البرنامج الوطني للذكاء الاصطناعي
NATIONAL PROGRAM FOR ARTIFICIAL INTELLIGENCE

United Arab Emirates

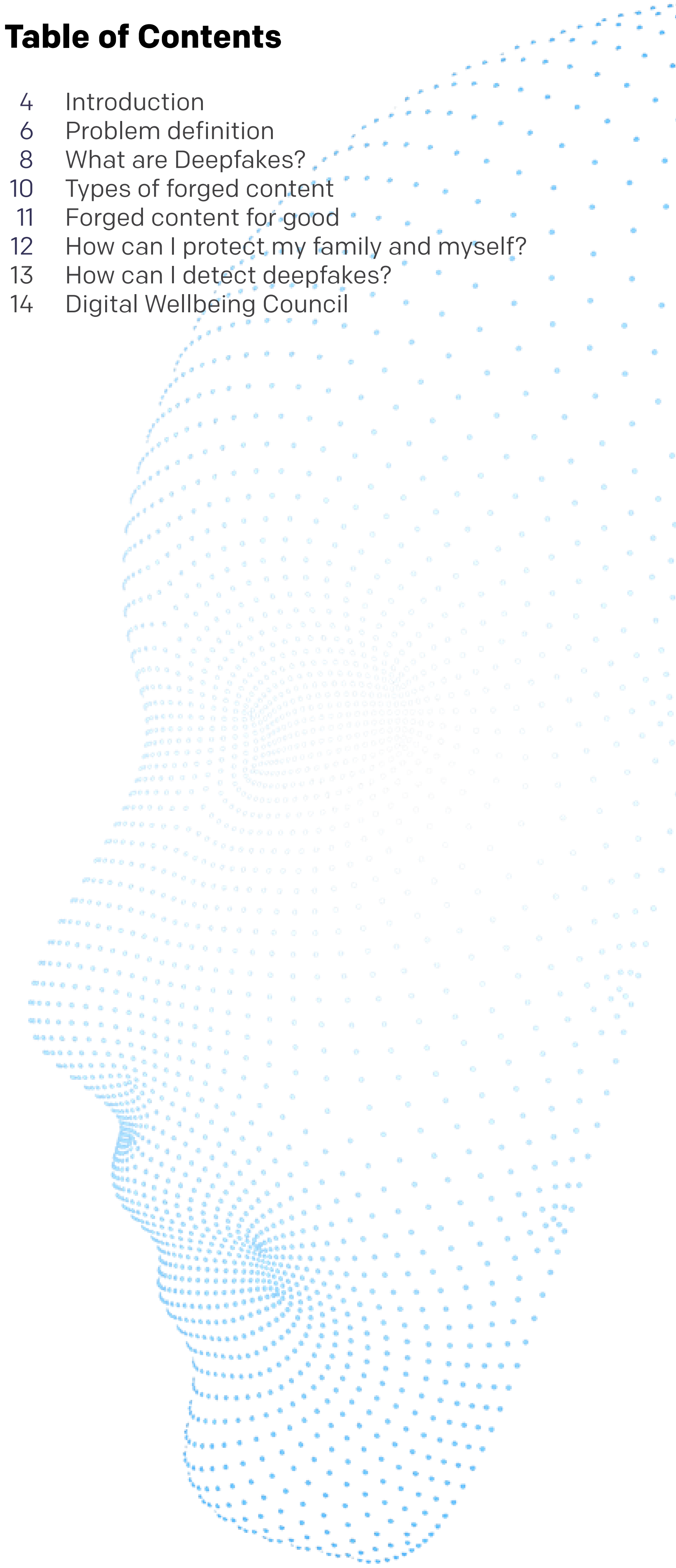
DEEPFAKE GUIDE

JULY 2021

To read this document digitally

www.ai.gov.ae

Table of Contents



INTRODUCTION

Artificial Intelligence (AI) technologies and Machine Learning (ML) have supported myriad applications in human life, of which the effects we see in our daily lives. However, there has been a more visible development in the field of AI and computer vision in recent years, focusing on using AI algorithms to generate or modify audio and video content. A unique issue about these newly developed AI algorithms is that they can generate real video and audio content that was never captured (nor occurred in reality). For example, AI experts can create video and audio content with public figures, celebrities and prominent political figures performing humorous acts. These technologies can now move beyond this superficial application. The entertainment industry heavily uses AI in creating special effects in movies, catchy advertisements, and enhanced visuals.

