographic interference. Likewise kurtosis measures data with more extreme outliers which could also detect steganographic content.

Co-occurrence Matrices: These matrices capture the frequency with which pairs of pixel values occur at specific spatial relationships. Co-occurrence analysis techniques could help reveal patterns and textures that can indicate the presence of hidden data.

Higher-order Statistics: Advanced statistical features that go beyond simple pairing relationships by capturing complex dependencies and anomalies across multiple pixels. Using features like Markov Random Fields or other sophisticated models that describe the spatial structure of images.

Local Binary Patterns (LBP): Detection capabilities of individual pixels in relation to their surrounding pixels to help detect subtle changes introduced by steganography.

Submodels

Rich models are not standalone models; instead, they are composed of several submodels which target different aspects of the image data. These submodels work together to provide a holistic analysis:

Spatial-domain Features: Focus specifically on the pixel values directly and their relationships with the pixels around them thus capturing changes made in an image.

Frequency-domain Features: By analysing the image based on patterns and how they change across the image, such as the Discrete Cosine Transform (DCT) or the Discrete Wavelet Transform (DWT), it may be possible to detect manipulations in frequency coefficients.

Hybrid Descriptors: By combining both of the above features we can take advantage of the strengths of both, providing a more resilient, in-depth and comprehensive form of detection.

Dimensionality Reduction

When working with a lot of features (high dimensionality), it is crucial to reduce the dimensionality so as to manage computational complexity and prevent the model becoming too tailored to the training data that it is no longer effective on new data. Techniques such as:

Principal Component Analysis (PCA): This form of analysis reduces the dimensionality by transforming the feature set into a new space defined by the new features while also capturing most of the variation of the original data.

Linear Discriminant Analysis (LDA): This analysis aims to maximise the separation between the different categories of data. It also helps by keeping the features that are best at distinguishing between these categories.

t-Distributed Stochastic Neighbour Embedding (t-SNE): t-SNE is mainly used for visualising data in 2 or 3 dimensions. It is particularly effective for exploring and understanding the structure of data. It does this by showing how the data points are represented. This is extremely useful for exploratory analysis and validation.

Machine Learning for Classification

In this context the detailed information that was extracted from the images are used as input for machine learning. This helps the machine learning algorithms to understand the difference between unaltered images and images with hidden data. Some of the more commonly used inputs are:

Support Vector Machines (SVM): Effective in handling data with many features (high-dimensional spaces) and known for their robustness and accuracy. Their strength becomes particularly evident in the binary classification tasks.

Ensemble Methods: Techniques such as Random Forests or Gradient Boosting combine multiple weak classifiers into one strong predictive model which effectively enhances detection performance and robustness.

Deep Learning Approaches: Although not traditionally used in Fridrich's models, modern adaptations could incorporate convolutional neural networks (CNNs) for automatic feature extraction and classification which could potentially improve accuracy and efficiency.

Fridrich's rich models have had a significantly positive effect in the area of steganalysis and have contributed substantially to elevated detection rates in the area. The layered approach and the extensive feature set of rich models allows for nuanced discrimination between cover and stego images, leading to higher accuracy and fewer false positives compared to models that preceded it.

Rich models partially solve the concern of Nguyen et al by creating a scalable and resilient detection model. One of the standout features of rich models is their versatility. They are effective against a broad range of steganographic techniques, including those that alter the pixels directly (the spatial domain) and those that manipulate images (the frequency domain). This adaptability makes rich models suitable for diverse steganalysis scenarios.

The modular structure and approach of rich models facilitates easy adaptation and enhancement, a key requirement into the detection of deepfakes. As new steganographic methods are developed, new features or submodels can be integrated into the existing framework, ensuring that detection capabilities remain robust and up-to-date. This provides an important point in developing countermeasures that have to rapidly evolve and improve to detect rapidly changing deepfake technologies.

Similarly to the technology behind the creation of deepfakes, rich models are computationally intensive. The process of extracting and processing a large number of features can be resource-intensive, which may be a limiting factor in the feasibility of deploying these models to capture deepfakes in a real-time environment.

Having a large feature set can help capturing intricate details however it can also create a risk of overfitting. Ensuring that the training data is sufficiently diverse can prevent overfitting is paramount. Cross validation and regularisation techniques are critical to ensure that this is prevented and that the model is effective against new data.

As defensive techniques continue to evolve the adversarial techniques will shift to adapt to new countermeasures creating an age old cat and mouse game of detection and evasion. This will create an environment of continuous improvement in both deepfake creation and deepfake detection.

There is no doubt that Fridrich's rich models have given the most important contributions in the field of steganalysis. They offer a strong framework to detect hidden information in digital images. Rich models are at a distinct disadvantage due to the dynamic development of steganographic methods and the computational power required to keep pace with this development. This aside, rich models should be considered a key tool in the detection of deepfakes due to their high detection rates. Rich models could play an important role in the detection of deepfakes by performing a deeper analysis of a digital sample that other detection techniques results are conflicting. Further studies into rich models and efforts to increase the efficiency of detection may also be key to addressing the computational barriers that currently exist.

2.1.3 Blockchain and Digital Watermarking

The erosion of the integrity of digital content is a key concern introduced by the creation of deepfakes. Considering one of the main features of blockchain technology is the integrity and immutability of data, it is unsurprising to see studies linked to the use of blockchain technology to combat the scourge of deepfakes. Indeed, the 2021 paper by Qureshi, Megías, and Kuribayashi (2021) introduces a novel approach to this challenge by combining digital watermarking with blockchain technology.

Digital watermarking is a technique where a signature is embedded into a video signal. This signature is resilient to modification by humans or computers and can be used to verify the authenticity and integrity of video content at a later date should its provenance be questioned. The paper by Qureshi et al. (2021) emphasises this technique as a way to trace and verify the integrity and authenticity of video content.

For this technique to be accepted as a trusted and effective source of deepfake detection the watermark embedding process needs to be robust. Qureshi et al (2021) suggest that the watermark is embedded into the video frames during the production phase. The alteration of the original content would then result in the removal, damage or exclusion of the embedded watermark thus allowing detection technologies to correctly identify the content as being modified.

