Billy To

Analysis of AI-generated Content and Deepfakes in social media

Metropolia University of Applied Sciences

Bachelor of Engineering

Information Technology

Bachelor's Thesis

6 May 2024

Abstract

Author:	Billy To
Title:	Analysis of AI-generated Content and Deepfakes in Social Media
Number of Pages:	37 pages
Date:	6 May 2024
Degree:	Bachelor of Engineering
Degree Programme:	Information and Communications Technology
Professional Major:	Mobile Solutions
Supervisors:	Toni Spännäri, Senior Lecturer

The advancement of Artificial Intelligence technology has led to an increase in deepfake and AI-generated content in social media platforms. Creating these AI content has been problematic for social media platforms due to the capability to manipulate users into believing in something that is not real, which is why these technologies pose a threat to society with the potential to create misinformation and fake news.

The objective of this thesis was to inform readers about the capabilities of manipulation techniques and help with detecting a possible use of such techniques on social media platforms. Also, raising awareness of the rapid growth of fake content on social media platforms should lead the users to double-check the source of the content.

This thesis was based on statistics and surveys from external sources, which presented the general idea of how, where, why, and what the AI content was being used for. Also, a few tests are conducted using Convolutional Neural Network model to demonstrate the ability to classify an object based on real and fake images. The outcome of the surveys and statistics was that the AI-generated content has increased exponentially, and awareness globally was less than fifty percent.

The outcome of this study was to raise awareness of the existence of deepfake and AI-generated content and comprehend the capabilities of AI tools and the impact they have on social media platforms.

Keywords: ai, artificial intelligence, deepfake, social media

---

Tiivistelmä

Tekijä:	Billy To
Otsikko:	Tekoälyn Luoma Sisällön ja Syväväärennöksen Analyysi Sosiaalisessa Mediassa
Sivumäärä:	37 sivua
Aika:	6.5.2024
Tutkinto:	Insinööri (AMK)
Tutkinto-ohjelma:	Tieto- ja viestintätekniikka
Ammatillinen pääaine:	Mobiilikehitys
Ohjaajat:	Lehtori Toni Spännäri

Tekoäly teknologian kehitys on johtanut syväväärennöksen ja tekoäly luoman sisällön lisääntymiseen sosiaalisessa mediassa. Tekoälyn luoma sisältö on ollut ongelmallinen sosiaalisessa mediassa, koska tekoälyn kyky manipuloida käyttäjiä uskomaan jotakin, joka ei ole aitoa. Tämän takia tekoäly teknologiat ovat uhka yhteiskunnalle, joilla on mahdollisuus luoda väärää tietoa ja valeuutisia.

Insinöörityön tarkoituksena oli informoida manipulointitekniikan kyvyistä ja auttaa havaitsemaan mahdollisesta tekniikan käytöstä sosiaalisessa mediassa. Lisäksi väärennetyn sisällön nopea kasvu sosiaalisessa mediassa pitäisi johtaa käyttäjiä tarkistamaan sisällön lähdettä.

Insinöörityö perustui tilastoihin ja kyselyihin ulkopuolisesta lähteestä, jotka esittävät yleiskuvan että, miten, missä, miksi ja miten tekoälyä käytetään. Lisäksi muutamia testejä tehtiin käyttäen konvoluutiohermoverkkoa, jonka tarkoituksena oli näyttää konvoluutiohermoverkon kyky luokitella objekteja todellisten ja väärennettyjen kuvien perusteella. Kyselyjen ja tilastojen tuloksena oli, että tekoälyn luoma sisältö on lisääntynyt eksponentiaalisesti, ja maailmanlaajuinen tietoisuus oli alle viisikymmentä prosenttia.

Tämän tutkimuksen tulos oli tietoisuuden lisääntyminen syväväärrennöksen ja tekoälyn luotu sisällön olemassaolosta ja käsitys tekoälytyökalujen kyvyistä ja niiden vaikutus sosiaalisessa mediassa.

Avainsanat: ai, tekoäly, syväväärennös, sosiaalinen media

Contents

List of Abbreviations

AI:	Artificial Intelligence
AAAI:	Association for the Advancement of Artificial Intelligence is an internation scientific society who is responsible of use of Artificial Intelligence
GAN:	Generative Adversarial Network
CNN:	Convolutional Neural Network
TTS:	Text-to-speech
NLP:	Natural Language Processing
VAE:	Variational Autoencoder
MNIST:	Modified National Institute of Standards and Technology database is a database of handwritten digits for training image processing systems.

1 Introduction

This thesis focuses on the growing AI-generated content and deepfakes in social media sites such as Facebook, X, TikTok, YouTube, and news source sites, which has caused a lot of problems with misinformation and fake news. These AI contents have become a major concern on social media, as they can be a threat to inexperienced people in technology by fooling them into believing they are genuine content.

AI-generated content and deepfakes have the power to fool users into believing in misinformation. As the AI-technology is advancing very quickly, which makes the new AI content even more difficult to notice and the tools to counter these must also keep up with the pace.

This thesis aims to learn about the AI-generated content and deepfakes by analysing the AI models they are using and researching their methods of usage and learning how to be aware when AI technology is being used.

2 Background

Artificial intelligence (AI) is a technology where information is being fed into a computer or machine to simulate human intelligence and problem-solving capabilities.

The founding fathers behind AI technology were John Von Neumann and Alan Turing in the 1950s. Alan Turing is one of the people who introduced the concept of the "game of imitation" in his article in 1950 called "Computing Machinery and Intelligence". In the article, Turing questioned that are the humans able to distinguish a conversation between a person and a machine. This concept is also known as the famous Turing test. Even though the AI technology was exciting and new, the popularity of AI in the 1960s was steep because of underperforming machines with very little memory, which made it difficult to use a computer language. (Coe, n.d.)

Even though interest in AI technology was low in the 1960s, the technology was rapidly gaining growth towards to 1980s, which fuelled more interest in AI. This timeline is labelled as the “AI boom”. In the 1980s, the government started supporting researchers by funding them for research purposes, which led to Deep Learning techniques and the use of Expert System becoming more popular. Despite the small period time of growth and interest in AI, the Association for the Advancement of Artificial Intelligence (AAAI) warned of an AI Winter in 1984. (Tableau, n.d.)

The warned AI Winter started in 1987, meaning that the interest in AI has lessened and therefore the funding for research has also decreased and paused. The reasoning behind the AI Winter was due to the high cost of funding with a low return. (Tableau, n.d.)

The AI Winter did not stop researchers from developing AI technologies. In 1997, IBM developed a chess playing machine as seen in Figure 1, called Deep Blue that was able to defeat the greatest chess player of all time, Garry Kasparov. This was a huge milestone for the AI technology. (IBM, n.d.)

Figure 1 A picture of Deep Blue machine that defeated the chess world champion in 1997. (Chess, n.d.)

2.1 Rise of AI Technology

Artificial intelligence is advancing at an unprecedented rate, surpassing human capabilities in performing tasks that were considered impossible even for machines in many different industries for example, medical and software development. Even though the technology seems beneficial, it also raises a lot of concerns due to its ethical use and possible risks to jobs. (Butazzo, 2023)

The boom of AI technology has caused a lot of consumers and industries to use easily accessible AI tools such as ChatGPT, which is one of the most used AI tools with 14.6 billion total website visits between September 2022 to August 2023, as shown in Figure 2. (Conte, 2024)

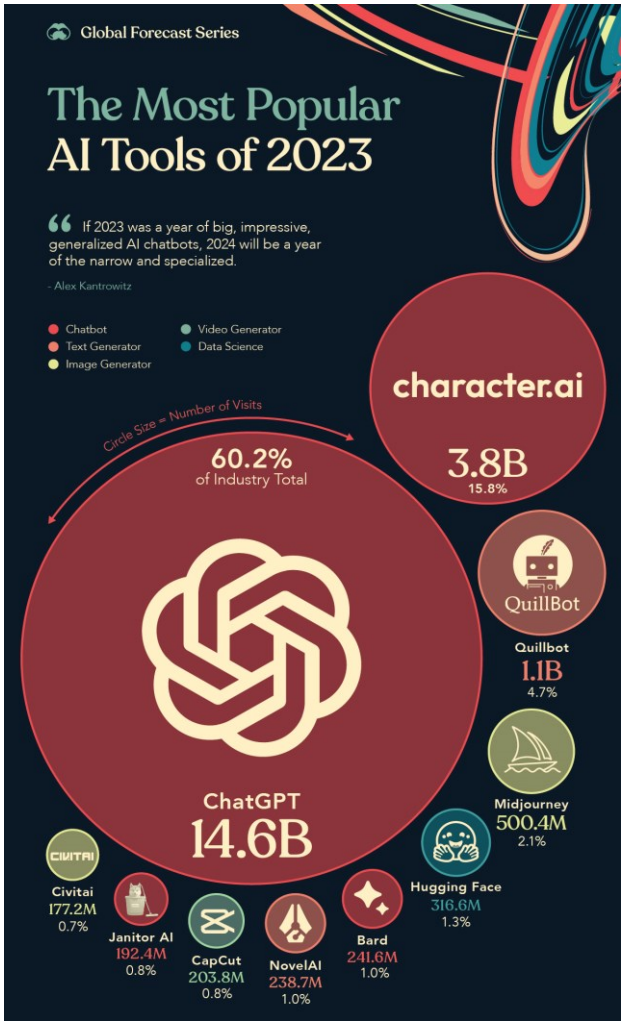

Figure 2 Statistics of the most used AI tools in 2023. (Conte, 2024)

The global market for AI is also growing exponentially. In 2022, the market size was valued at 135.55 billion USD, and it is projected to reach 1811.8 billion USD by 2030. (Maheshwari & Jain, 2024) The country with the most economic gains from AI will be China with 7.0 trillion USD, followed by North America with 3.7 trillion USD, totalling 10.7 trillion USD, which is 70% of the global economic impact. Economic gains for other regions can be seen in Figure 3.