These new technologies also found use in the medical industry, creating synthetic audio for patients who suffer from vocal cord damage that affects their ability to talk. In this case, the AI algorithm "imagines" how the patient sounds and creates audio files that sound as if the patient is speaking naturally.

The most common term that describes AI systems that generate these imaginary video and audio files is called "**deepfake**". A deepfake signifies content that is fake and created using deep learning[1]. It has no basis in reality, even though it appears very realistic and convincing. Looking at the history of AI, never before were AI algorithms able to generate content; they used to be simple tools that analyzed and learned from data such as speech recognition and robotics. However, in 2014[2] there was a breakthrough. A complex deep learning-based technique codenamed Generative Adversarial Networks (GANs) was developed; it learned from video and audio data before generating its own. This advancement was the key to allow AI systems to produce new data (imaginary data). GANs opened the door for many applications, including what is now known as "deepfakes".

When deepfakes first appeared, it was a complicated process and hard to create. However, with the advancement of software technology, it became easier for non-professionals, to create modified audio and video files. The technology became more accessible and cheaper, requiring less data and computational power. It was also much faster to create. On the one hand, this contributed to the increased spreading of modified audio and videos for entertainment. On the other hand, it also allowed to exploit such technologies for creating offensive or fake audio/video, fake news, and contents dubbed deepfakes, mostly as a form of cyberbullying.

As this technology is hard to detect and identify, this guide aims to define the problem of deepfakes. It also presents protection measures and how to report identified deepfakes to the appropriate authority.

[1] Deep learning is a machine learning technique that has proven very useful in Audio/Visual applications including detection of objects, recognition, and video tracking applications.

[2] Based on MIT Technology Review and BBC available on: https://www.technologyreview.com/2014/12/29/169759/2014-in-computing-breakthroughs-in-artificial-intelligence/ and https://www.bbc.co.uk/teach/ai-15-key-moments-in-the-story-of-artificial-intelligence/zh77cqt

PROBLEM DEFINITION

Deepfakes are a new form of an old problem related to the distribution of fake media content. Compared to the previous image and audio editing tools used by average individuals, this media is now being developed using Artificial Intelligence Machine Learning technology, making it much closer to reality. The utilization of deepfake technology makes it possible to generate content (video and audio) to impersonate people, offering false information on their behavior, their activities, and the environment.

Deepfakes become a threat when the technology is used as a tool to create and distribute fake media and false information about individuals, officials, and key figures doing and saying things that never happened. People can be placed in imaginary environments, situations, and events.

Deepfakes could be exploited for malicious purposes, including:

- Causing damage to the reputation of individuals and countries.
- Manipulating public opinion for the intention of causing disruptions.
- Imposing a lack of trust by using deepfake reality as part of a plausible truth.
- Creating fabricated evidence to influence a legal judgment.

Deepfake videos have gone viral in recent years, giving millions around the world their first taste of this new technology. They were very impactful as the subject of such bullying was celebrities, politicians, and many other public figures. Creating deepfakes of celebrities is surprisingly easy because all the data needed are readily available from various media sources.

The rapid development in deepfake technology made it more available, accessible, easier to use, cheaper, and faster. Moreover, newer and powerful tools are increasingly available as less data and computational power are needed, resulting in a rise in the quality and quantity of deepfakes. Deepfakes affect individual reputations and harm an organization's stability and national interest, inflicting a reputational threat by falsifying behaviors, activities, and events. They can show a person or people are involved in an act that never happened. Additionally, deepfakes can negatively impact at a national level by establishing a public agenda to influence communities. It can manipulate and influence public opinion, especially before an event. It also may impose a reputational threat to nations as well as cause disruptions to international and diplomatic relations if not verified promptly by concerned governments.