The authors suggest that the creation of fragile watermarks based on the speech content of the media is generated. This speech content can then be hashed and embedded as a fragile watermark. The metadata of the created watermark is then committed to the blockchain for tamper-proof recording and verification.

By utilising algorithms to insert a watermark into less perceptible parts of the video, such as the least significant bits of pixel values the process can ensure there is minimal visual impact. Likewise, including markers to facilitate the accurate extraction of the watermark even if the video undergoes common processing operations like compression or resizing.

Robustness and Imperceptibility

When creating and embedding a fragile watermark it is important that this watermark is resilient to typical computational manipulation that is common for digital media. It must withstand compression, conversion or minor alterations all while maintaining its invisibility to viewers. Some methods put forward by Qureshi et al. (2021) include Redundant Embedding which consists of embedding the watermark information redundantly across multiple frames or within different spatial regions. They also suggested the use of Adaptive Algorithms. This technique uses adaptive algorithms that adjust the embedding strength based on local video content, ensuring the watermark is invisible while maintaining resilience.

Detection Mechanism

When a video is suspected of being a deepfake, the embedded watermark is extracted and analysed. The detection mechanism includes Watermark Extraction and Integrity Analysis. This consists of techniques for accurately retrieving the embedded watermark from the video, even after it has been subjected to various manipulations and comparing the extracted watermark with the original information to verify the video's authenticity. If the watermark is intact, the video is considered authentic; if altered, the video is flagged as potentially manipulated.

Blockchain for Authentication

To enhance the trustworthiness of the watermarking system, the paper integrates blockchain technology. Blockchain provides a decentralised and immutable ledger for storing watermark-related information, ensuring that the authentication process is secure and tamper-proof. Key aspects of blockchain in deepfake detection are the Decentralised nature and Immutability. By eliminating central points of failure by distributing the storage of watermark data across a network of nodes. Blockchain can provide a robust and resilient verification network while also ensuring that once watermark information is recorded on the blockchain, it cannot be altered, providing a reliable audit trail for video authenticity.

Impact and Advances

The combination of digital watermarking and blockchain technology provides a high level of security and integrity. This dual approach makes it extremely difficult for attackers to alter the watermark without detection, thereby safeguarding the authenticity of video content.

Blockchain effectively addresses the scalability concern raised by Nguyen et al. (2019). By committing the watermark details to the blockchain the information can be stored and accessed globally, mitigating risks associated with centralised systems and enabling widespread adoption across various platforms and industries.

Using blockchain to store and verify data is a practical approach to detecting deepfake content. It leverages existing and resilient technology. By embedding watermarks during production and verifying them before distribution, stakeholders can ensure the integrity of their media.

Criticisms and Limitations

Certain blockchains have faced criticism in the media for the amount of power they consume to stay operational. In recent years the emergence of more efficient blockchains (proof of stake) have proven to be as effective and resilient as more power intensive technologies. By leveraging efficient blockchains to verify embedded watermarks we can effectively overcome the computational overhead that comes with other solutions.

To ensure the integrity of media the watermark needs to be embedded into the media during the production phase. This limits watermarking to new and future media and excludes media that was created in the past.

Embedding watermarks into already created media may become a trivial process however embedding watermarks into live video streams could be a more challenging prospect. The analysis of video content to determine synchronisation markers for the addition of watermarks may result in weaknesses or gaps that can be targeted by adversaries to undermine the entire process.

3 Analysis of Detection Technologies

To ensure detection capabilities keep pace with deepfake technology it is imperative that detection mechanisms become more sophisticated over time. By employing deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) there is a fighting chance of staying abreast of deepfake technologies to quickly, easily and accurately detect fake content. This chapter explores the technical intricacies of these models, their effectiveness in detecting deepfakes, and the challenges in training them, including robustness against adversarial attacks.

Convolutional Neural Networks (CNNs) in Deepfake Detection

CNNs are a subset of deep learning models extremely well-suited to image and video analysis. They are made up of three different layers that work together to produce a detailed analysis of an image or video. The layers in a Convolutional Neural Network consist of a Convolutional layer which scans images to detect features such as textures and edges in a digital specimen. There are pooling layers that compress the data while preserving important information about the specimen. Compressing the data allows for faster processing. Finally there are fully connected layers that combine all the features learned by the previous layers to come to a final conclusion regarding the authenticity of the digital artefact.

Convolutional Neural Networks are extremely effective at detecting deepfakes content due to their ability to learn, adapt and recognize complex patterns. These patterns can be used to differentiate real content from fake content. Studies have shown that CNNs can detect subtle artefacts and inconsistencies introduced during the deepfake generation process (Dhar).

Image-based CNNs are models that analyse individual frames of videos to detect inconsistencies in the frames that are typical of inconsistencies that occur in deepfakes. Techniques like XceptionNet have been particularly successful, achieving high accuracy in various benchmarks (Rossler et al., 2019).

Video-based CNNs work by processing sequences of frames. These models can capture changes over time that single-frame analysis might miss. This approach enhances detection robustness, especially when combined with detecting changes and patterns that occur over time (temporal dynamics).

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks

Recurrent Neural Networks (RNNs) were specifically designed to handle sequential data, which makes them ideal for analysing video data where temporal dynamics are crucial. RNNs face unique challenges when analysing video data. One of these challenges is the vanishing gradient problem [Appendix A]. This problem is addressed using Long Short-Term Memory (LSTM) networks, a type of RNN that can remember information for longer periods.

This is important as it allows the network to understand and make sense of sequences and patterns where the context is spread out over time (effective in video analysis).

RNNs and LSTMs are effective in deepfake detection as they can analyse and identify inconsistencies over time. They are excellent at capturing motion anomalies that are indicators of deepfake manipulations.

> Temporal Analysis: LSTM-based models can detect deepfakes by analysing the consistent progression over time in videos. They can identify unnatural movements and inconsistencies in facial expressions that occur across individual video frames.
> When we combine CNNs for spatial feature extraction with LSTMs for temporal analysis we can enhance the detection capabilities of the network. While more resource intensive it does provide a more comprehensive approach to identifying deepfakes.

