



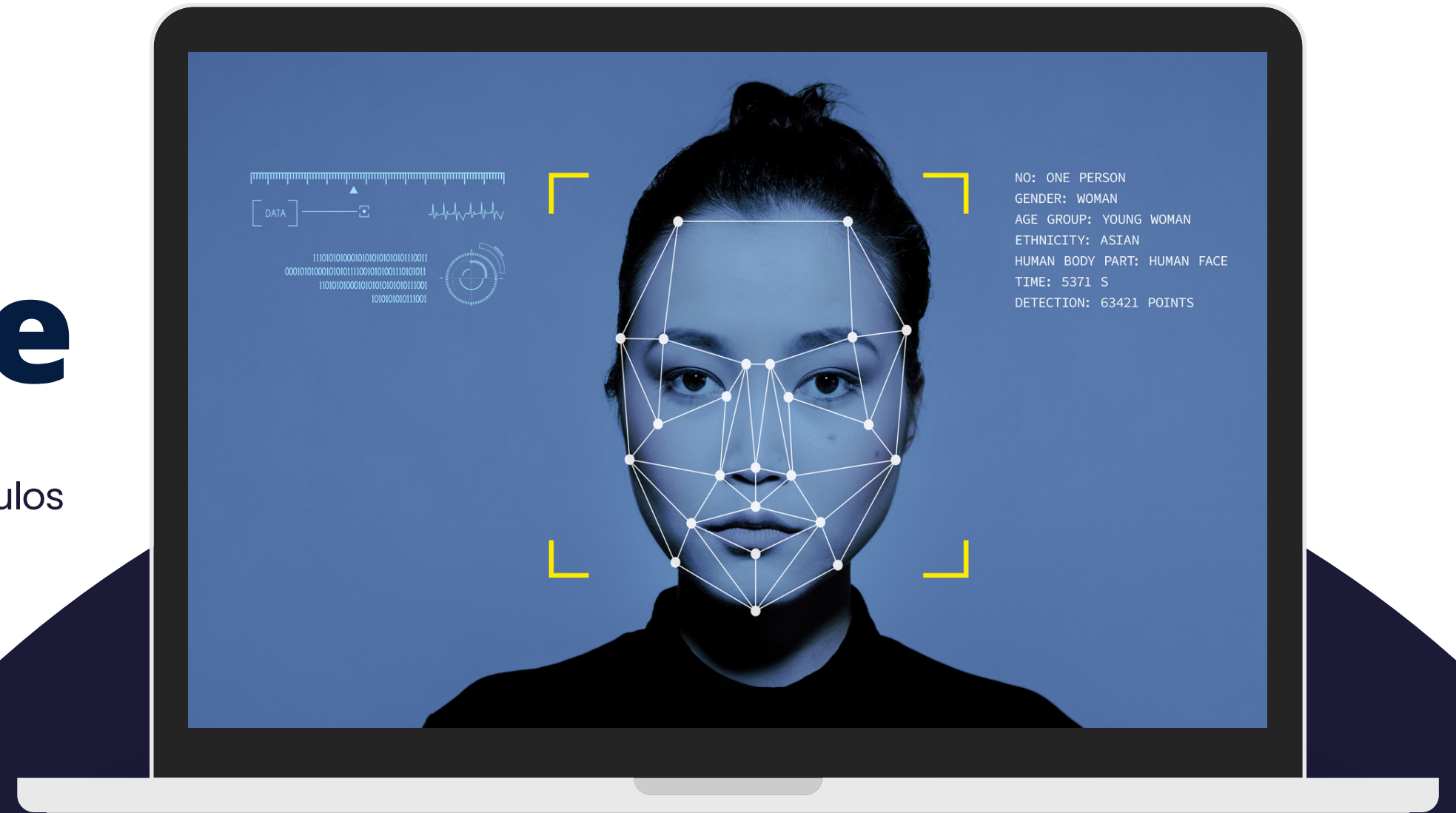
**FACT-CHECK**  
by MedDMO

**Monitoring Disinformation  
in Cyprus, Greece & Malta**

# The Deepfake knowledge base

By: Nikos Sarris, Efthimios Hamilos, Zoi Palla and Symeon Papadopoulos

*Media Analysis, Verification and Retrieval Group (MeVer)  
Information Technologies Institute (ITI)  
Centre for Research and Technology-Hellas (CERTH)*



# What are deepfakes?

Deepfakes are manipulated media, including videos, images, or audio, that have been created or altered using artificial intelligence (AI) and deep learning techniques. The term "deepfake" is derived from the combination of "deep learning" (a subset of machine learning) and "fake."