Existing UAE national laws prohibit cyberbullying, actual malice, and identity impersonation. Also, specific criteria exist for producing, distributing, and publishing or broadcasting media content that contains fake news, disregard for individual privacy, and religious norms and traditions.

Internationally, some jurisdictions lack such legislation and face an outbreak of deepfakes that disrupt political systems and harm individual privacy and personal lives. Therefore, nations started to issue legislation that prohibits deepfakes, but it is overly simplistic to assume it will be enough of a deterrent or easy to enforce.

Face swaping applications process the user's facial data and replaces celebrities' faces with the user's utlising advanced AI algorithms.

Imagined by a GAN (generative adversarial network)StyleGAN2 (Dec 2019) - Karras et al. and Nvidia (https://thispersondoesnotexist.com)

Face swapping has become popular entertainment on mobile devices; however, it carries risks offending viewers feeling and causing gender or racial dilemmas.

WHAT ARE DEEPFAKES?

It has become very easy to produce deepfake content and it is getting harder to determine if a celebrity in a video is acting him/herself.

The legal definition of deepfakes is broad. A recent bill in the state of Illinois defines a deepfake as a video "created with the intent to deceive, that appears to depict a real person performing an action that did not occur in reality [3]" Moreover, deepfakes can be defined as manipulated visual and/or audio content based on Artificial Intelligence and advanced software technology to misrepresent individuals, objects, environments, and events. These manipulations seem close to reality, and the general public may find them hard to detect.

The concept of manipulating photos, audio, and videos is not new and has been available for years. But with the recent technological advancements of AI, ML, and software commercialization, it has become easier to access and create forged content using desktop and mobile apps in a myriad of ways.

Forged contents can be categorized mainly as:

1. **Shallow fakes:**
   a. Slow-motion clips: Clips using video editing software to slow down the manner of speech without changing the audio pitch. This may indicate an abnormality of the targeted person in the video or place stress on certain words or voice inflection to imply false perspectives.
   b. Changing dates and locations: manipulating dates and locations to appear as modern clips and in different places. This leads to the spread of false news that harms the safety of society and the individual.
2. **Deepfakes:** The act of face swapping and replacement using ML and AI technologies through extracting images of the source and targeted individuals into frames, training the AI model of these extracted images into two neutral networks, then reconstructing the new face, creating the desired video. The same process can be used for audio.

[3] State of Illinois – United States Deepfake bill: 10 ILCS 5/29-21

Good quality deep fakes require repetitive training and more data

with the recent technological advancements of AI, ML, and software commercialization, it is easy to access and create forged content using desktop and mobile apps in a myriad of ways.

Photo editing applications powered by machine learning (ML). None of these people are real. All these images were fabricated using AI and GANs.

TYPES OF FORGED CONTENTS

The most common applications of deepfakes are associated with the following:

1. **Visual**: describes the use of deepfakes in creating images and videos.

   a. Face swapping using encoder/decoder algorithms to digitally map a person's face onto another:

   The encoder algorithm runs thousands of screenshots to study the facial features of two people. It then finds the similarities between those two faces, reduces them to the shared features, and compresses the images. A second AI algorithm called a decoder is then taught to recover the faces from the compressed images. Because the faces are different, one decoder is programmed to recover the first person's face, and another decoder is trained to recover the second person's face. To perform the face swap, the encoded images are simply fed into the "wrong" decoder.

   b. Facial Manipulating such as expression modification and lip-syncing using Generative Adversarial Networks (GANs):

   This method pits two artificial intelligence algorithms against each other. The first algorithm, known as the generator, is fed random noise and turns it into an image. This synthetic image is added to a stream of real images—of celebrities, for example— and then fed into the second algorithm, known as the discriminator. At first, the synthetic images will look nothing like faces. But repeating the process countless times, with feedback on performance, and the discriminator and generator improve. Given enough cycles and feedback, the generator starts producing utterly realistic faces of completely non-existing people.

2. **Audible**: Mainly in voice synthesis and modification by either generating a voice-over and creating a new speech that has never been said or by controlling the tone of a person to show unreal emotions or behavior.