Figure SEQ Figure * ARABIC 2: Architecture of a Recurrent Neural Network

Input Layer: The sequence of inputs is fed into the network.

Hidden Layers: The hidden layers represent the RNN's internal memory. The hidden state from each layer is fed back into itself, allowing the network to maintain a temporal understanding of the input sequence.
Output Layer: This layer provides the final output after processing the entire input sequence.

The diagram shows the flow of data through the RNN, including how the hidden states are looped back into the hidden layers, which is a key feature of RNNs enabling them to handle sequential data.

Figure SEQ Figure * ARABIC 3: Architecture of a Long Short-Term Memory Network

Input Gate: Controls the extent to which the new information is added to the cell state.
Forget Gate: Decides what information from the cell state should be discarded.
Cell State: The memory of the network that carries information across different time steps.
Output Gate: Determines the output based on the cell state and the input.
Hidden State: The output that is passed to the next time step and the next layer in the network.

3.1 Challenges in Training Deepfake Detection Models

To ensure that deep learning models are trained effectively, and to prevent bias, it is important that they are trained using large and diverse datasets. The models must be exposed to a wide variety of real and fake content to help them learn effectively. Publicly available datasets such as FaceForensics++ have been instrumental in providing learning material, but there is a constant need for new and diversified data to keep pace with evolving deepfake technologies.

Adversarial attacks pose a significant challenge to deepfake detection models. These attacks involve poisoning inputs specifically designed to trick the learning models. Ensuring that learning models are resilient to these types of attacks is paramount to their effective operation. There are two approaches that we can follow to protect the integrity of detection models. Firstly we can incorporate adversarial training into the training process. This can aid the model into identifying and detecting such attacks and in turn taking protective measures against the attack.
As discussed earlier, the compute power required to train and utilise deep learning models can be prohibitive meaning that this capability may only be deployed by well funded sources.

Unlike the ease of access and use of deepfake creation tools, detection capabilities remain elusive to casual internet users. This can significantly hamper widespread adoption.

3.2 Digital Forensics Techniques

By utilising digital forensics we can interrogate the metadata of digital content to detect anomalies that could provide clues as to the authenticity of the specimen. Spectral analysis of images or sound may also provide insights to the investigator. In this chapter we will look at the techniques and methodologies of digital forensics and their effectiveness in detecting deepfakes.

Metadata is information about data. This information might include details such as file creation dates, modification history, camera settings, geographic location of a photograph and more. This information can play an important role in detecting inconsistencies in digital media. Through the examination of attributes like creation and modification dates, file formats, and software used we may be able to reveal signs of manipulation. For example, discrepancies in timestamps or the use of editing software can indicate that a video has been altered. Likewise the examination of camera metadata such as information about the device that captured the video, including make and model, GPS location, and exposure settings, can help verify the authenticity of the image. Inconsistent metadata can signal that the media has been tampered with.

Metadata analysis can be performed easily and without requiring significant processing power. It can be a very useful first step in deepfake detection. By identifying anomalies in the metadata, forensic analysts can identify and flag suspicious content for further investigation. Integrating metadata analysis into commonly used technologies such as browsers or smart devices could provide an initial layer of deepfake detection. This technology currently exists in the form of a browser plugin called TinEye that can detect manipulation in an image.

Forensic Analysis of Physiological Signals

Physiological signals refer to the subtle movement in facial expression such as blinking, pulse, mouth movements, facial micro-expressions or tics. These signals can be used to detect deepfakes, as they are difficult to replicate accurately with current deepfake technology. Natural eye movements in response to situational proximities can be difficult to recreate accurately by current deepfake technology. Abnormalities in these natural movements can be indicators of manipulated or synthetically generated content. Likewise the subtle changes to skin tone caused by blushing, temperature change or stress can create other signals that can be used to detect image manipulation.

Advanced forensic tools in this area are not new. In fact techniques to detect and amplify subtle physiological signals have been around for over ten years. Wu et al. (2012) introduced a method for amplifying subtle changes in video, known as the Eulerian Video Magnification. This method consists of applying spatial decomposition and temporal filters to video frames and the second step is to magnify these frames. This technique enables the visualisation of an event that may be so minute as to generally go unnoticed. Amplifying subtle changes in physiological reactions can introduce detection markers to video content allowing manipulations to be more easily detected.

Challenges in Forensic Analysis

As with Deep Learning Models, digital forensics is also at the mercy of the quality and quantity of data that is available to it. Low resolution images and videos are less likely to contain the forensic matter and physiological signals that are required for a highly confident result. Likewise a low resolution specimen may lack the data that will show the micro-expressions that can provide clues and indicators of manipulation.

Considering the Eulerian Video Magnification technique is not new there is a significant chance that adversarial techniques to defeat it have already been identified. It is already known that low quality samples render this technique invalid. Coupled with the technological advancements in deepfake technology, the ability to successfully recreate these physiological micro-expressions may also render this method defunct.

3.3 Blockchain and Digital Watermarking

The development and application of blockchain technology has seen significant growth in the past ten years. This technology, coupled with the advancement in digital watermarking offer promising solutions for ensuring the authenticity and integrity of digital content. Here, we will explore how we can use these technologies along with existing systems to provide a robust defence against deepfakes. By layering the strengths of digital watermarking with the immutability of blockchain we can effectively prove the integrity and provenance of digital images and video content to refute the misinformation that may be proliferated by deepfake material.

Blockchain Technology

Blockchain technology is decentralised distributed immutable digital ledger. It records and stores transactions in a way that the information cannot be changed once committed to the blockchain. This immutability of data provides an excellent way to ensure the integrity and security of a record thus ensuring a permanent and verifiable record of the committed data. By committing the details of a fragile watermark of an unmodified authentic digital sample to the blockchain it can be used as a reference against all other copies of that digital sample to ensure that it has remained unmodified.