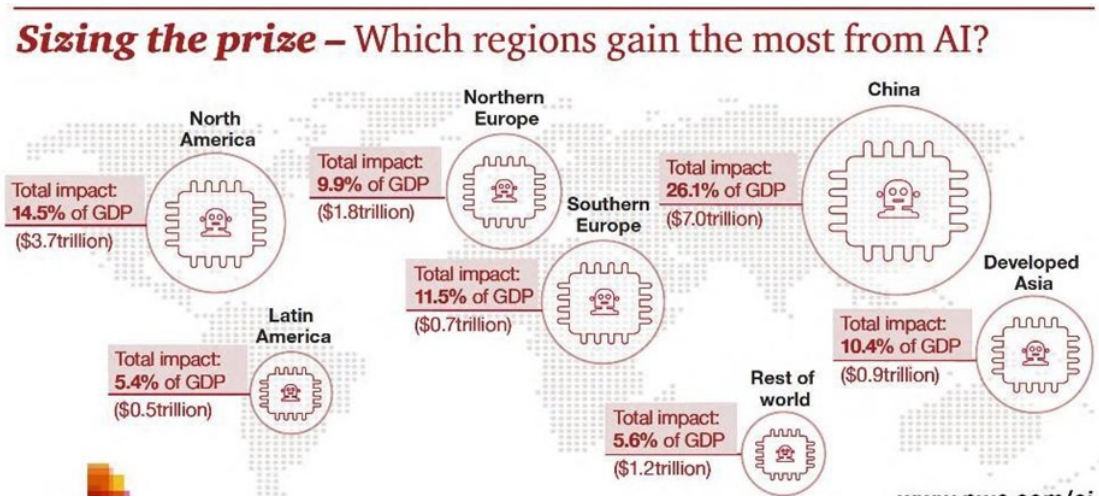

Figure 3 Economic gains from AI in different regions. (Jain & Maheshwari, 2024)

The AI technology in healthcare is gaining a lot of growth, estimating 19.27 billion USD in 2023 due to demand for enhanced efficiency, accuracy, and better patient outcomes. In the United States of America, 79 percent of healthcare organizations are using AI technology. The future market will grow exponentially as shown in Figure 4. (Grand View Research, 2024)

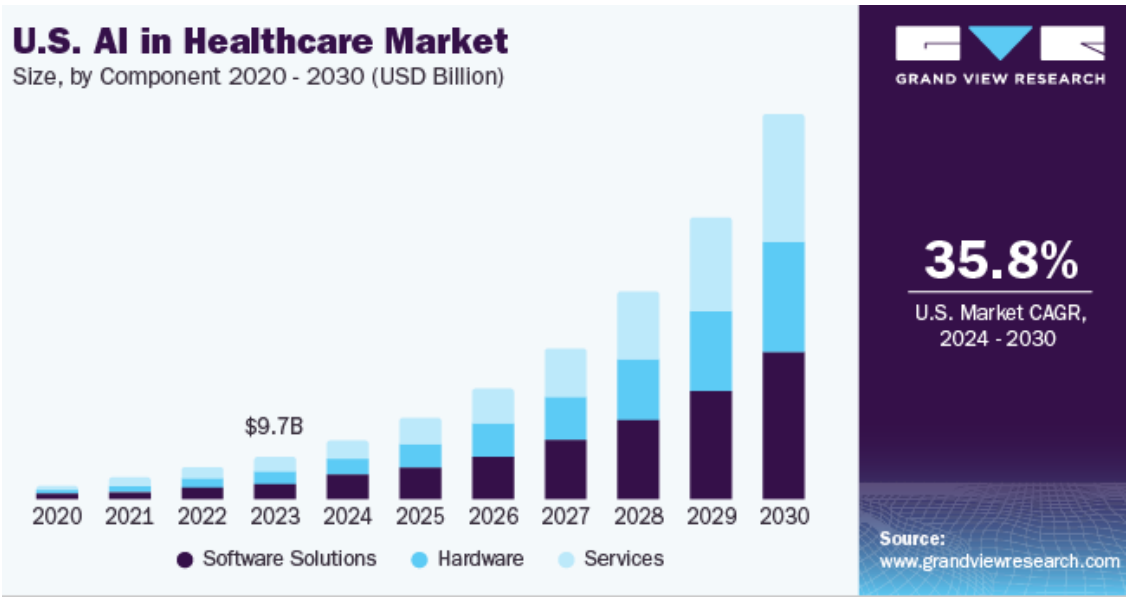

Figure 4 Predicted growth in healthcare market in the United States. (Grand View Research, 2024)

Increased use of AI technology raises questions about its effect on jobs. According to an analysis, almost 40 percent of global employment is exposed to AI, especially in high-skilled jobs. The advanced economy is impacted by AI by 60 percent, as half of the jobs may take advantage of AI integration and the other half can perform tasks that are performed by humans therefore, it could lead to lower labour demand, followed by lower wages and less hiring. Figure 5 shows

the statistics of exposure by job type, where low-income jobs are less likely to be replaced by AIs. (Georgieva, 2024)

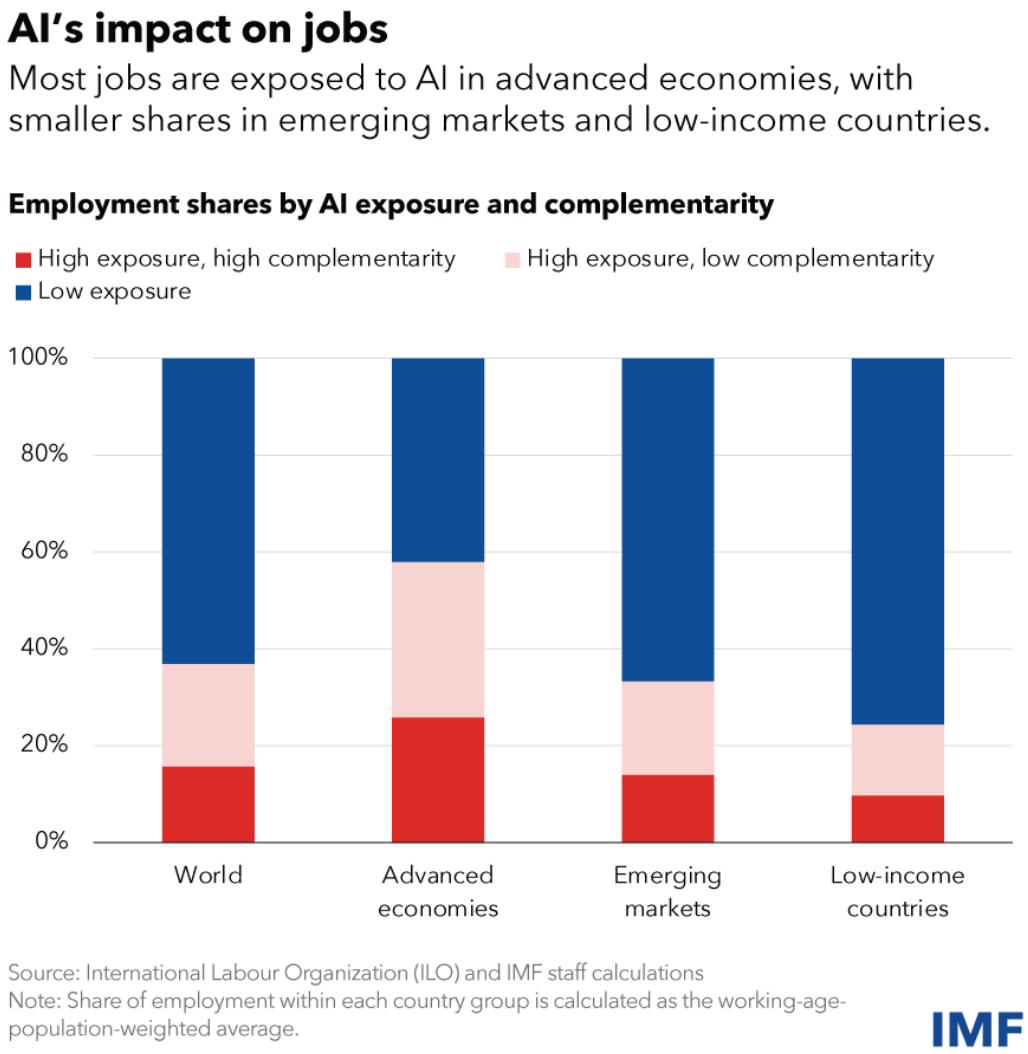

Figure 5 Statistics of jobs that are exposed by AI. (Georgieva, 2024)

2.2 AI-generated Content and Deepfakes

AI-generated content is machine made content where the machines analyse large amounts of data to produce human-like content based on machine learning. They can generate a full text-based story based on your inputs and they can also produce images. (Zaluski, 2023) This type of AI technology is also known as generative AI, which is not brand-new since it was already introduced in the 1960s in chatbots. With the introduction of generative adversarial networks

(GANs) in 2014 the generative AI was able to create highly realistic images, videos, and audio of real people. One of the most popular generative AIs is ChatGPT. Using these tools are very beneficial, they can be used to help create a job application for example. (Lawton, 2024).

Deepfakes are also a type of AI-generated content, but instead of relying on user inputs, this technology is used to manipulate faces and synthesize voices. Just like text-based generated AI content, deepfakes also needs to be trained by feeding a large amount of data of images and audio to create realistic deepfakes. Figure 6 shows an example of deepfake technology.