Deepfakes leverage powerful AI algorithms, particularly generative adversarial networks (GANs), diffusion models, or AI algorithms based on neural radiance fields (NeRF), to replace or superimpose the original content with new content that appears realistic but is actually synthesised or manipulated. These algorithms analyse and learn from large datasets of images or videos to understand patterns, features, and context. They then generate new content by combining and altering elements from the source material.

While deepfakes have some legitimate applications, such as in the entertainment industry or creating realistic virtual avatars, they also raise concerns due to their potential misuse. Deepfakes can be used to spread disinformation, defame individuals, manipulate political events, or deceive people by creating convincing fake content.

As a response to the ethical and security concerns associated with deepfakes, researchers are developing detection techniques and working on ways to mitigate the negative impact of this technology.


It is crucial for individuals to be mindful of the existence of deepfakes and to approach online media with a critical mindset when evaluating its authenticity.

# How are deepfakes generated?

Deepfakes are generated using applications based on artificial intelligence and particularly deep learning techniques. Several applications have been developed for that purpose such as [Midjourney](#), [DALL-E 2](#), [Stability.ai](#), [Synthesia](#), [DeepBrain](#), [Runway](#), [Craiyon](#), [Reface](#), [Face Swap](#), [DeepFaceLab](#), and [Face Swapper](#), to name a few.

More specifically, to generate a deepfake, deep neural networks called generative adversarial networks (GANs) or autoencoders, are trained on the collected dataset. A GAN is a machine learning model in which two neural networks compete with each other to become more accurate in their results. These models learn to understand the facial features, expressions, and other characteristics of the target person. Another approach to create synthetic content is through diffusion models that focus on the process of iteratively refining a random noise input to generate high-quality samples. There is also a third emerging category to create synthetic media based on Neural Radiance Fields (NeRF). The key idea behind NeRF is to model a continuous volumetric function that represents the radiance (color and lighting) at any given 3D spatial location. This function is parameterized by a neural network, which takes as input the 3D coordinates and outputs the corresponding radiance values. By training the network using a large dataset of 2D images or videos, NeRF can infer the underlying 3D scene geometry and appearance.

Furthermore, there are also several “traditional” editing tools or techniques for creating manipulated content that involve manual editing and visual effects, without using AI. Such tools could include:

- 
- ▶ Image Editing Software: Programs like GIMP, or Pixlr enable users to perform tasks such as resizing, cropping, adjusting colours, and blending elements together.
  - ▶ Video Editing Software: Applications like Adobe Premiere Pro, Final Cut Pro, or DaVinci Resolve allow you to edit and manipulate video content. You can cut and splice footage, apply effects, adjust colours, and overlay elements.
  - ▶ Masking and Compositing: Techniques like masking, where specific areas of an image or video are selectively hidden or revealed, can be used to blend elements together. Compositing involves combining multiple visual elements into a single image or video frame. Some AI based methods also achieve the same result.
  - ▶ Motion Tracking and editing: Motion tracking software can be used to track the movement of objects or faces and apply manual adjustments. Some AI based methods also achieve the same result.