The decentralised nature of blockchain can also provide a robust defence against an attack on the system. Adversarial techniques would need to target each node of the blockchain to effectively disrupt it. This would require considerable resources and effort to achieve. Blockchains are also resilient to DDos attacks. To commit data to a blockchain a fee must be paid. An entity that attempts a DDoS attack will need to pay for every transaction it commits which could become very expensive and make a disruptive attack like DDoS ineffective. Decentralisation also ensures that the blockchain is not dependent on one node to continue to function which provides resiliency and availability, both important attributes for the instant verification of data.

Blockchain can be used to record the provenance and authenticity of digital media, ensuring that any alterations are easily traceable. By storing hash values of original content on the blockchain, any subsequent changes can be detected by comparing the current hash with the stored value. Blockchain can provide a transparent and immutable record of a digital file's history, including its creation, modifications, and distribution thus creating a digital trail that can be analysed by forensic methods.There are several projects already leveraging blockchain technology to provide immutability of data. For example MediLedger uses the blockchain to verify pharmaceutical products which allows for traceability of pharmaceutical supplies.

Digital Watermarking

Digital watermarking involves embedding information into digital media that can be used to verify its authenticity and detect tampering. Watermarks are designed to be imperceptible to users while being detectable by specialised tools. Unlike forensic analysis techniques that are affected by the quality of the digital specimen, watermarks should not affect or be affected by the usability or quality of the digital content. Likewise, watermarks should withstand common manipulations, such as compression, resizing, and minor editing.

Digital watermarking can be used to embed verification data into videos, ensuring that any alterations can be detected by analysing the integrity of the watermark. Watermarks can contain metadata about the content's origin, creator, and modification history. Any alterations to the video will disrupt the embedded watermark, signalling potential tampering.

Integration of Blockchain and Digital Watermarking

Combining blockchain and digital watermarking provides a multi-layered defence against deepfakes, enhancing the security and authenticity of digital content. By embedding unique watermarks into the content at the time of production the provenance and integrity of the content can be verified. This value can then be committed to the blockchain creating an immutable record of authenticity. The watermark can continue to be verified for its entire lifecycle ensuring that a true and accurate account is committed to history..

Using blockchain to store immutable identifiers has very obvious benefits but it also has its limitations. Blockchain technology is relatively new. As the blockchain continues to grow there is no way to predict the future problems it may encounter. These issues may affect the verification of historical data and allow for alternative theories to gain prominence. There is also the concern regarding the interoperability of different blockchains and the issues that may cause.

4 Societal and Legal Implications

We introduced this paper talking about the benefits and implications of deepfake technology. In this chapter we will discuss in more detail the social and legal implications for society as we know it. The use of deepfakes poses unique moral and ethical challenges. These challenges will shape modern society for years to come. Putting the positive uses of the technology to one side and focussing on the negative we will explore the implications of deepfakes, and explore the need for comprehensive strategies to guard against their adverse effects on information integrity, privacy, and legal frameworks.

The spread of misinformation and disinformation, made exponentially worse by the growth and popularity of social networks, can be a powerful weapon when combined with deepfakes. The ability to create fake content and post it to social media is effortless. By the time the false information is proven to be fake it will have already spread to multiple different sources. We have seen the growth in deepfakes to spread disinformation and misinformation over the years. It has been weaponized by political campaigners to sway the electorate. It has been used to undermine political opponents and it has been used to create false narratives. Dispatches, the Channel 4 show, recently performed an experiment where deepfakes were used to simulate political leaders of both Conservative and Liberal party leaders ("Deepfake audio of Sir Keir Starmer released on first day of Labour conference"). In this experiment left and right leaning members of the were shown different deepfakes which challenged tr allegiances. When the experiment was revealed to them a majority confessed that the very real feelings they had from watching the content overrode the revelation that the content was fake. This admission has serious consequences as we might find that the damage deepfakes can do cannot be undone regardless of proof to the contrary.

We have seen in recent times the undermining of the media by political challengers. By challenging the integrity of digital records, it becomes possible to brand any disparaging content "fake news". It undermines the content produced by media outlets and introduces bias where the public start to believe only the content that aligns to their values, branding any alternative viewpoints to be false. This creates another significant challenge regarding the proof of authenticity. If digital content can be easily labelled as fake news, the credibility of media creators who use watermarks to establish authenticity can also be undermined. This results in a loss of trust in both the creators and the documented real-world events.

The proliferation of deepfakes can also lead to a "post-truth" society where the distinction between reality and fiction becomes increasingly blurred. This erosion of trust affects not only media and journalism but also interpersonal relationships and social norms. As people become aware of the potential for media manipulation, they may become more cynical and less likely to trust information, regardless of its authenticity.

Deepfakes can be used to create non-consensual explicit content, causing severe psychological and reputational harm to individuals. As highlighted by Westerlund (2019), such content can be used for blackmail, extortion, and harassment, leading to long-term emotional and social consequences for victims. The unauthorised use of personal images and videos raises significant privacy concerns and highlights the need for robust legal protections. In recent years major film studios have attempted to convince actors to sign away their rights to their likeness meaning that studios can create digital content using their likeness without having to employ or pay them. By losing control of their own image it could be used in any format leading to psychological or reputational damage. Deepfakes have also created moral dilemmas regarding the legality of synthetically generated pornographic media that depicts minors or other subjects. While no minors may have been harmed in the creation of the content, the subject matter of the content itself has significant implications on the evolution of society if it is not carefully managed.

Legal Implications

Technology is a fast moving sector and the rapid advancement of deepfake technology far outpaces current legal frameworks. This divide presents unique challenges for lawmakers. Maras and Alexandrou (2019) discuss the difficulties in authenticating video evidence in legal contexts, given the sophisticated nature of deepfakes. Existing laws already struggle to adequately address the creation and distribution of malicious deepfakes. The use of deepfake technology needs regulators to step in and create frameworks, guidance and laws governing the technology and its applications. Indeed Chesney and Citron (2019) advocate for the implementation of laws specifically targeting the malicious creation and distribution of deepfakes. They argue that such legislation should criminalise harmful uses and provide clear frameworks for victims to seek redress. They suggest that lawmakers criminalise the production and publication of deepfakes used for malicious means as well as harmonising legal standards governing deepfake technology across the globe to address the ubiquitous threat that deepfakes bring.