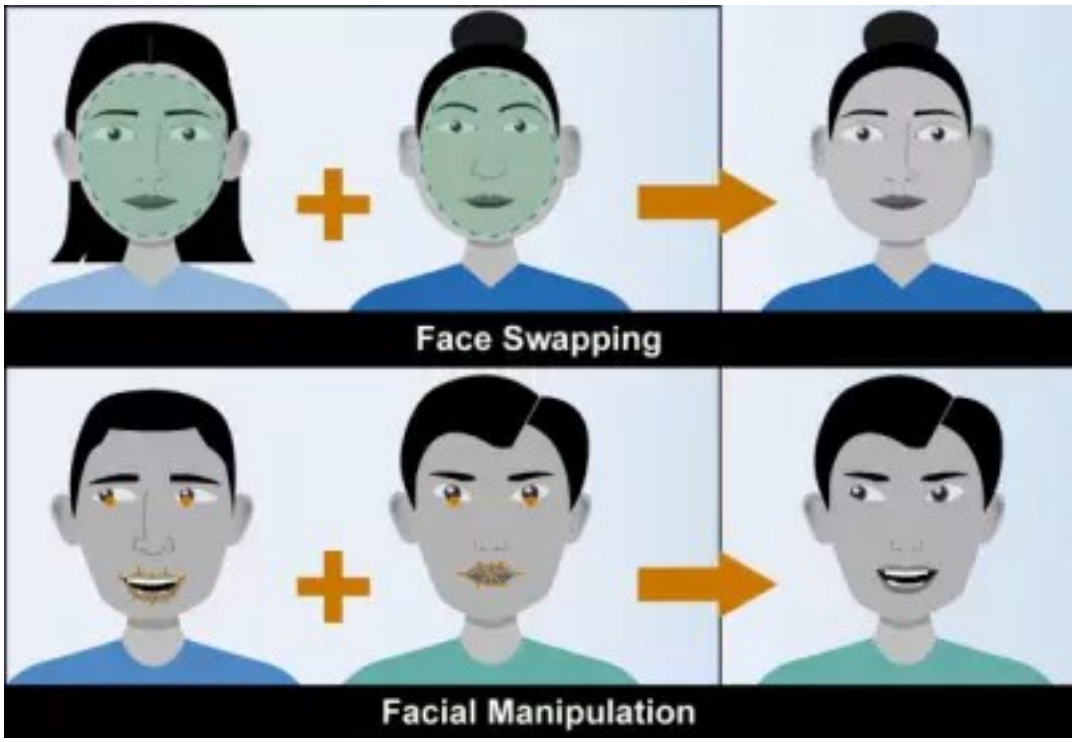

Figure 6 A screenshot of swapping and manipulating faces using deepfake tools. (GAO, 2020)

The deepfake technology was used in a non-harmful way as in jokes, but after realizing the potential of deepfake technology, it can be used to trick people into believing in false information. (Frolov et al., 2022)

The potential of deepfake is frightening when used maliciously during unprecedented events such as war and presidential elections. For example, a deepfake video of President Zelensky surrendering to Russia was spread through social media during the war between Russia and Ukraine. (The Telegraph, 2022) According to a survey with 16,000 participants in 2022, 71 percent of people worldwide don't know what a deepfake video is and 57 percent of people could differentiate between a deepfake and a genuine video globally. (Iproov, n.d.) Figure 7 shows the awareness of deepfake videos from selected countries.

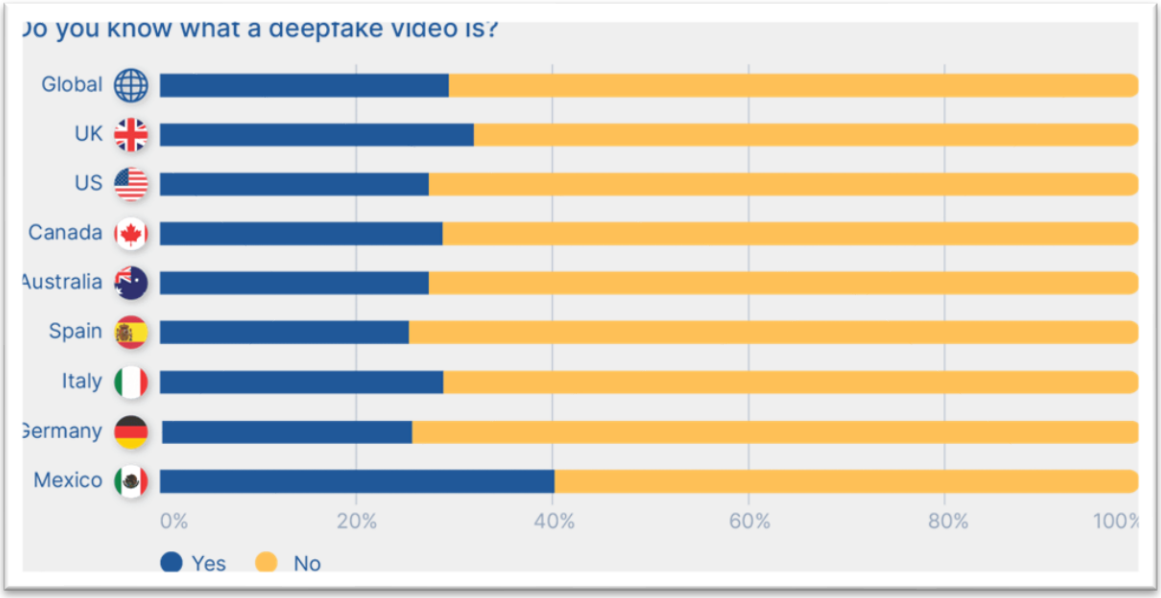

Figure 7 A screenshot of a graph showing the awareness of deepfake videos from different countries. (Iproov, n.d.)

With how low the global awareness of deepfake videos is and the rapid increase of deepfake videos on social media platforms can be very alarming. According to deepfake statistics, the amount of deepfake videos has increased by around 650 percent from 2019 to 2023. Figure 8 shows the number of deepfake videos by

year.

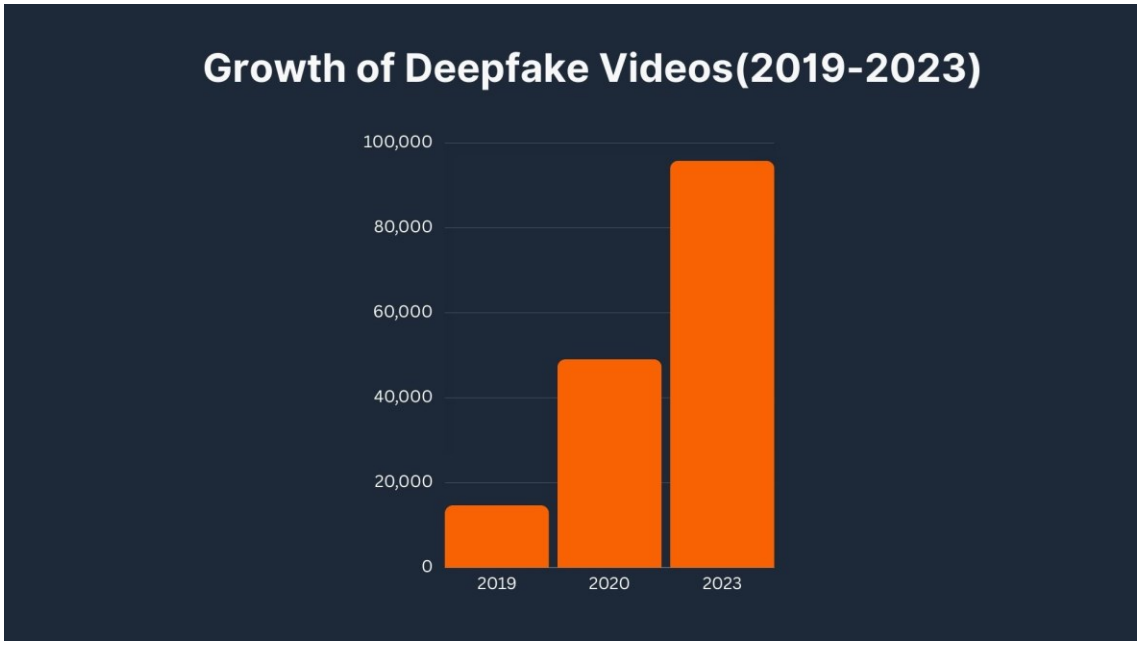

Figure 8 A graph showing the growth of deepfake videos on the internet. (Reddy, 2023)

The most dominant use for deepfake is pornographic content, where 96 percent of said technology is used for adult content. (Reddy, 2023) Which has led to a major issue for women that has fallen into unconsented usage of their likeness in pornographic content.

2.3 Impact on Social Media

AI-generated content and deepfakes have mostly caused a lot of problems on social media. With the tools to manipulate things, it is easier to exploit the people who are unaware of the possibilities of AI technologies or AI in general.

Misinformation from AI-generated content and deepfakes are very effective when it comes to political issues especially during election times. Recently in the United States of America, there was a deepfake robocall featuring President Joe Biden in New Hampshire presidential primary election, where the call told the

Democratic voters not to vote in the primary. Figure 9 portrays a video of the robocall. (Teale, 2024)

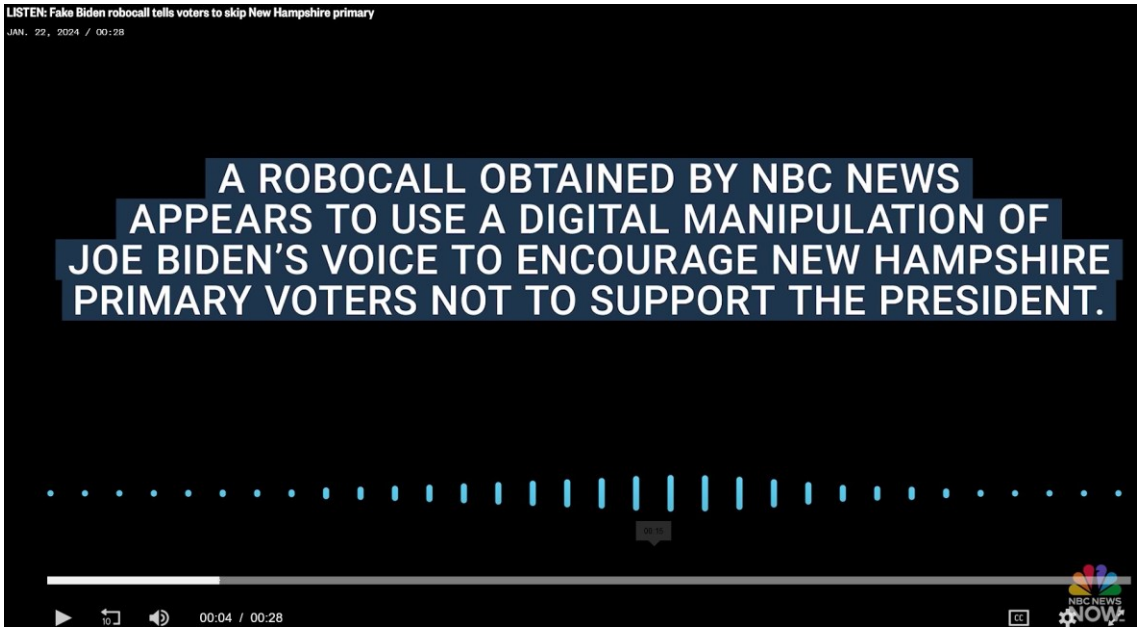

Figure 9 A voice recording of Joe Biden's robocall during presidential elections in New Hampshire. (NBC News, 2024)

Public opinions can also be easily manipulated by generating disinformation about health in multiple languages and producing a lot of fake patient and clinician testimonies and generating fake images of health products that cause harm to people. (Davey, 2023)

With the amount of misinformation and disinformation in social media, it causes users to not trust the authenticity of the content on these platforms. This could discourage users from using these media sites. Because of this problem, there needs to be moderation to detect all the fake content.