# How can you spot a deepfake?

01

## Visual deepfake detection guide

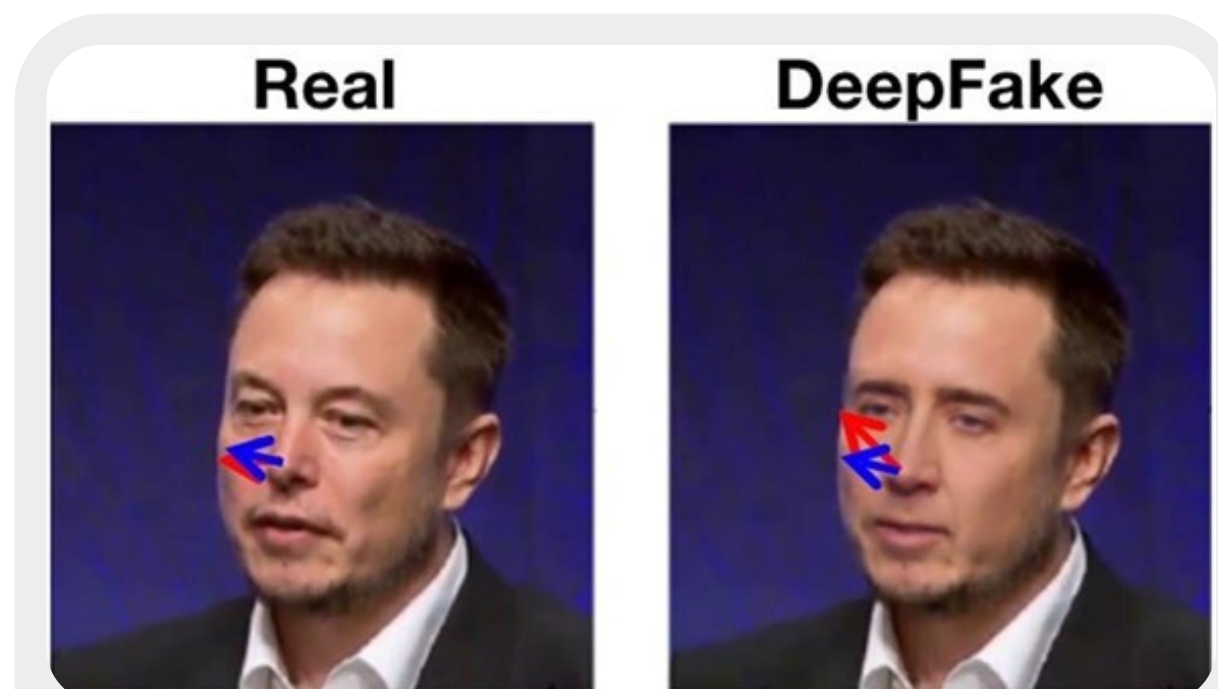
Detecting deepfakes without specific tools or software can be challenging, as deepfake techniques are designed to be convincing and the technology behind them is in constant advance. However, there are several techniques you can use to try to identify potential deepfakes.

02

## IT tools that can help in deepfake detection

There are several tools that can assist users in assessing the possibility of an image, video, or audio item being a deepfake. These range from simple tools that can help you find earlier (untampered) published versions of the item in question, to advanced tools that utilise the same deepfake-production technology to identify if it has been used to manipulate an item. Some of these tools are free for anyone to use and some require user registration or even payment.

- **Pay attention to facial expressions and body movements**



**When a computer puts Nicolas Cage's face on Elon Musk's head (Source)**

Deepfakes often struggle to perfectly mimic natural human body movements and facial expressions. Look for awkward or inconsistent positioning of head and body, and inconsistencies or unnatural body movements, such as blinking irregularities, strange head movements, or odd facial reactions. In the following example we can observe an attempt to create a deepfake that blends Nicolas Cage's face on Elon Musk's head. The arrows show the inconsistencies as the face and the head are not lined up correctly.



**Spot a deepfake through the facial expression (Source)**

Another example is depicted in the following figure, where Mona Liza and other well-known paintings have been manipulated by altering the original facial expressions. In cases where the image resolution allows a closer inspection by zooming into the facial characteristics, unnatural expressions and/or odd misalignments within the facial area, can raise suspicions on the potential of dealing with a deepfake.