The responsibility and success of deep fake detection mechanisms lies in the ability of law makers, academics and technology to collaborate and support efforts for the greater good. The development of advanced technology should be supported and financed by the government to ensure that detection technology keeps pace with deepfake creation technology. The development of protocols and standards in digital forensic techniques will be paramount if their findings are going to have any standing in a court of law.

Regulation is absolutely necessary to protect against the misuse of deepfake technology but it is also equally important to balance the regulations with the potential benefits that the technology brings. Excessive legal or regulatory barriers could stifle innovation and legitimate uses in fields like medicine, entertainment, education, and creative arts. It is imperative that policymakers ensure that regulations are flexible enough to accommodate the positive applications of deepfakes while preventing their abuse.

Ethical Considerations

A report published in 2019 by a Dutch company called Deeptrace estimated that 96% of the deepfakes that were published online at that time contained pornographic images, Ajder et al

(2019). The sheer volume of deepfake material poses a privacy and consent dilemma. The creation of non-consensual deepfakes, especially deepfakes containing explicit content , can cause severe emotional and reputational damage to an individual. Additionally the use of images of minors for a similar purpose could cause severe emotional distress and psychological harm. It is not difficult to see from the misuse and possible repercussions the need for ethical frameworks and their emphasis on the importance of obtaining consent from individuals before using their images or videos in any synthetic media.

Educating the public on the danger and harm of deepfakes is crucial to limiting the damage of deepfake material. Teaching members of the public to critically assess any media by looking for common indicators and characteristics of deepfake material can help contain the spread of misinformation. Likewise, encouraging a sceptical and analytical approach to consuming media can also help prevent negative consequences. Finally having a clear and accessible channel for reporting suspicious content can make the identification and awareness of fraudulent content more efficient. Media literacy campaigns and training programs for journalists and media professionals can empower individuals to critically evaluate the authenticity of media content.

It is clear that an unregulated approach to deepfake detection will not work. The greatest chance of success is to force content makers, news outlets and social media platforms to integrate deepfake detection technologies into their platforms. This approach relies heavily on the technology and the ability to work seamlessly and accurately to improve confidence in the system and provide immutable evidence of the authenticity of digital content. A transparent approach to the technology is also required to ensure the public's trust in it. Once deployed, deepfake detection solutions must be free from the influence of governments, law makers, policymakers or others that could benefit from its manipulation. A concerted effort to supplement this technology with public awareness campaigns, media literacy campaigns and interdisciplinary collaboration will give society the greatest chance of counteracting the negative effects of deepfakes and the chaos that they can bring.

4.1 Misinformation and Disinformation

The 2019 article by Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics," published in ForeignAffairs.com, addresses the emerging threat of deepfake technology and its implications for world politics and the integrity of information. This review incorporates additional insights from a recent experiment by Channel 4’s Dispatches (Channel 4) on the use of deepfakes in political contexts, highlighting the urgent need for effective countermeasures. The study found that deepfakes can create very real and convincing false narratives which undermine public trust in media and other digital information sources. The recent use of deepfakes in politics, such as the Channel 4 experiment, demonstrates how these technologies can influence public opinion and disrupt the democratic process.

Deepfakes depicting individuals in compromising situations can also be used for blackmail and extortion, affecting the personal reputations, mental wellbeing and have other real social implications. Additionally, the overall awareness of deepfakes can lead to a general erosion of trust in digital media, eroding the credibility of genuine media content.

The authors argue that deepfakes represent a new front in the disinformation war with significant implications for global politics. Deepfakes can interfere in elections by creating false statements or actions attributed to candidates, influencing voter opinions and disrupting democratic processes. False videos of world leaders making inflammatory statements or engaging in illegal activities can exacerbate tensions between countries and lead to diplomatic crises. Authoritarian regimes can use deepfakes to spread propaganda, discredit opposition, and manipulate public perception.

Chesney and Citron discuss various legal and policy measures to combat the threat of deepfakes:

> Legislation: Implementing laws that specifically target the malicious creation and distribution of deepfakes, criminalising harmful uses and providing frameworks for victims to seek redress.
>
> Technological Solutions: Developing advanced detection tools to identify deepfakes, involving collaboration between governments, tech companies, and researchers.
>
> Public Awareness and Education: Increasing public awareness about deepfakes and educating people on how to critically evaluate media content can mitigate the effectiveness of disinformation campaigns.
>
> International Cooperation: Fostering international cooperation to create standards and practices for addressing the global threat of deepfakes, including intelligence sharing and jointly developing detection technologies.

While the article provides a comprehensive overview of the threats posed by deepfakes and potential responses, some limitations are noted:

> Detection Challenges: The continuous advancement of deepfake technology makes it a moving target for detection tools. The arms race between deepfake creators and detectors is ongoing, with no definitive solution in sight.
>
> Enforcement Issues: Enforcing laws against deepfakes can be challenging, especially in jurisdictions with weak legal frameworks or where the origin of the deepfake is outside the jurisdiction.
>
> Balancing Regulation and Innovation: There is a need to balance regulation with the potential benefits of deepfake technology, ensuring that legitimate uses are not stifled.

Chesney and Citron’s article, supported by recent developments highlighted in the Dispatches piece, outlines the significant threat posed by deepfakes to information integrity and political stability. By identifying potential risks and suggesting a multifaceted response involving legal, technological, and educational measures, the research contributes significantly to the discussion on navigating the challenges of the post-truth era. Their work emphasises the importance of proactive measures to protect the credibility of media and maintain public trust in digital content.

4.1.2 Privacy and Consent

The unauthorised use of individuals' images and videos in deepfakes raises significant privacy concerns. Non-consensual explicit content created using deepfakes can cause psychological and reputational harm, as highlighted by Westerlund (2019). The use of deepfakes also brings moral considerations such as the depiction of minors or other vulnerable people in deepfake images. As discussed earlier chapters, the pace of the legal and justice system is severely lagging behind that of technological innovation rendering existing laws in privacy and consent severely lacking. The personal consequences to the use of unauthorised depictions of an individual's identity can be severe therefore gaining and giving consent must be properly considered. Current data protection legislation could be modified or used to protect the data of an individual in this case. Unfortunately the reputational damage may already be done.