Because of the surge in AI usage on social media sites, it is important to learn media literacy in the current year. In the United States of America, there are currently 18 states that are teaching media literacy in schools to identify disinformation and misinformation on the internet. Even though there are no

federal guidelines for teaching media literacy, the Congress are trying to change that. (Klawans, 2024)

2.4 Challenges and Concerns

With the use of AI technology and deepfakes. it may cause some ethical concerns. When using AI technology, there needs to be a guideline on how to use these tools. When you are developing an AI technology you need to address the following: Bias and fairness where you must minimise biases in training data and algorithms, ensuring fairness and non-discrimination. Transparency where you developed the AI with transparency in mind, making it traceable and explainable in the AI's decision-making process. And lastly, accountability where the developers or users of AI must take accountability for the outcomes of their technologies when they have caused an impact on public opinion or infringed on personal rights. (Ahvanooey et al., 2023)

The ethical use of deepfakes can be positive if it is used for educational purposes, medical simulations or in therapeutic settings. (Ahvanooey et al., 2023)

Regulating AI content and deepfakes has been a challenge in the United States of America because they can sometimes be considered as a legitimate content if they are not used deceptively. The fast pace of deepfake creation and usage has gained a response from the U.S. government, which led to the proposal of legislation to regulate AI. (Norden & Weiner, 2023) While the European Commission have published their proposal to regulating AI in April 2021 with the aim of enabling trustworthy and secure application of AI and respecting the values and fundamental rights of EU citizens. The EU regulatory framework is a risk-based model that has four different risk levels as follows: unacceptable risk, high risk, limited risk, and minimal risk. (Boheemen et al., 2021)

Detecting deepfakes has been very challenging. For example, there was a fake image shared from a verified Twitter account (known as X) of an explosion near the Pentagon in Washington, DC. shown in Figure 10, with a message saying

“Large explosion near the Pentagon complex in Washington, DC. – initial report,” which caused the stock market to dip for a moment. (O’Sullivan & Passantino, 2023)

Figure 10 An AI image of an explosion near the Pentagon that spread through Twitter causing confusion in Washington, DC. (O’Sullivan & Passantino, 2023)

Even though there are a lot of tools to detect deepfakes, for example, a software that can detect AI output are not always reliable, because of technical limitations as the AI technology keeps advancing and becoming harder to detect. Therefore, it is also important to use common ways of detecting deepfakes using reverse image search, checking for inconsistencies, examining metadata, and analysing the source of media. (Basheer, 2023)

The deepfake technology also brings a lot of cybersecurity risks, as shown in Figure 11. The attacker could easily falsify a video or audio to scam a person, giving the criminal what they want. One of the major threats that deepfake has caused is nonconsensual pornography, where most celebrities are used for pornography using the deepfake technology. Another risk is identity theft, where the attacker can create a new identity based on real people, which they can use to create false documents or purchase products by pretending to be that person. (Fortinet, n.d.)

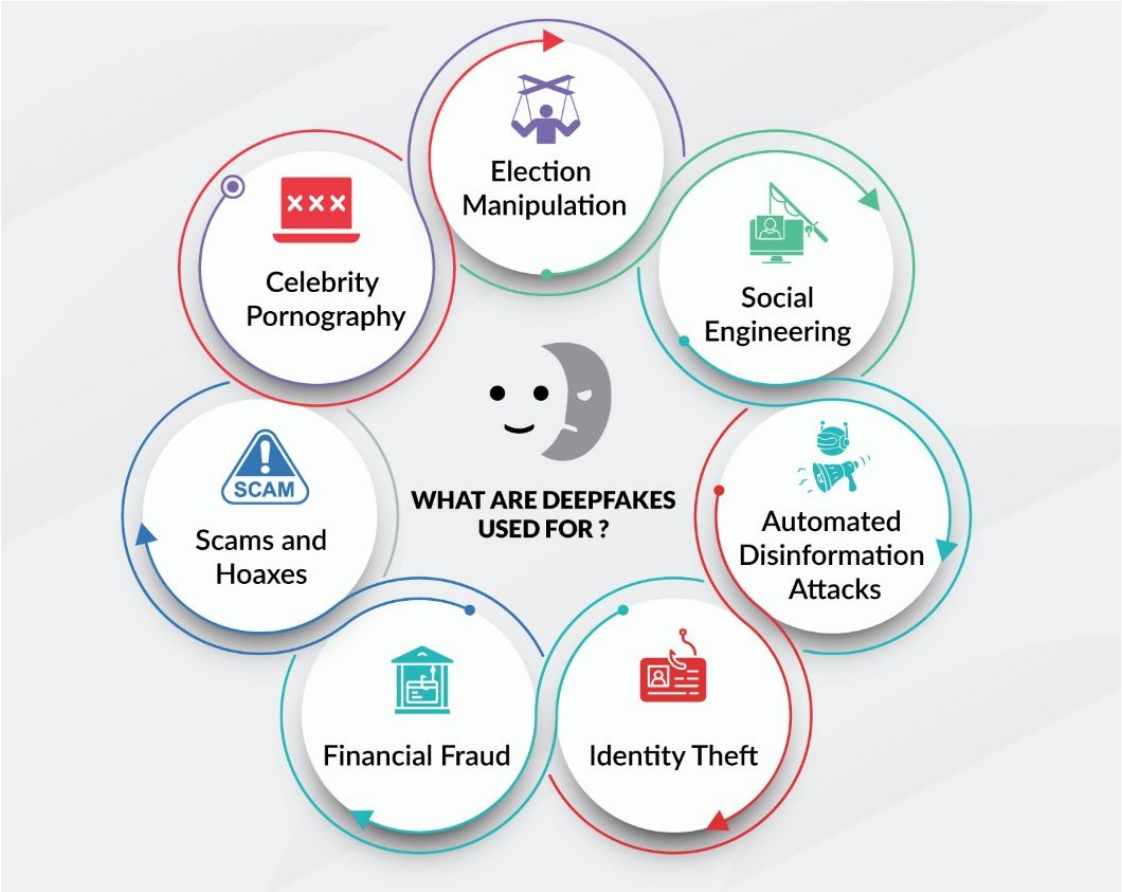

Figure 11 An infographic on how deepfakes are used in malicious ways. (Fortinet, n.d.)

3 Analysis

This chapter analyses AI technology by giving a better picture of how some of the popular AI technologies operate and the results they can produce. Comprehending the capabilities of manipulation technologies gives a better view on the potential of previously mentioned technologies and the impact they have had caused on society and social media platforms.

3.1 Analysis of Manipulation Techniques

This thesis aims to explore popular manipulation techniques that are used to create AI-generated content and deepfakes. By learning these techniques, we can understand the capabilities and risks of AI technology and its effects on social

media. The techniques that are commonly used are Face Swapping, Voice Synthesis and Text-to-Image Synthesis.

3.1.1 Face Swapping

Face Swap is a deepfake technology that is mostly implemented on GAN. The meaning of face swapping is to literally swap a person's face in an image with another person, as shown in Figure 12, while preserving the details of the original image for example, expressions and facial features.

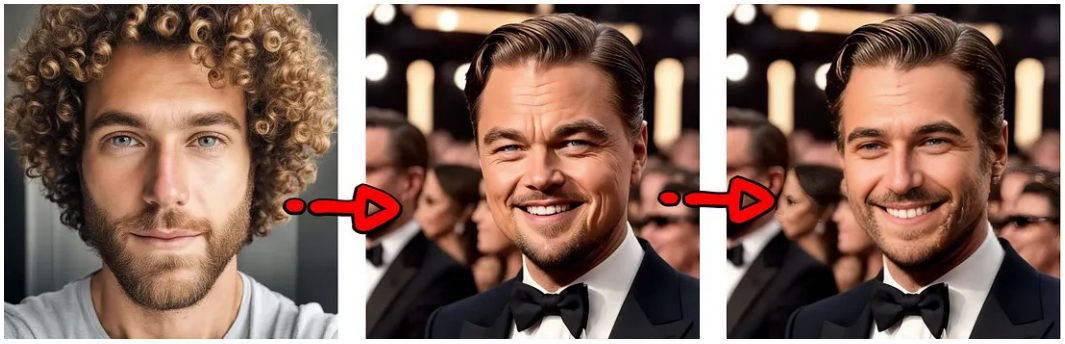

Figure 12 An example of a highly detailed face swapping technique combined with facial landmark alignment. (Negoita, 2024)

To achieve the details on an image, facial landmark alignment does the job by marking out all the positioning of the person’s facial features such as the eyes, nose, mouth, and chin shown in Figure 13. (Negoita, 2024)

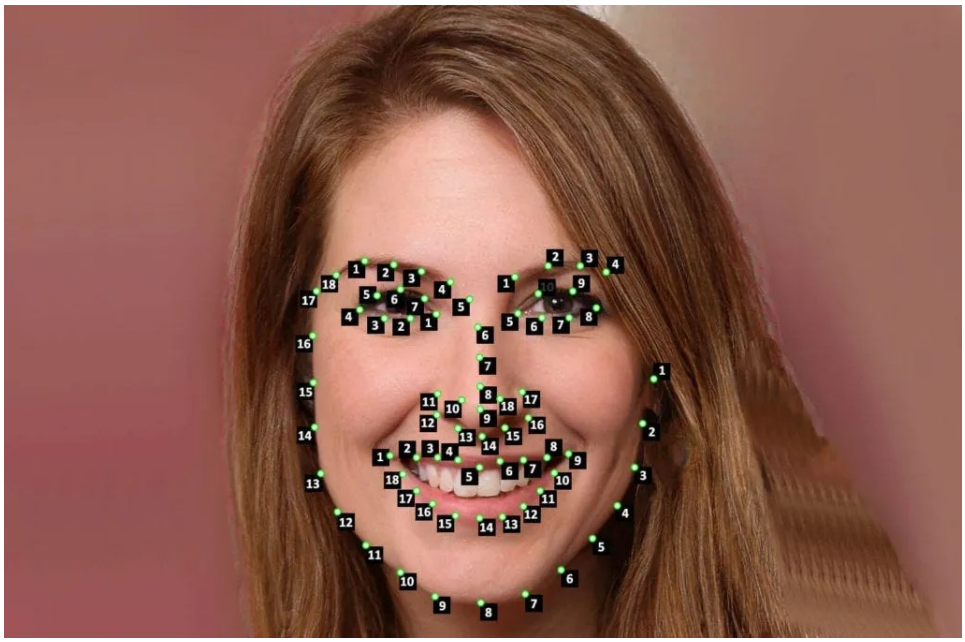

Figure 13 An image of facial landmark alignment on a person's face. (Negoita, 2024)