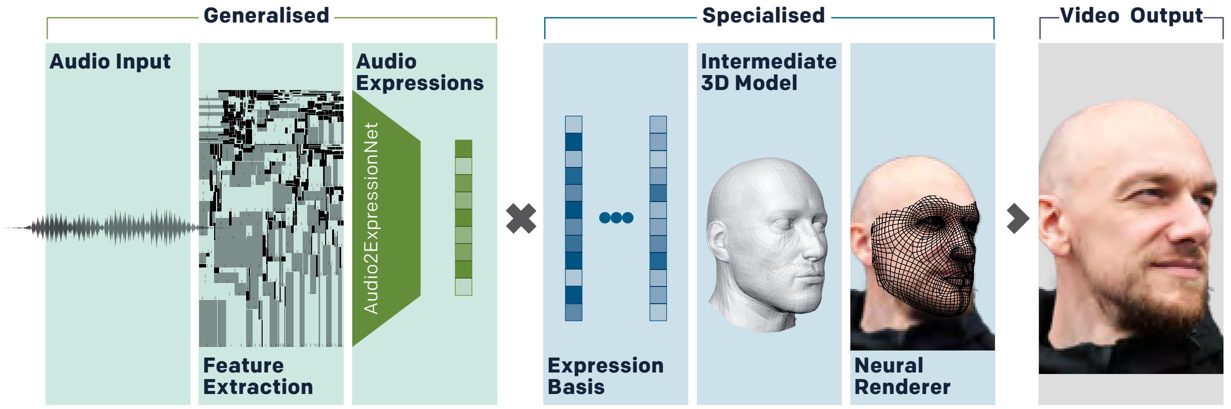
Deepfakes impose significant risks due to manipulating truths and harming reputations by broadcasting such messages on various media channels anonymously.

While training AI algorithms to create deepfake technologies may seem complicated, the internet offers people several easy-to-use applications that create instant deepfakes on their phones and computers for entertainment and other purposes. The constant usage of such applications allows continous training for deepfake AI algorithms.

FORGED CONTENTS FOR GOOD

While we are led to believe that deepfakes are evil, it is important to note that deepfakes cannot be categorized as good or bad. Deepfakes are merely a tool that can be used for different purposes. There are many positively focused applications of deepfakes in different industries, some of these examples include the following:

1. **Medical applications**
   - Generation of new images such as MRI images for training purposes
   - Synthesize Audio for Articulatory Based Speech (ALS) for those who lost their ability to speak due to cancer and other medical conditions that affect the vocal cords.
2. **Entertainment**
   - Film industry and advertisements: enhancement of media content and creation of special visual effects to create motion scenes and face manipulation.
   - News anchor - virtual news presenter
3. **Customer services**
   - Virtual assistant - there is a trend of using virtual assistant audio and video to provide customer service for call centers.



Using voice generator and facial expression manipulation to generate believable videos of the target.

HOW CAN I PROTECT MY FAMILY AND MYSELF?

Deepfake technology relies on obtaining a large amount of data to train an AI system and develop deepfake audio and video clips. This data can be in the form of:

Pictures of people

Video clips

Audio clips

This data is often spread widely on social media. Generally speaking, the more data a malicious person obtains, the greater the quality of the deepfake audio and video. Resulting in a greater chance of experiencing a deep falsification that is close to reality. Like any algorithm, AI deepfake technology needs to be trained to perform. Therefore, people must be aware that the more they use these different applications and the more they post pictures online, the more data they are providing for potential deepfake framing material.

> **the more data a malicious person obtains, the greater the quality of the deepfake audio and video quality**

It is important to raise awareness about the dangers of exposing one's identity on different internet platforms, especially to the younger generation.