Policymakers will need to put careful consideration into the protection of a human likeness and give people the opportunity to withdraw consent where previously given. Punitive measures must be a sufficient deterrent to those that do not comply with the protection laws.

4.1.3 Legal Challenges

Maras and Alexandrou (2019) examine the challenges of authenticating video evidence in the context of deepfake technology. The invention of the smartphone brought with it the saturation of video content. Essentially allowing any event to be captured instantaneously. Consider the footage of a serious crime being submitted into evidence in a criminal trial. The legal system needs to have faith that the integrity of the footage is unaltered and reflective of the events it captured. This evidence might be the difference between an acquittal or a conviction. Without being able to prove the provenance of the footage it's legitimacy could be undermined and thrown out as evidence having devastating consequences to the prosecution or defence. A robust forensic framework must be established to ensure that footage can be verified forensically to provide an unwavering faith in the authenticity of the media.

5 Future Directions and Proposed Framework

Through the comprehensive literature review this paper explores different methods for deepfake detection using machine learning, digital forensics, digital watermarking and blockchain technology. One common theme was constant in all of these areas and that is the requirement of significant compute resources to accurately detect deepfake content. This intensity of analysis may prevent the mass adoption of deepfake detection capabilities.

To help foster mass adoption and acceptance of deepfake detection an instantaneous checking mechanism is required. This check may utilise multiple different technologies that when combined allow for immediate and accurate detection.

The challenge is clear, with humankind uploading an estimated 14 billion images daily it is impossible to check the authenticity and integrity of each image prior to posting. The requirement for an initial, resource light, integrity check or rating system may reduce the amount of content that is flagged for further analysis.

A legal or regulatory mandate requiring social media platforms to perform integrity and authenticity checks on media being shared and uploaded to their platforms may also help to prevent the spread of misinformation and disinformation through their platforms. These checks could incorporate an initial pass or screen which, depending on the content, is passed into a CNN or RNN for further screening.

Hardware manufacturers could embed immutable signatures to the media they produce that can be verified through blockchain technology. This would at least ensure that the media being shared was produced using imaging hardware and not produced using a deep learning algorithm.

Figure SEQ Figure * ARABIC 4: Image analysis and detection flow

The image above suggests an integrity data flow for an image being posted to social media.

The premise of this data flow is that an image will go through the least resource intensive checks earlier in the process. At each stage the image will be given an integrity rating.

Once the integrity rating has reached a certain score it will be posted as is or it will be posted with a warning stating that the image has failed integrity checks and should be treated with caution.

Using a framework such as this ensures that legitimate images get through the check quickly whereas doctored or faked material does not. This provides an efficient and less intensive check to deliver online digital content.

This framework has several areas for future study such as:

Real-time Steganalysis -

> Efforts into reducing the computational complexity of existing algorithms using techniques such as quantization and pruning.

Development of lightweight models that can run on mobile and edge devices, facilitating real-time applications.

Practical Applications of real time Steganalysis could enhance social media monitoring tools to detect steganographic content in images and videos uploaded in real-time and also to improve digital forensic tools for quick analysis and identification of hidden data in digital media.

Enhancing Digital Watermarking and Deepfake Detection

The studies of Qureshi, Megías, and Kuribayashi provides us with further areas of study most interestingly to:

Enhancing Watermark Robustness by developing algorithms that embed watermarks within media files, making them resistant to tampering and the use adaptive watermarking techniques that adjust based on the content characteristics.

The developments of real-time detection tools that can quickly verify the presence and integrity of watermarks in media files as they are shared or streamed. This can aid in the application of an integrity rating.

Intuitive Tool Development:

Providing detection tools to consumers of media is an effective and resource efficient way of identifying deepfakes (where detectable).

Design verification tools with simple user interfaces that require minimal technical knowledge to operate and ensure tools provide clear and actionable feedback, helping users quickly determine the authenticity of content.

Develop a check to prevent false positives by detecting continuous abuse of these tools.

Accessibility and Trust Promotion:

Make verification tools widely available through web and mobile applications, encouraging broad adoption.

Promote the use of these tools through educational campaigns and partnerships with content creators and platforms.

This research offers a robust solution for video authentication, addressing limitations like computational overhead and dependency on initial watermarking. Future advancements will build upon this framework, enhancing digital media security.

6 Conclusion & Final Thoughts

It is clear that the threat of disinformation and misinformation perpetuated by synthetic media poses an immediate risk to societal norms. Allowing deepfake technology to go on unregulated is a significant risk to the integrity of information sharing and public discourse.

The technology, if weaponised, has the real capability to unsettle governments, disrupt financial markets and effect legal trials. The negative consequences emerging from the abuse of deepfakes cannot be understated. It is a threat to the accuracy and dissemination of legitimate and historic events. In the wrong hands it can undermine the integrity of the justice system across the globe. The ramifications of a successful deepfake campaign can change the course of history. The trajectory of which is based on fictitious events underpinned by synthetic media accurately depicting fantasy. There are positive applications to deepfake tech however the consequence of a negative application of the technology far outweighs the positive applications.

Mark Twain once said that "A lie can travel halfway around the world while the truth is still putting on its shoes". This is certainly true with the invention of the internet and the proliferation of social networks. The ease of sharing content across platforms has never been easier. Social media platforms do not discriminate between fake and real content meaning a doctored or fabricated image or video can be shared to millions over the course of a few hours. As of today, social media platforms have not managed to get a handle on the publication of false content although X (formerly Twitter) has made attempts to notify users of fake or misleading content [Appendix B]. Ensuring that content is quickly and accurately flagged as fake is paramount to preventing the spread of misinformation and disinformation.

To accurately depict the challenge that countermeasures face we need to quantify the effort involved. According to Photutorial (Broz), it is estimated that users share approximately 14 billion images daily across social media. If only 0.5% of these images were fake that still means that 70 million images would need to be checked daily. Considering images are shared instantaneously it means that to be accepted by the general public, integrity and accuracy checks would need to perform analysis without interrupting the flow of data.