For this type of face swap technique, we use the GAN method, which consists of two parts: a generator and a discriminator. These two parts have their own operation. The generator tries to create new images that look genuine, and the

discriminator tries to identify genuine images from fake ones. See Figure 14 for the GAN structure. (Negoita, 2024)

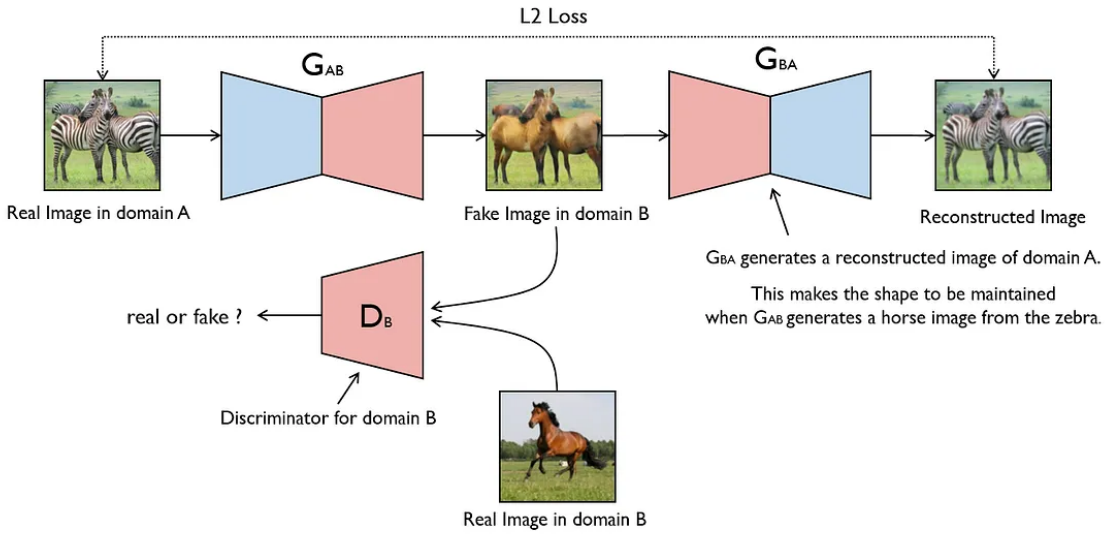

Figure 14 A picture of how the GAN architecture works. (Negoita, 2024)

In social media, face swapping is very concerning especially when it is used to misinform, creating fake news, and manipulating politically to spread false information. As shown in Figure 15, Alec Baldin is imitating Donald Trump on a show called Saturday Night Live and the deepfake of Donald Trump is very indistinguishable from the actual President, which can be very dangerous if the

people are not aware that this is indeed a deepfake image. (Smith, 2018)

Figure 15 A screenshot of a skit from Saturday Night Live, where Donald Trump's face (right) is face-swapped with Alec Baldin's face (left). (Smith, 2018)

3.1.2 Voice Synthesis

Voice synthesis, also known as text-to-speech (TTS), focuses on generating human-like speech by using advanced algorithms and machine learning. TTS is mainly being developed using three main methods, which are machine learning algorithms, Natural Language Processing (NLP), and speech synthesis techniques. (Podcastle, 2023)

NLP allows machines to understand human language. This technique finds important details in written words and sentences, which allows the AI to interpret and speak in complex sentences. (Podcastle, 2023)

Speech synthesis technique allows machines to turn processed text into coherent and expressive speech. There are different ways to use this technique, such as recording a sample of sounds and fusing them together (concatenative synthesis). (Podcastle, 2023)

Voice synthesis in social media has its own unique way of using it, whether it's used in a good or bad way. For example, music can be sung by a different artist using the voice synthesis, such as Kanye West singing Hey There Delilah by Plain White T's, or the AI technique can be used maliciously by scamming your loved ones after simulating the person's voice to sound as close as possible (Belanger, 2023). And of course, for political use as this thesis has previously mentioned about the fake robocall from the President Joe Biden, which was done with TTS. Since these voice synthesis tools are easy to use, they are being misused to make fake celebrity audio clips saying unhinged or fraudulent statements. (Moon, 2023). Figure 16 showcases the potential of voice synthesis.

Figure 16 A music track called Hey There Delilah by Plain White T's is sung by Kanye West using voice synthesis. (YeezyBeaver, 2023)

3.1.3 Text-to-Image Synthesis

Text-to-image synthesis generates a visually equivalent image from descriptive user text inputs. This allows users to create appealing images for their own use.

(Alhabeeb & Al-Shargabi, 2024). Figure 17 shows a simple screenshot of how an image is created through the text-to-image method.

Figure 17 A screenshot of text-to-image architecture. (Alhabeeb & Al-Shargabi, 2024)

To create these text-to-image images, it is advised to use generative AI models, which helps with generating these images. There are a few generative AI models that are being used the most, such as diffusion models, GANs, and Variational Autoencoders (VAEs). (Gadwal, 2023). Figure 18 shows a few of the generative models.

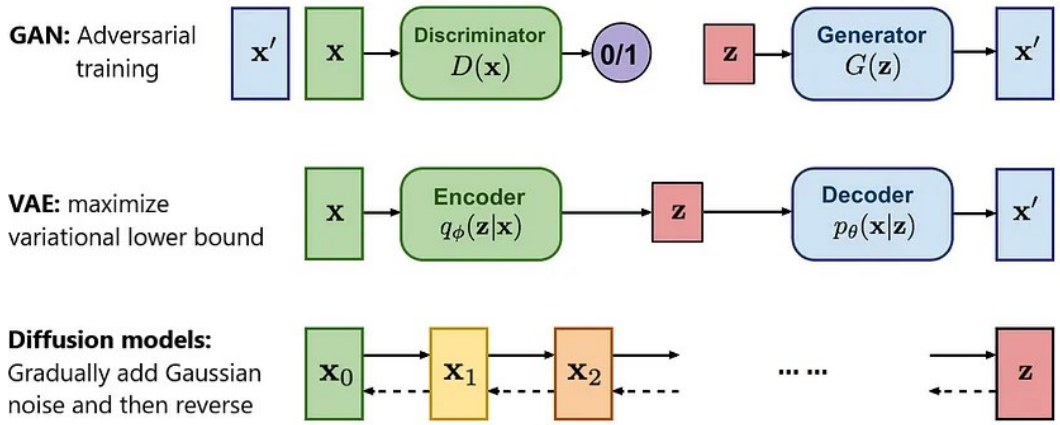

Figure 18 A screenshot of the generative model architecture. (Gadwal, 2023)

Diffusion models are a type of a generative model that works by erasing noise from trained data and then recovering the data back by reversing the noising

process, which then the model can generate data by passing randomly sampled noise through the learned denoising process. (O'Connor, 2022) Figure 19 shows a generated image from noise.

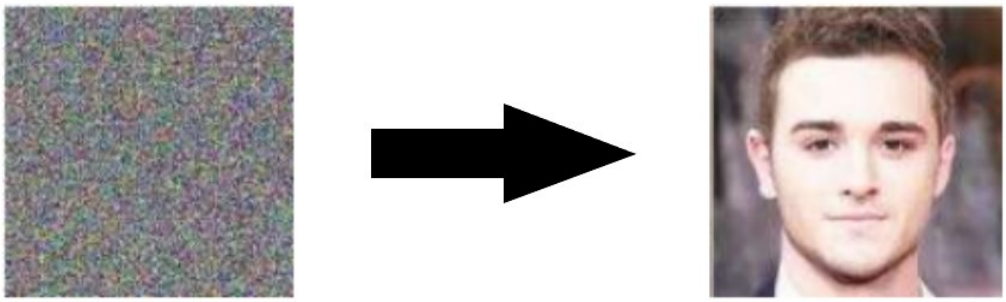

Figure 19 A screenshot of a noise that is generating into an image.

The easiest way to generate images from a diffusion model is by using online generators, for instance DALL-E, as seen in Figure 20, or by creating a simple Python script using torch libraries, as shown in code snippets.

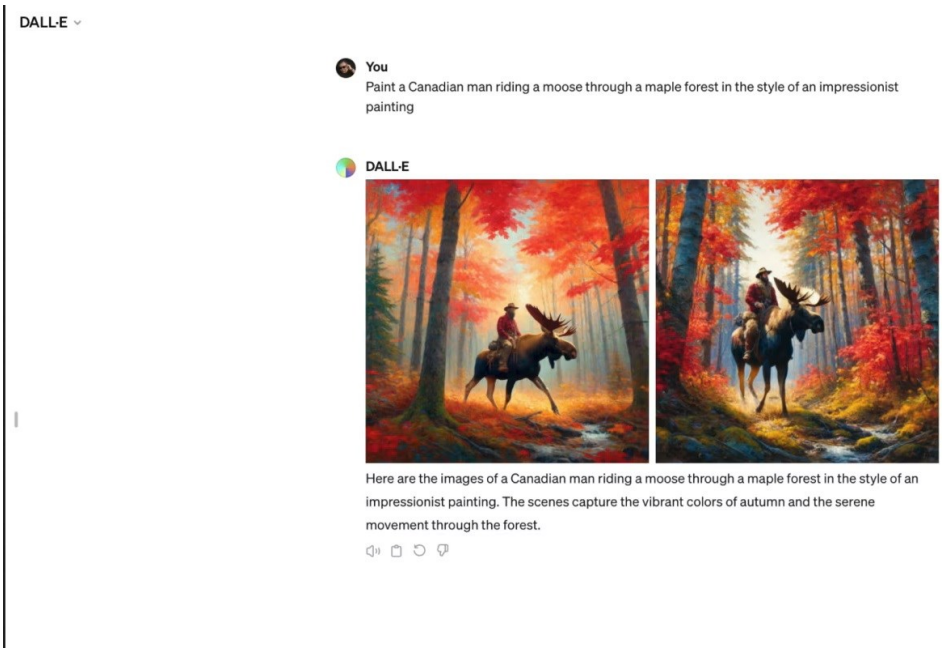

Figure 20 A screenshot of DALL-E generating an image through user text inputs. (Guinness, 2024)

At first, let's define a diffusion model (listing 1).

```
class DiffusionModel(nn.Module):
    def __init__(self, input_size, hidden_size):
        super(DiffusionModel, self).__init__()
        self.encoder = nn.Linear(input_size, hidden_size)
        self.decoder = nn.Linear(hidden_size, input_size)

    def forward(self, x):
        x = self.encoder(x)
        x = torch.relu(x)
        x = self.decoder(x)
        return x
```

Listing 1. A diffusion model that is defined with layers and forward pass.