- **Examine the eyes and reflections**



**Detecting a deepfake by observing misaligned eye reflections (image created with the use of DALL-E 2)**

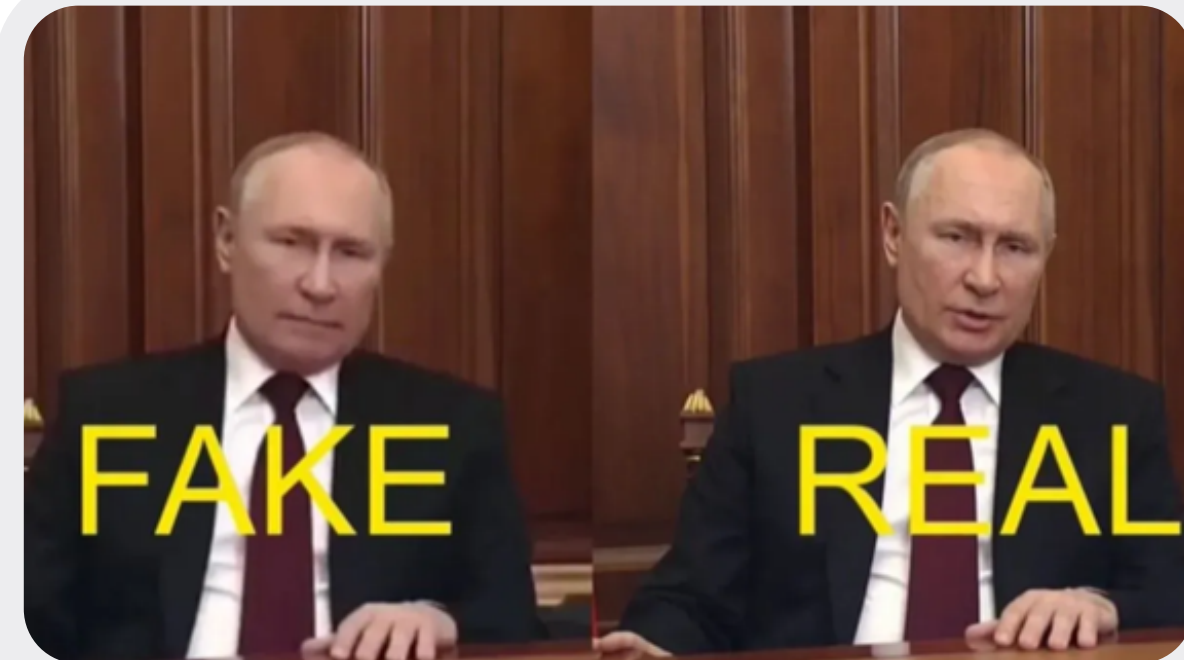
The eyes are often challenging to recreate realistically in deepfakes. Pay close attention to the eyes of the person in the video or image. If they appear unnatural, lack, or have mismatched reflections, it could be a sign of manipulation. In the following example we can observe mismatched eye reflections on a facial image created by DALL-E 2.

- **Contradictory facial expressions:** Spot facial morphing or image stitches if someone's face doesn't seem to exhibit the emotion that should go along with what they're supposedly saying.
- **Observe the overall quality (blurring or misalignment):** Deepfakes may exhibit lower quality compared to original videos or images. Look for anomalies like blurred edges, inconsistent sharpness, or artefacts around the subject's face.



- **Analyse the audio**

AI voice cloning technology has improved significantly in recent years, making it easier to create realistic-sounding voice replicas. These advancements have made it possible for scammers to mimic the voices of individuals with high accuracy, including celebrities and public figures, and increase the likelihood of their victims complying with their requests. Despite the fact that there are several deepfake voice cloning tools such as [Resemble](#), [Fakeyou](#), [Descript](#), [VoiceAI](#), etc, there is not the same progress in online available tools such as [AI Voice Detector](#), that detect voice manipulations and help users to verify audio authenticity. Of course sometimes deepfakes do not synthesise audio convincingly. If the audio appears disconnected or the lip-sync is off, it may be a sign of manipulation. In the following example, during the video where the real Putin is silent, the fake one is speaking.



**Deepfake video with Putin announcing the end of Russia's war with Ukraine ([Source](#))**

• **Look for visual glitches or anomalies**

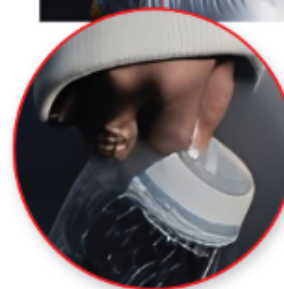


**Bill Hader channels Tom Cruise (Source)**

Deepfakes can introduce visual artefacts or glitches, particularly around the face or edges. Pay attention to areas that seem distorted, have inconsistent lighting, or display unusual colours. The following example shows Bill Hader's face transformed to Tom Cruise's face. If we take a closer look at the video we can observe that when Tom Cruise's face appears or disappears, the lighting in the face differs.

**A closer look at the Balenciaga Pope image**

His **eyelid** appears to merge into his glasses then flow into their own shadow



His **fingers** are closed around thin air rather than the coffee cup he carries



The **crucifix** is held inexplicably aloft with the other half of the chain missing

An [article from the TIME magazine](#) also examined a very characteristic case of an AI generated image depicting Pope Francis allegedly wearing a puffer jacket. Close examination of the image reveals several visual anomalies around his glasses, crucifix and fingers that clearly show this image is not authentic.