Final Thoughts

Deepfakes have already caused untold damage to many lives either through an invasion of privacy or spread of disinformation. Lawmakers cannot wait for a catastrophic event to take action. A concerted effort must be made by all affected stakeholders to come together to effectively address the threat of deepfakes. Through a layered approach a solution can be found to limit the negative effects and enhance the positives of this incredible technology.

References

Chesney, R. and Citron, D.K. (2019). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. Foreign Affairs. Available at: https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war [Accessed 18 Jun. 2024].

Nguyen, T.T., Nguyen, C.M., Nguyen, D.T., Nguyen, D.T. and Nahavandi, S. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey. arXiv preprint arXiv:1909.11573. Available at: https://arxiv.org/abs/1909.11573 [Accessed 18 Jun. 2024].

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11), pp.39-52. Available at: https://timreview.ca/article/1282 [Accessed 18 Jun. 2024].

Maras, M.H. and Alexandrou, A. (2019). Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos. The International Journal of Evidence & Proof, 23(3), pp.255-262. doi:10.1177/1365712718807226.

Fridrich, J. (2012). Rich Models for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security, 7(3), pp.868-882. doi:10.1109/TIFS.2012.2190402.

A. Qureshi, D. Megías, and M. Kuribayashi (2021). Detecting Deepfake Videos using Digital Watermarking. *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC), Tokyo, Japan, 2021, pp. 1786-1793.

Channel 4. (2024). Watch Can AI Steal Your Vote? Dispatches | Stream free on Channel 4. *Channel 4*, 27 June 2024. Available at: https://www.channel4.com/programmes/can-ai-steal-your-vote-dispatches [Accessed 3 July 2024].

The Telegraph. “How deepfake AI could swing the General Election.” *The Telegraph*, 26 June 2024, https://www.telegraph.co.uk/tv/2024/06/26/deepfake-ai-general-election-cathy-newman-channel-4/. [Accessed 3 July 2024].

Tonkin, S. (2024) “Want to spot a deepfake? Look for the stars in their eyes.” *The Royal Astronomical Society*, 17 July 2024, https://ras.ac.uk/news-and-press/news/want-spot-deepfake-look-stars-their-eyes. [Accessed 1 August 2024].

Sjouwerman, S., 2024. How a North Korean Fake IT Worker Tried to Infiltrate Us. Incident Report Summary: Insider Threat. KnowBe4. Available at: https://blog.knowbe4.com/how-a-north-korean-fake-it-worker-tried-to-infiltrate-us [Accessed 1 August 2024].

BBC (2024) ‘Trump supporters target black voters with faked AI images’, 3 March. Available at: https://www.bbc.com/news/world-us-canada-68440150 [Accessed: 5 August 2024].

Dhar, A., Agrawal, E. (2024). Detecting AI-Generated Deep Fakes Using ResNext CNN and LSTM-Based RNN: A Robust Approach for Real-Time Video Manipulation Detection. In: Chaturvedi, A., Hasan, S.U., Roy, B.K., Tsaban, B. (eds) Cryptology and Network Security with Machine Learning. ICCNSML 2023. Lecture Notes in Networks and Systems, vol 918. Springer, Singapore. https://doi.org/10.1007/978-981-97-0641-9_37

A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1-11, doi: 10.1109/ICCV.2019.00009. keywords: {Face;Videos;Forgery;Benchmark testing;Forensics;Three-dimensional displays;Databases},

Wu, H.Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F. and Freeman, W., 2012. Eulerian video magnification for revealing subtle changes in the world. ACM transactions on graphics (TOG), 31(4), pp.1-8.

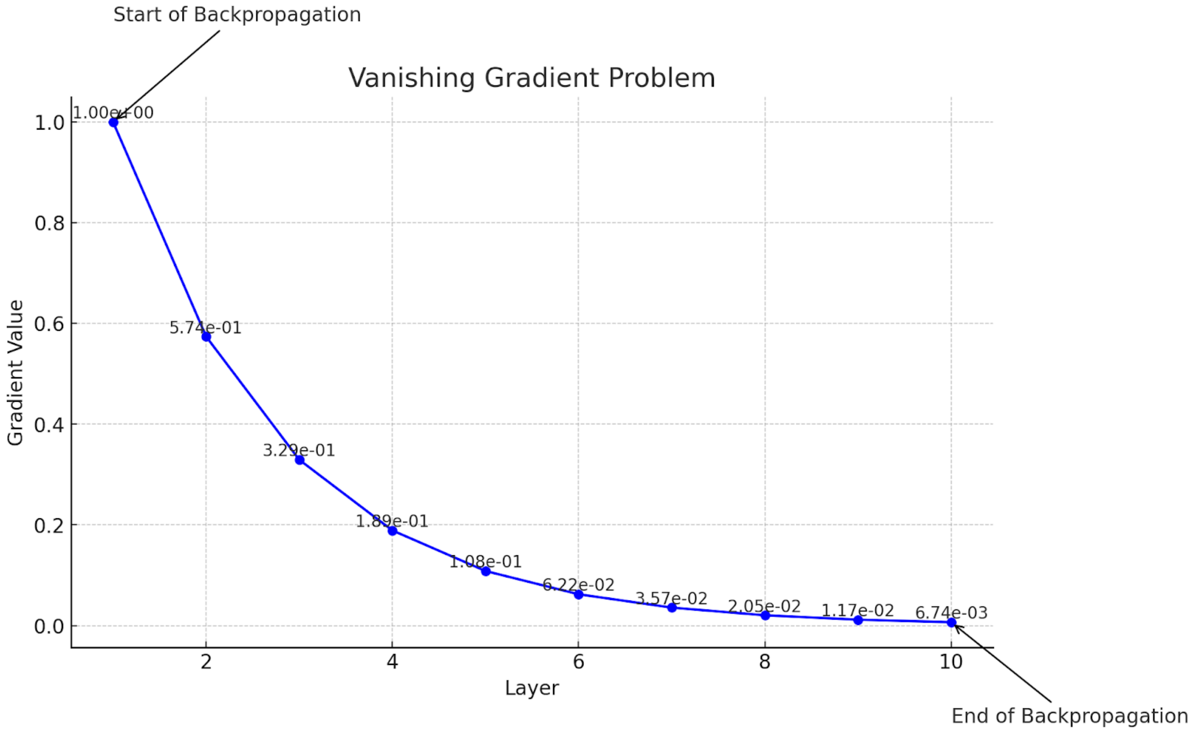
Sky News, 2023. Deepfake audio of Sir Keir Starmer released on first day of Labour conference. [online] 9 October. Available at: <https://news.sky.com/story/labour-faces-political-attack-after-deepfake-audio-is-posted-of-sir-keir-starmer-12980181> [Accessed 8 August 2024].