After defining the model, it needs to load and preprocess the dataset (listing 2).

```
transform = transforms.Compose([
    transforms.ToTensor(),
    transforms.Resize((28, 28)),
    transforms.Normalize((0.5,), (0.5,))
])

train_dataset = datasets.MNIST(root='./data', train=True, download=True,
transform=transform)
train_loader = torch.utils.data.DataLoader(train_dataset, batch_size=64,
shuffle=True)
```

Listing 2. Modified National Institute of Standards and Technology database (MNIST) dataset is being used for converting images to tensors, resizing, and normalising.

Now the model needs to be initialised and then the loss function and optimizer are defined (listing 3).

```
input_size = 28 * 28
hidden_size = 128
model = DiffusionModel(input_size, hidden_size)
criterion = nn.MSELoss()
optimizer = optim.Adam(model.parameters(), lr=0.001)
```

Listing 3. The diffusion model is being initialised with an input and hidden layer size. Criterion and optimizer are defined to train the diffusion model.

The model needs to be trained with a training loop (listing 4).

```
num_epochs = 10
for epoch in range(num_epochs):
    for batch_idx, (data, _) in enumerate(train_loader):
        data = data.view(data.size(0), -1)  # Flatten the images
        optimizer.zero_grad()
        output = model(data)
        loss = criterion(output, data)
        loss.backward()
        optimizer.step()

        if batch_idx % 100 == 0:
            print(f"Epoch [{epoch+1}/{num_epochs}], Batch
[{batch_idx+1}/{len(train_loader)}], Loss: {loss.item():.4f}")
```

Listing 4. The model is being trained for 10 epochs, where each epoch is loaded in batches.

After training the model, it can now generate and visualize a diffusion image (listing 5).

```
with torch.no_grad():
    sample_data, _ = next(iter(train_loader))
    sample_data = sample_data[0].view(1, -1)
    reconstructed_data = model(sample_data)
    reconstructed_data = reconstructed_data.view(1, 28, 28)
    plt.figure()
    plt.subplot(1, 2, 1)
    plt.title('Original Image')
    plt.imshow(sample_data.view(28, 28), cmap='gray')
    plt.subplot(1, 2, 2)
    plt.title('Diffusion Image')
    plt.imshow(reconstructed_data.view(28, 28), cmap='gray')
    plt.show()
```

Listing 5. A sample image is being reconstructed and then visualized side-by-side with the original image, as seen in Figure 21.

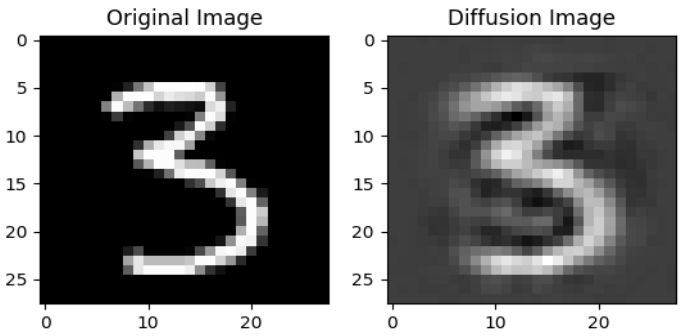

Figure 21 A screenshot of a result from diffusion using the torch library.

VAEs involves an encoder and a decoder, where the encoder takes input data to map it out to a latent space, while the decoder takes input data from the encoder and then reconstructs the original data see Figure 22. (Shende, 2023)

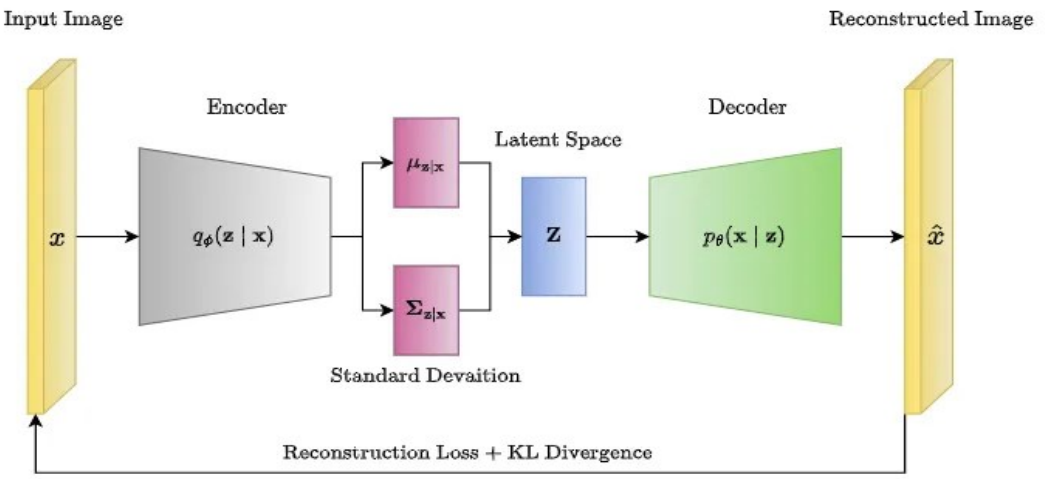

Figure 22 A screenshot of VAE architecture. (Shende, 2023)

When this technology is used in social media, it can raise a lot of ethical concerns. With the ability to create amazing content without putting in any effort and rapidly, it could lead to massive amounts of deceptive or malicious content.

Users can use TTS to make art that is influenced by known artists, which leads to a debate about the authenticity and originality. For example, a popular digital artist Greg Rutkowski's name has been used in TTS tools for more than 400 000 times since September 2022 without his consent, and he is also worried about his works in the future. (Hutchinson & John, 2023)

When it comes to copyrighting an image using TTS, in the United States of America, the images cannot be copyrighted according to the US Copyright Office (USCO), as the production is created by the technology and not the human user. (Holt, 2023)

3.2 Identification of AI Models

Identifying AI is becoming more challenging to spot as the technology keeps advancing at a rapid rate, making the tools more powerful.

When it comes to detecting text-based AI tools, the main things to look for are perfect grammar and spelling, where the AI uses a lot of rare phrases or odd words that might be out of place for the context. The importance of analysing the core text structure is to find anything unusual that is related to what the content is supposed to represents, such as a lack of context and specific details. (AIContentfy, 2024)

There are AI detection tools that can be used to detect AI-generated text. Although these tools are in an experimental state, they can still be used, but not reliably with an accuracy of 84 percent for premium tools and 68 percent for free tools to detect AI-generated text. (Caulfield, 2023)

Detecting AI-generated images or videos is not an easy task with a single glance. Digital art that looks a little bit suspicious, as in very unnaturally realistic or inconsistent with colours or details, it is advised to double-check the image and perform an image reverse search for pictures to check the original image if it exists, and for videos, it is advised to spot anything unnatural for humans for example, facial expressions, body movements, and body shape or posture. (Stouffer, 2023) Figure 23 explains simply how to spot a deepfake.

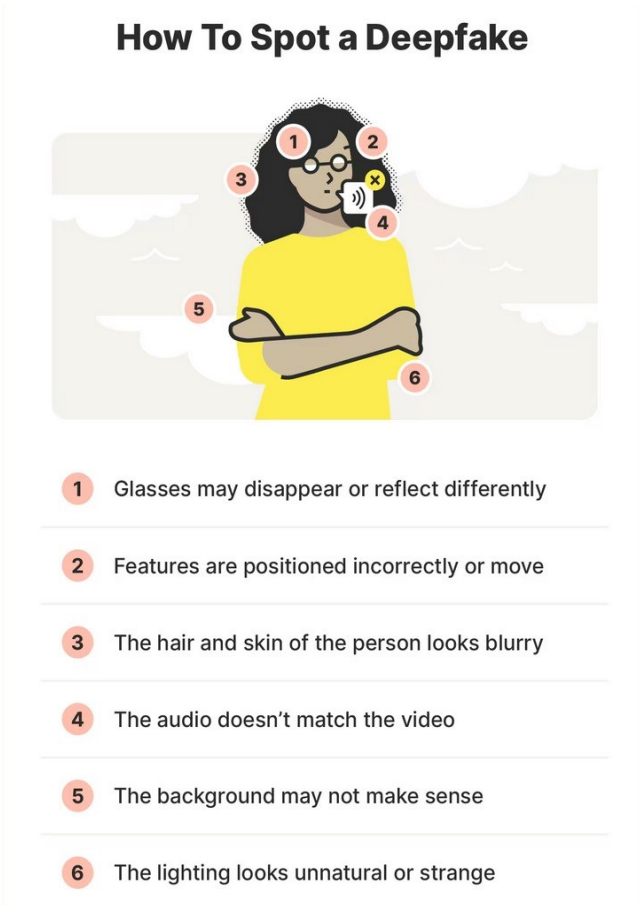

Figure 23 A screenshot of spotting deepfakes with a guide.

This is a reason why AI generative tools are very alarming on social media if a person is not able to differentiate between real and AI content. One study made a survey consisting of a thousand participants ranging from different ages to try and identify between real and AI-generated content of pictures and text copies, but with a restriction on not allowing using human images as they were easily

identifiable as AI-generated because of a drawback of AI. Figure 24 shows the participants of the survey.