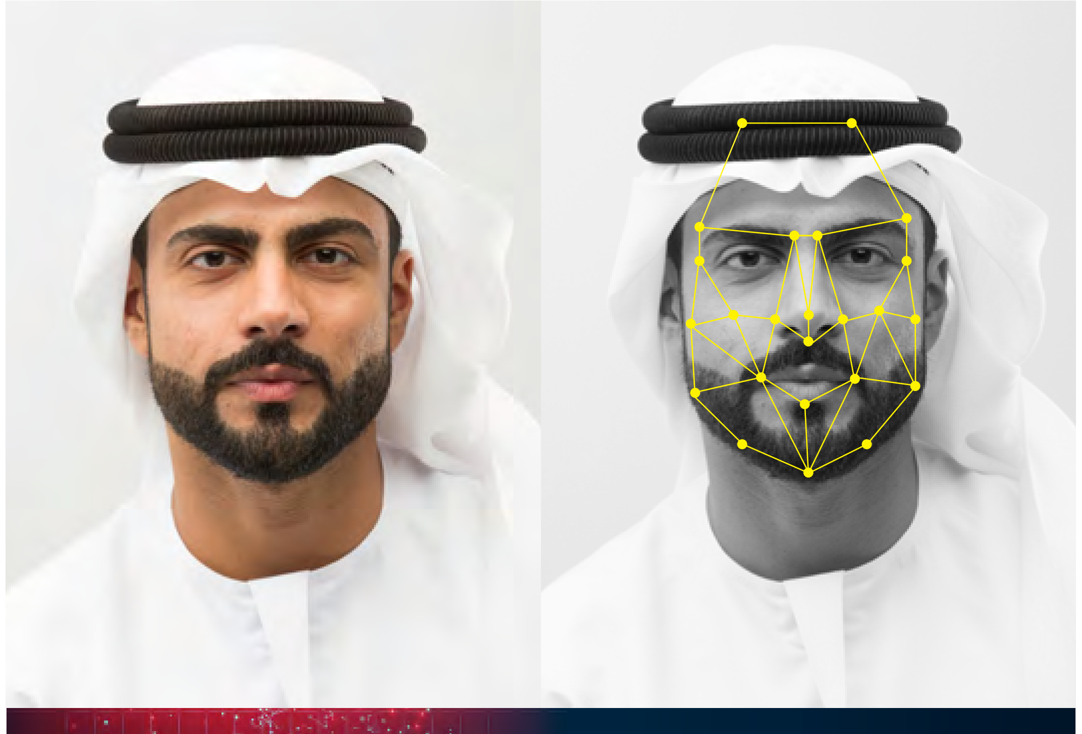
HOW CAN I DETECT DEEPFAKES?

As mentioned in this document earlier, the availability of huge amounts of data in the form of pictures and videos has allowed AI systems to be trained to create deepfakes that are so close to reality. However, it is possible for a human eye to detect signs that suggest whether a video content is forged or not. Some of these signs include the following:

1. Disorganized and irregular facial movement of the person
2. A sudden difference in the lighting directed at the person
3. Skin color change during the clip
4. Repetitive blinking or no blinking at all
5. Lip movements do not match the audible speech
6. Distortion in the area surrounding the face

Although physiological examination of videos is possible, it can be slow and unreliable. The most accurate approach to detect forged contents is through a systematic screening of the deepfakes using AI-based tools that need to be regularly updated.

Numerous research is being conducted on using AI to detect deepfakes.



DIGITAL WELLBEING COUNCIL

The Digital Wellbeing Council was established based on the directives of His Highness Sheikh Saif bin Zayed Al Nahyan, UAE Deputy Prime Minister and Minister of Interior, consists of several members on federal and local government levels. In which National Program for Artificial Intelligence is considered as one of its key members with a vital role in enabling digital wellbeing through the field of AI, digital economy and the different areas of remote work. The aim of the council is to create a safe digital community in the UAE, with a positive and productive identity in the digital life.

To learn more about the Digital Wellbeing Council's role, visit the website:

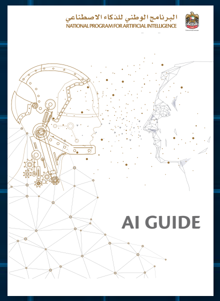
www.digitalwellbeing.ae

جودة الحياة الرقمية
DIGITAL WELLBEING

OTHER PUBLICATIONS

To read other publications released by the National Program for Artificial Intelligence please scan the relevant barcode.

AI GUIDE	BLOCKCHAIN GUIDE	UAE NATIONAL STRATEGY FOR ARTIFICIAL INTELLIGENCE 2031	AI HARDWARE INFRASTRUCTURE REPORT UAE



DEEPFAKE GUIDE