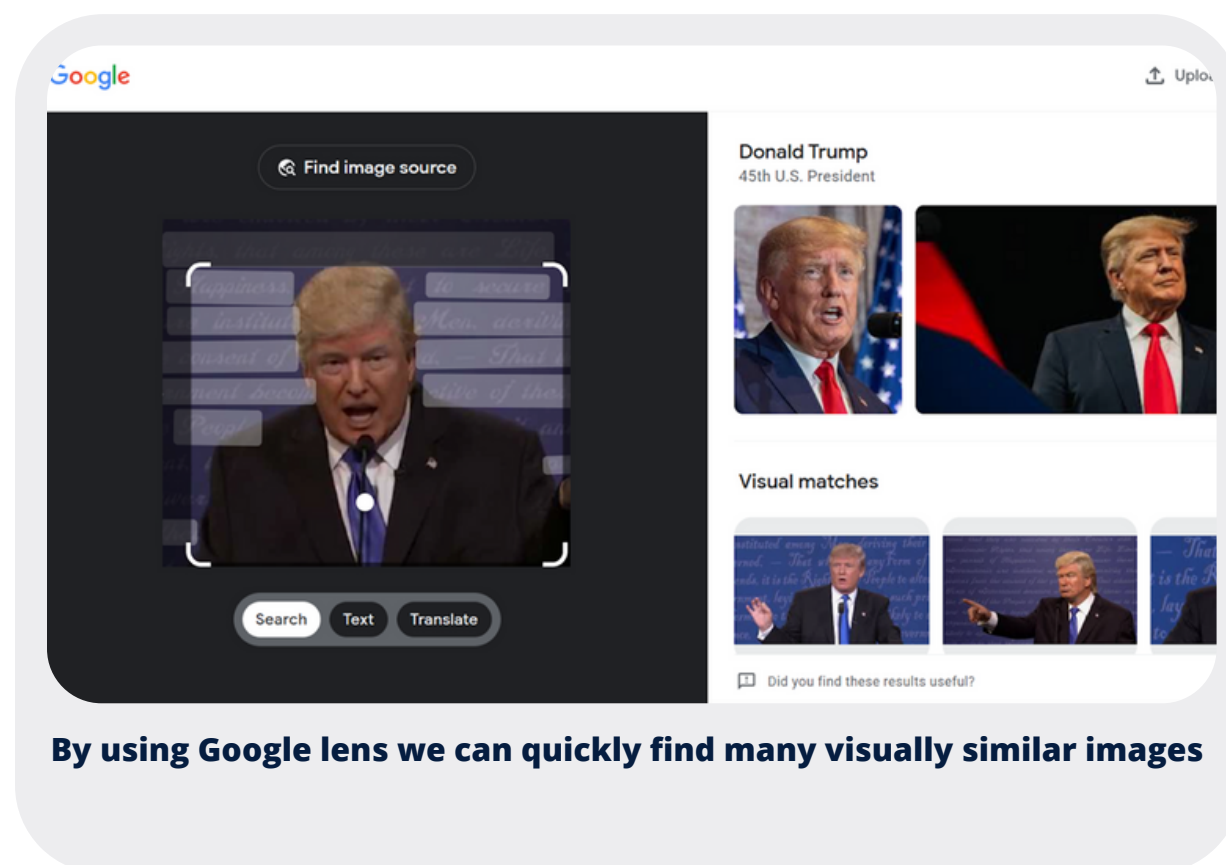
**Spot a deepfake through the facial expression (Source)**

- **Research the source and context:** Investigate the original source of the video or image. Consider the context in which it was shared, and check for any corroborating evidence or multiple angles of the same event.
- **Consult experts:** If you're unsure about the authenticity of a video or image, consult professionals or experts in the field of digital forensics or media analysis. They can provide valuable insights and help determine if manipulation has occurred.