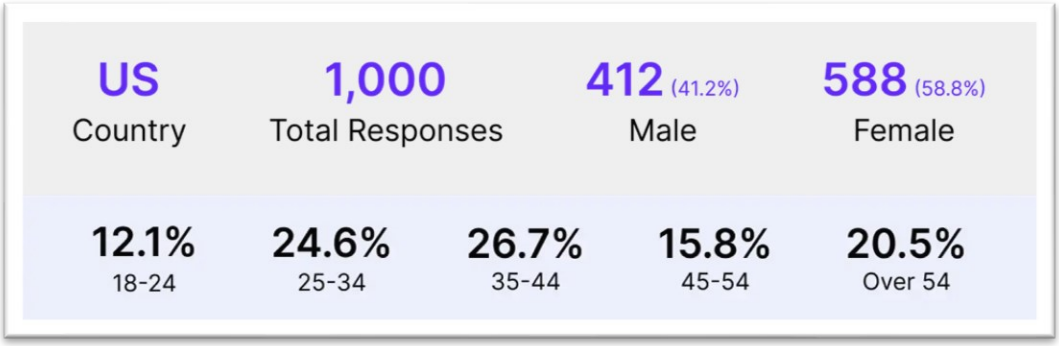

Figure 24 A screenshot of survey participants divided into five different age groups to differentiate images and texts between AI created and genuine. (Nexcess, n.d)

The result of this study is that the youngest age group were able to identify more accurately than other age groups, with 61.3 percent accuracy. For the overall results between the participants, it was easier to spot a text copy with 57.3 percent accuracy rather than images with 53.4 percent accuracy. (Nexcess, n.d.)

3.3 Classification of AI-generated Content

The answer to spotting real images on social media sites is to classify AI-generated content using classification models that has been developed for a long time to counter the spread of misinformation.

In this study to classify AI-generated content, we are collecting a few samples of real and AI-generated images from the internet, and we are going to preprocess the images and classify them by using a VGG16 Convolutional Neural Network (CNN) model.

The VGG16 is a good model to use for classification due to its effectiveness in image recognition. The model is not by any means perfect, it has its own flaws for instance it cannot always tell a very realistic looking AI image from a real one.

Therefore, these CNN models are always trained to keep up with the absurd number of high-quality AI images.

To use the VGG16 model, it is necessary to install libraries like TensorFlow and Keras and then write a script to preprocess the image and then classifying them, which will give you a prediction and accuracy of the image.

At first, we will load up the VGG16 model (listing 6).

```
model = VGG16(weights='imagenet', include_top=True)
```

Listing 6. A Python script to load up a classification model named VGG16.

After loading up a classification model, we will make a function to preprocess an image (listing 7).

```
def preprocess_image(img_path):

    img = image.load_img(img_path, target_size=(224, 224))
    img_array = image.img_to_array(img)
    img_array = np.expand_dims(img_array, axis=0)
    img_array = preprocess_input(img_array)
    return img_array
```

Listing 7. A function that is preparing for classification by resizing imags and converting them into a NumPy array.

And the last step is to classify the pre-processed images to give a prediction with accuracy (listing 8).

```
def classify_image(img_path):

    img = preprocess_image(img_path)
    predictions = model.predict(img)
    decoded_predictions = decode_predictions(predictions, top=3)[0]

    return decoded_predictions
```

Listing 8. A function that classifies preprocessed images and returns the top three predictions of the image.

For this image classification, we will be using three pictures of real and AI-generated images of fruits which are banana (Figure 25), cucumber Figure (26) and strawberry Figure 27.

Figure 25 A screenshot of an AI-generated banana (left) and a picture of a real banana (right).

Figure 26 A screenshot of an AI-generated cucumber (left) and a picture of a real cucumber (right).

Figure 27 A screenshot of an AI-generated strawberry (left) and a picture of a real strawberry (right).

Below are the results of the classification with the top three predictions, shown in Figure 28.

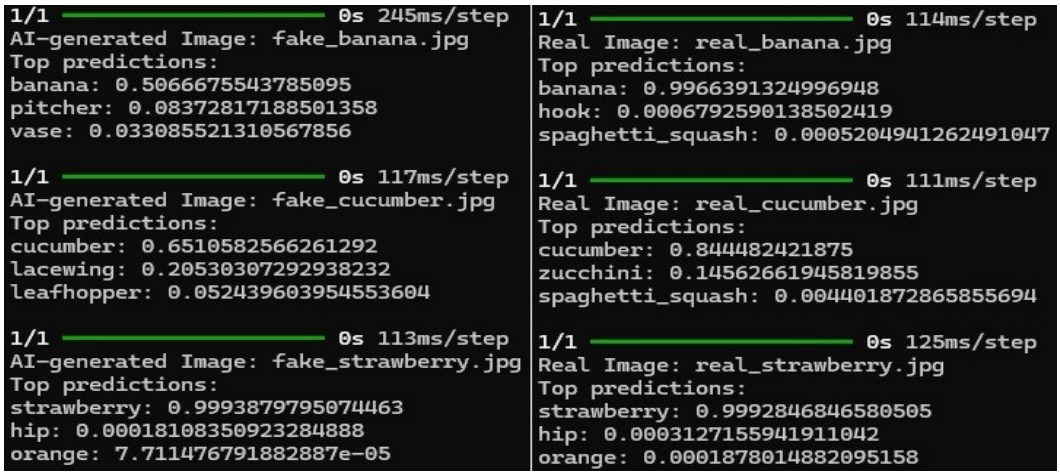

Figure 28 A screenshot of the classification results of AI-generated (left) and real images (right) with top the three predictions.

With a small sample size of images, we can conclude that the classification has its inconsistency in differentiating between AI-generated images and real images, as seen in Figure 28, where the results of the bottom row shows that the classification predicted fake and real images of strawberries with 99 percent accuracy as strawberry, which means that the tool was not able to detect the AI-generated picture of strawberries. Surprisingly, the AI-generated banana was predicted with 50 percent accuracy, as seen in the top left row of the results in Figure 28, considering how indistinguishable the fake banana's resemblance is from the real one within a single glance. Even though the real images have a high accuracy prediction compared to their counterparts, it is still advised not to always rely on classification tools. Therefore, these tools need to be trained a lot for better accuracy and reliability.

4 Results and Discussion

The VGG16 classification results shown in Figure 28 shows the capabilities of predicting the authenticity of an image, with certain accuracy depending on how much effort has been put on these models to be trained. Even with the small sample size that was demonstrated in this thesis, it shows that the tool is at least capable of predicting what the image could be. According to a research paper,

the VGG16 model's prediction accuracy resulted as the lowest accuracy compared to the other three models and it also took the most time to train. The test was conducted by predicting 22688 images and only accounting for the top-one accuracy. (Liu, 2023) Figure 29 shows the results for each model.

Model	Accuracy (%)	Training Time (minutes)
VGG-16	96.90	39
ResNet-50	98.85	28
MobileNet	97.95	25
SC-3	97.81	34

Figure 29 A screenshot of the accuracy and training time results for each model. (Liu, 2023)

The results from the research paper in Figure 29, shows that a highly trained CNN model delivers better accuracy in predicting images. Therefore, a new test was conducted using the previous implementation of classification in Chapter 3.3, where the VGG16 model was used to classify images of objects. Rather than using the VGG16 model, this test utilised the MobileNet model to classify 5 000 images of real and AI-generated pictures of faces. (Siddyn, 2022) Figure 30 shows the results of the prediction accuracy that was done using the MobileNet CNN model.

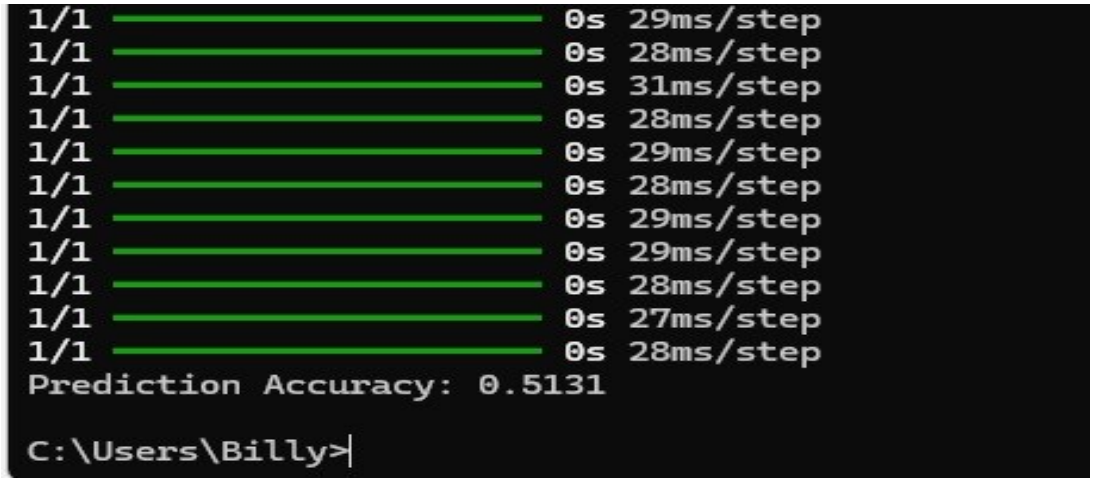

Figure 30 A screenshot of a result with 51 percent prediction accuracy.

The test prediction accuracy was very low at 51 percent, implying that the model was not able to differentiate between AI-generated and real images consistently. Therefore, these models require fine-tuning to produce a better result.

The diffusion results from Chapter 3.1.3 Figure 21 demonstrates that the reconstructed image of the original almost resembles each other with great fidelity, which suggests that the diffusion model successfully represents an input data and generates a faithful reconstruction. However, there are slight inconsistencies with the diffusion image for example, the blurriness in Figure 21, due to the limitations of the diffusion model.

5 Conclusion

In conclusion, this thesis explored the evolution of AI and its boom in present times. With the rapid growth of AI technology, users' awareness has also increased towards to AI content on the internet, as these contents are being generated daily due to the ease of access to free AI tools, which could lead to a possible malicious use. Therefore, regulating AI technology could help prevent the misuse of AI tools. Furthermore, the global AI market has grown exponentially with the purpose of boosting labour productivity, which could lead to possible replacement of jobs.

Detecting highly detailed fake content is not an easy task. However, one of the best tools to check the originality of possible manipulated content is the internet and using it to search for the source of the content, while low effort AI contents are usually detectable with bare eyes since they are usually distorted.

The future goals are researching different AI models and techniques in more depth to have a better understanding of the architecture and experiment with them further to produce more accurate results, with the aim of showcasing the potential of the tools.