Ajder, H., Patrini, G., Cavalli, F. and Cullen, L., 2019. The State of Deepfakes: Landscape, Threats, and Impact. [online] September. Available at: <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf> [Accessed 8 August 2024].

Broz, Matic. "How many photos are there? (Statistics & trends in 2024)." Photutorial, 5 July 2024, https://photutorial.com/photos-statistics/. Accessed 8 August 2024.

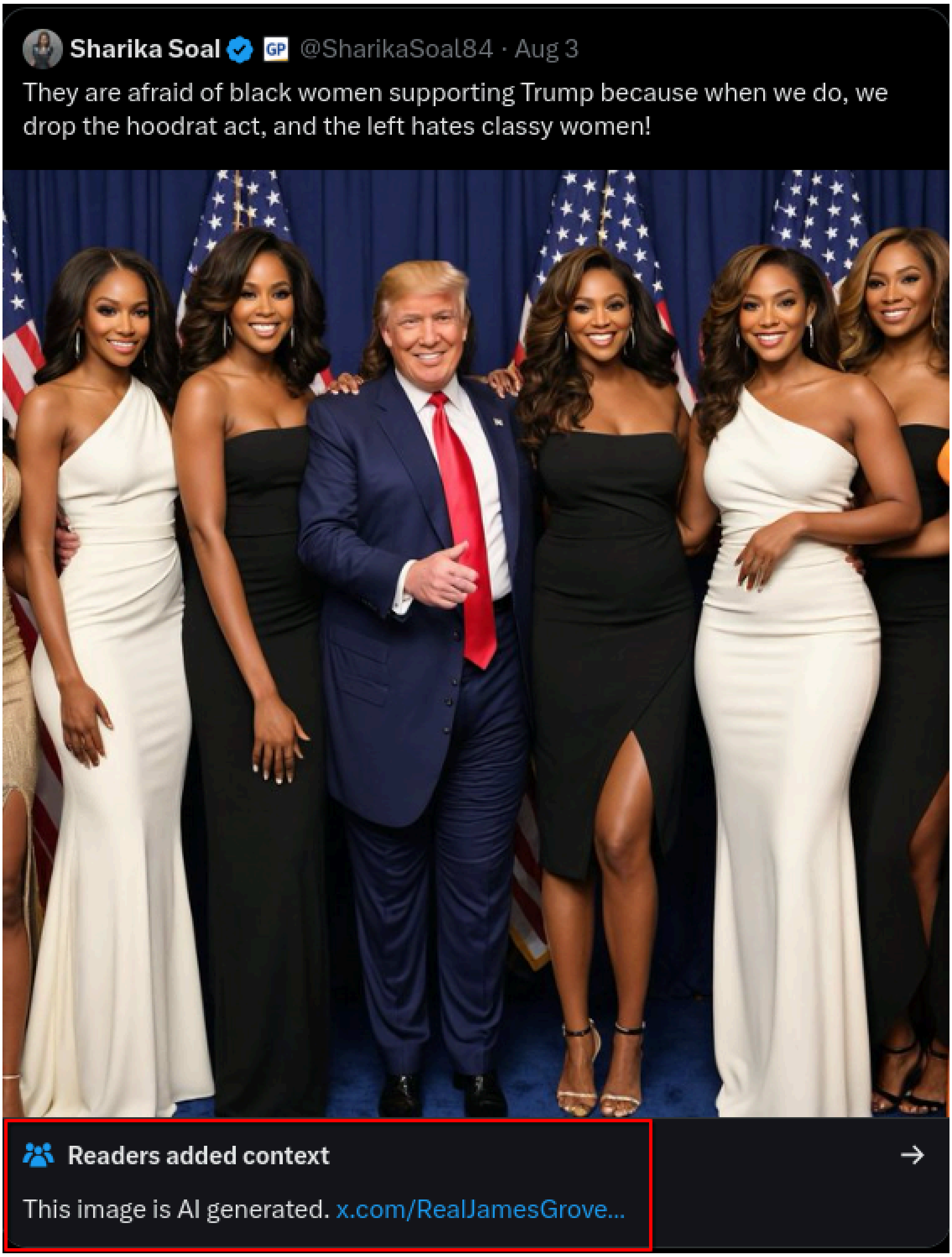
Appendix

Appendix A - The Vanishing Gradient Problem. This is an issue that arises when training neural networks. It is particularly prominent in deep learning models that have many layers. To understand the vanishing gradient problem we need to understand how a neural network is trained. When a neural network is in training it will answer a question then adjust its weights to minimise the error between their prediction and the actual answer. This adjustment is done using a method called backpropagation i.e. the degree of error is calculated and weights are adjusted to minimise the degree of error. In deep networks as we move backwards through the layers the adjustments to the weights become very small and learning slows down significantly.



In the above diagram the x-axis represents the layers of a neural network. The y-axis are the gradient values. As the backpropagation goes through the layers the gradient value decreases exponentially. This graph shows how gradients can vanish in deep networks making it difficult for the initial layers to learn effectively.

Appendix B - A fake image showing Donald Trump in the presence of 5 women with a warning to users regarding the legitimacy of the image.



Sharika Soal @SharikaSoal84 · Aug 3

They are afraid of black women supporting Trump because when we do, we drop the hoodrat act, and the left hates classy women!

Readers added context

This image is AI generated. x.com/RealJamesGrove...

Appendix C - A deepfake used as part of a fraudulent job application that resulted in a company hiring a North Korean hacker. Sjouwerman (2024) with the original image on the left and the doctored image on the right.