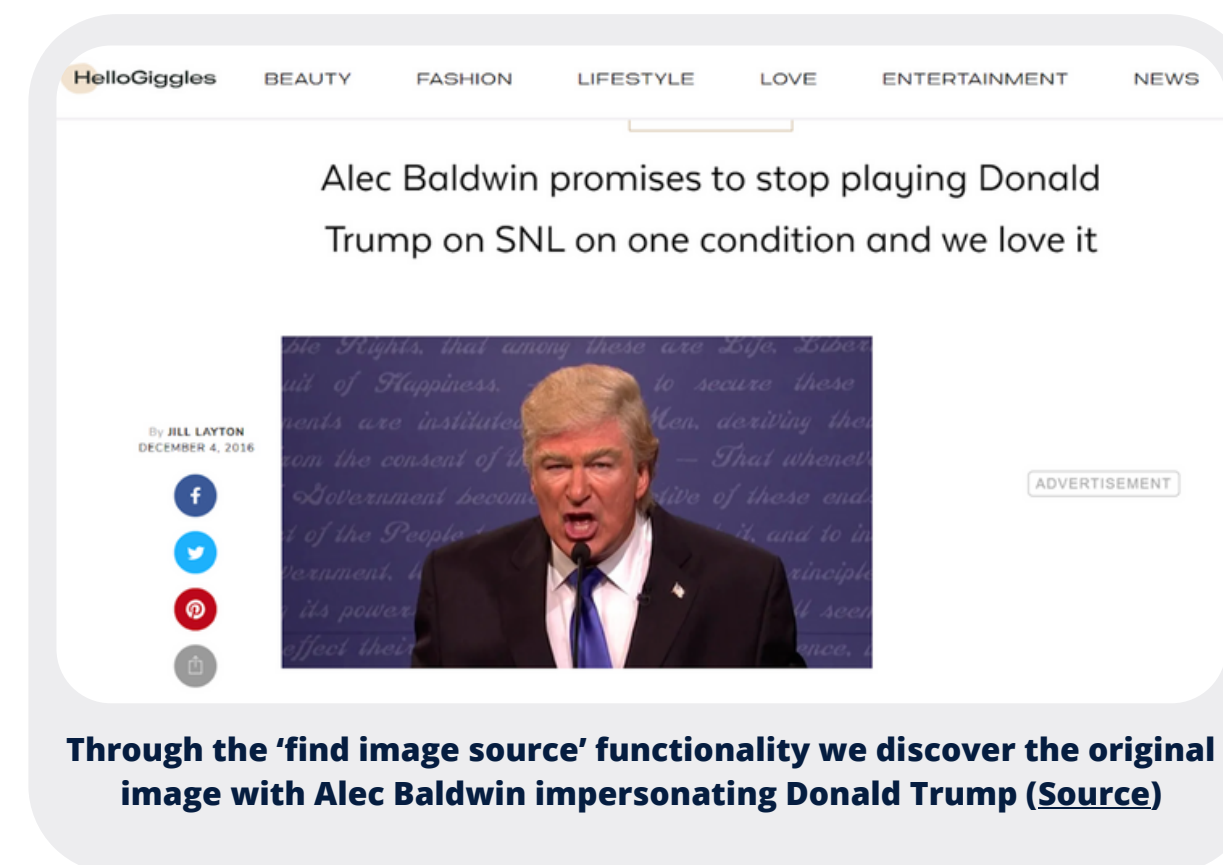
 **Remember, no single method can guarantee 100% accuracy in detecting deepfakes. It's crucial to remain vigilant and combine multiple techniques to increase your chances of spotting potential manipulations.**

## • Reverse image search

Searching for visually similar images on the web can prove highly effective in spotting deepfake images or videos, as one can discover the original image and easily determine if it has been manipulated. There are many free online reverse image search tools such as [Google lens](#), [TinEye](#), [Yandex reverse image search](#), and [Bing visual search](#). In the following example we can see how we can detect a deepfake image by discovering the image source through the use of reverse image search.