Finally, the purpose of this thesis is to raise awareness about the large amount of deepfakes and AI-generated content on social media platforms that has the potential to manipulate users into something that is not real and how to be more aware of manipulated content.

References

1 Ahvanooey, M. T, Shoaib, M. R, Wang, Z., Zhao, J. (2023) Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. 2023 International Conference on Computer and Applications (ICCA) (pp. 1-7). IEEE.

2 AIContentfy. (2024) Unmasking AI: A Guide to Detecting Artificially Generated Content. [Internet]. AIContentfy. Available from: https://aicontentfy.com/en/blog/unmasking-ai-guide-to-detecting-artificially-generated-content [Accessed 29,april].

3 Alhabeeb, S. K, Al-Shargabi, A. A. (2024) Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction. In IEEE Access (vol. 12, pp. 24412-24427).

4 Basheer, K. C. S. (2023) How to Detect and Handle Deepfakes in the Age of AI? [Internet]. Analytics Vidhya. Available from: https://www.analyticsvidhya.com/blog/2023/05/how-to-detect-and-handle-deepfakes-in-the-age-of-ai/ [Accessed Apr. 2024].

5 Belanger, A. (2023) Thousands scammed by AI voices mimicking loved ones in emergencies. [Internet]. Ars Technica. Available from: https://arstechnica.com/tech-policy/2023/03/rising-scams-use-ai-to-mimic-voices-of-loved-ones-in-financial-distress/ [Accessed Apr. 2024].

6 Boheemen, P. V., Das, D., Fatun, M., Gerritsen, J., Huijstee, M. V., Jahnel, J., Karaboga, M., Kool, L., Nierling, L. (2021) Tackling deepfakes in European policy. [Internet]. European Parliament. Available from: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf [Accessed April, 2024].

7 Butazzo, G. (2023) Rise of Artificial general intelligence: risks and opportunities. [Internet]. Frontiers. Available from: https://www.frontiersin.org/articles/10.3389/frai.2023.1226990/full [Accessed Mar. 2024].

8 Caulfield, J. (2023) How Do AI Detectors Work? | Methods & Reliability. [Internet]. Scribbr. Available from: https://www.scribbr.com/ai-tools/how-do-ai-detectors-work/. [Accessed Apr. 2024].

9 Chess. (N.d.). Deep Blue (Chess Computer). [Internet]. Available from: https://www.chess.com/terms/deep-blue-chess-computer [Accessed Apr. 2024].

10 Coe. (N.d.). History of Artificial Intelligence. [Internet]. Available from: https://www.coe.int/en/web/artificial-intelligence/history-of-ai [Accessed Mar. 2024].

11 Conte, N. (2024) Ranked: The Most Popular AI Tools. [Internet]. Visual Capitalist. Available from: https://www.visualcapitalist.com/ranked-the-most-popular-ai-tools/ [Accessed Mar. 2024].

12 Editor, Davey, M (Ed.). (2023) 'Alarming': convincing AI vaccine and vaping disinformation generated by Australian researchers. [Internet]. The Guardian. Available from: https://www.theguardian.com/australia-news/2023/nov/14/alarming-convincing-ai-vaccine-and-vaping-disinformation-generated-by-australian-researchers [Accessed Apr. 2024].

13 Editor, Jain, A., & Editor, Maheshwari, R (Eds.). (2024) Top AI Statistics And Trends. [Internet]. Forbes. Available from: https://www.forbes.com/advisor/in/business/ai-statistics/ [Accessed Mar. 2024].

14 Editor, Smith, M (Ed.). (2018) This face-swap video of Donald Trump could be the terrifying future of fake news. [Internet]. Mirror. Available from: https://www.mirror.co.uk/news/politics/face-swap-video-donald-trump-12059215 [Accessed Apr. 2024].

15 Fortinet. (N.d). What is A Deepfake? [Internet]. Available from: https://www.fortinet.com/resources/cyberglossary/deepfake [Accessed].

16 Frolov, D. B, Makhaev, D. D, Shishkarev, V. V. (2022) Deepfakes and Information Security Issues. 2022 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS) (pp. 147-150). IEEE.

17 Gadwal, P. S. (2023) The Next Frontier in AI Creativity: Text-to-Image Synthesis. [Internet]. Medium. Available from: https://medium.com/@premsaig1605/the-next-frontier-in-ai-creativity-text-to-image-synthesis-aab9f40ec4bf [Accessed Apr. 2024].

18 GAO. (2020) Deconstructing Deepfakes—How do they work and what are the risks? [Internet]. Available from: https://www.gao.gov/blog/deconstructing-deepfakes-how-do-they-work-and-what-are-risks [Accessed Apr. 2024].

19 Georgieva, K. (2024) AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity. [Internet]. IMF. Available from: https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity [Accessed Apr. 2024].

20 Grand View Research. (2024) AI In Healthcare Market Size, Share & Trends Analysis Report By Component (Hardware, Services), By Application, By End-use, By Technology, By Region, And Segment Forecasts, 2024 – 2030. [Internet]. Available from: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market [Accessed Apr. 2024].

21 Guinness, H. (2024) How to use DALL·E 3 to create AI images with ChatGPT. [Internet]. Zapier. Available from: https://zapier.com/blog/dall-e-3/ [Accessed Apr. 2024].

22 Holt, K. (2023) AI-generated images from text can't be copyrighted, US government rules. [Internet]. Engadget. Available from: https://www.engadget.com/ai-generated-images-from-text-cant-be-copyrighted-us-government-rules-174243933.html [Accessed Apr. 2024].

23 Hutchinson, C., John, P. (2023). AI: Digital artist's work copied more times than Picasso. [Internet]. BBC. Available from: https://www.bbc.com/news/uk-wales-66099850 [Accessed Apr. 2024].

24 IBM. (N.d.). Deep Blue. [Internet]. Available from: https://www.ibm.com/history/deep-blue [Accessed Apr. 2024].

25 Klawans, J. (2024) The push for media literacy in education amid the rise of AI. [Internet]. The Week. Available from: https://theweek.com/tech/media-literacy-AI-schools [Accessed Apr. 2024].

26 Lawton, G. (2024) What is generative AI? Everything you need to know. [Internet]. TechTarget. Available from: https://www.techtarget.com/searchenterpriseai/definition/generative-AI [Accessed Apr. 2024].

27 Liu, K. (2023) Comparison of different Convolutional Neural Network models on Fruit 360 Dataset. [Internet]. Research Gate. Available from: https://www.researchgate.net/publication/369470221_Comparison_of_different_Convolutional_Neural_Network_models_on_Fruit_360_Dataset [Accessed May. 2024].

28 Moon, M. (2023) A new AI voice tool is already being abused to make deepfake celebrity audio clips. [Internet]. Engadget. Available from: https://www.engadget.com/ai-voice-tool-deepfake-celebrity-audio-clips-094648743.html?guccounter=1 [Accessed Apr. 2024].

29 NBC News. (2024). LISTEN: Fake Biden robocall tells voters to skip New Hampshire primary. [Internet]. Available from: https://www.nbcnews.com/video/listen-fake-biden-robocall-tells-new-hampshire-not-to-vote-in-primary-202609733664 [Accessed Apr. 2024].

30 Negoita, J. (2024) Deepfake AI Face Swap. [Internet]. Medium. Available from: https://medium.com/@codingdudecom/deepfake-ai-face-swap-e11edc7d67e1 [Accessed Apr. 2024].

31 Nexcess. (N.d.). AI vs. human study: Can consumers tell the difference between AI and human-generated content? [Internet]. Available from: https://www.nexcess.net/resources/ai-vs-human-study [Accessed Apr. 2024].

32 Norden, L., Weiner, D. I. (2023) Regulating AI Deepfakes and Synthetic Media in the Political Arena. [Internet]. Brennan Center. Available from: https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena [Accessed Apr. 2024].

33 O'Connor, R. (2022) Introduction to Diffusion Models for Machine Learning. [Internet]. AssemblyAI. Available from: https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/ [Accessed Apr. 2024].

34 O'Sullivan, D., Passantino, J. (2023) 'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion. [Internet]. CNN. Available from: https://edition.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html [Accessed Apr. 2024].

35 Podcastle. (2023) The Complete Guide to AI Voices: Everything You Need to Know. [Internet]. Available from: https://podcastle.ai/blog/the-complete-guide-to-ai-voices-everything-you-need-to-know/ [Accessed Apr. 2024].

36 Reddy, R. (2023) 24 Deepfake Statistics – Current Trends, Growth, and Popularity (December 2023). [Internet]. Contentdetector AI. Available from: https://contentdetector.ai/articles/deepfake-statistics [Accessed May. 2024].

37 Shende, R. (2023) Autoencoders, Variational Autoencoders (VAE) and β-VAE. [Internet]. Medium. Available from: https://medium.com/@rushikesh.shende/autoencoders-variational-autoencoders-vae-and-%CE%B2-vae-ceba9998773d [Accessed Apr. 2024].

38 Siddyn. (2022) batched AI generated and real images (140k FvR). [Internet]. Kaggle. Available from: https://www.kaggle.com/datasets/siddyn/new-fake/data [Accessed May, 2024].

39 Stouffer, C. (2023) What are deepfakes? How they work and how to spot them. [Internet]. Norton. Available from: https://us.norton.com/blog/emerging-threats/what-are-deepfakes [Accessed Apr. 2024].

40 Tableau. (N.d.). What is the history of artificial intelligence (AI)? [Internet]. Available from: https://www.tableau.com/data-insights/ai/history [Accessed Apr. 2024].

41 Teale, C. (2024) AI misinformation a 'whole new area' for elections officials to deal with. [Internet]. Route Fifty. Available from: https://www.route-fifty.com/digital-government/2024/02/ai-misinformation-whole-new-area-elections-officials-deal/393962/ [Accessed Apr. 2024].

42 The Telegraph. (2022) Deepfake video of Volodymyr Zelensky surrendering surfaces on social media. [Internet]. Available from: https://www.youtube.com/watch?v=X17yrEV5sl4 [Accessed May. 2024].

43 YeezyBeaver. (2023) Hey There Delilah but it's Kanye's Voice (So Vits SVC). [Internet]. Available from: https://www.youtube.com/watch?v=-9Ado8D3A-w [Accessed Apr. 2024].

44 Zaluski, A. (2023) Using AI-generated content: An in-depth analysis. [Internet]. Notion. Available from: https://www.notion.so/blog/ai-generated-content [Accessed Apr. 2024].