By using Google lens we can quickly find many visually similar images



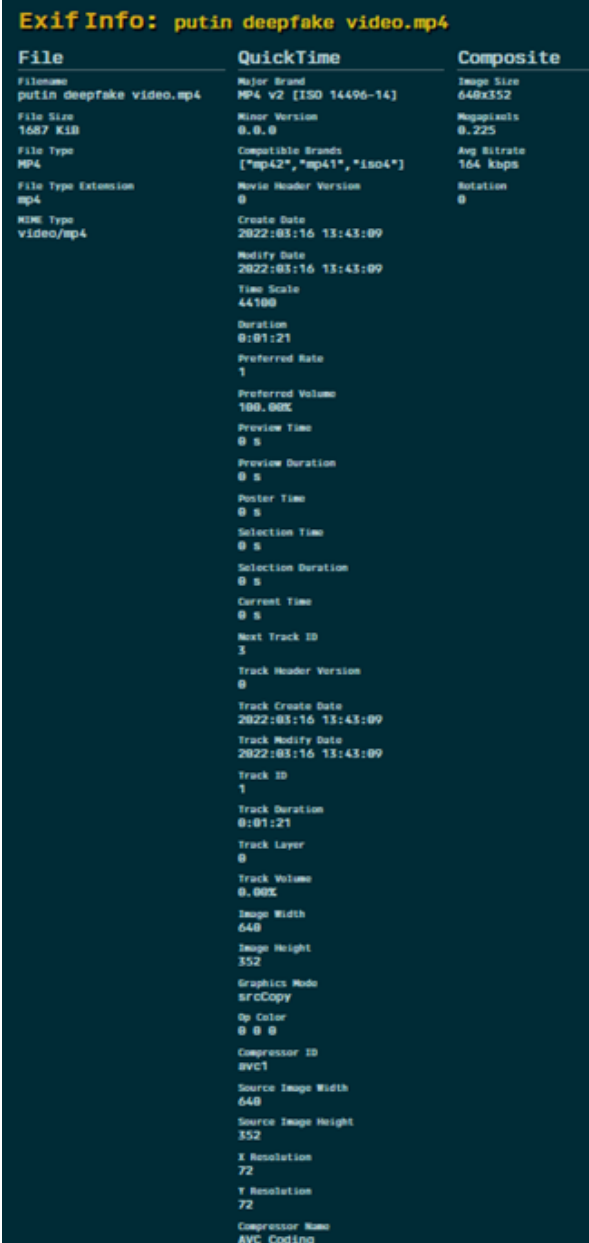
Through the 'find image source' functionality we discover the original image with Alec Baldwin impersonating Donald Trump ([Source](#))

- **Inspect metadata**

Image and video metadata is a valuable resource that can help you verify the authenticity of multimedia content. Metadata contains provenance information about the multimedia file, such as the date it was created, the camera used, and the encoding format. By analysing metadata, you can often discover discrepancies between what the multimedia content claims to show and the details hidden within the file.

For example, metadata may show that an item has been produced, or modified utilising an encoder-decoder algorithm commonly associated with deepfake detection technology. This revelation would strongly suggest that the image or video is not authentic.

To examine the metadata of multimedia content, there are several available online tools and software applications, such as [Exif Info](#), [Jimpl](#), [METADATA2GO](#), and [Brandfolder – Metadata Extractor](#). These tools can extract and present metadata information for analysis. By comparing the metadata of suspicious videos with that of known genuine videos, you can identify inconsistencies and potentially expose deepfakes. In the following example we illustrate the metadata analysis results from the Exif Info tool.



The screenshot shows the output of the Exif Info tool for a file named 'putin deepfake video.mp4'. The output is organized into three columns: File, QuickTime, and Composite. The File column lists basic file information. The QuickTime column lists video-specific metadata such as creation and modification dates, duration, and track information. The Composite column lists image-specific metadata like image size, resolution, and compression details.

File	QuickTime	Composite
Filename putin deepfake video.mp4	Major Brand MP4 v2 [ISO 14496-14]	Image Size 648x352
File Size 1687 KiB	Minor Version 0.0.0	Magazines 0.225
File Type MP4	Compatible Brands ["mp42", "mp41", "iso4"]	Avg Bitrate 164 kbps
File Type Extension mp4	Movie Header Version 0	Rotation 0
MOV Type video/mp4	Create Date 2022:03:16 13:43:09	
	Modify Date 2022:03:16 13:43:09	
	Time Scale 44100	
	Duration 0:01:21	
	Preferred Rate 1	
	Preferred Volume 100.00%	
	Preview Time 0 s	
	Preview Duration 0 s	
	Poster Time 0 s	
	Selection Time 0 s	
	Selection Duration 0 s	
	Current Time 0 s	
	Next Track ID 3	
	Track Header Version 0	
	Track Create Date 2022:03:16 13:43:09	
	Track Modify Date 2022:03:16 13:43:09	
	Track ID 1	
	Track Duration 0:01:21	
	Track Layer 0	
	Track Volume 0.00%	
	Image Width 648	
	Image Height 352	
	Graphics Mode srcCopy	
	Op Color 0 0 0	
	Compressor ID AVC1	
	Source Image Width 648	
	Source Image Height 352	
	X Resolution 72	
	Y Resolution 72	
	Compressor Name AVC Coding	

Example metadata analysis table provided by [Exif Info](#)

- **Detect the use of deepfake technology (1)**

Various tools have been developed to identify deepfakes by detecting the use of the relevant technology. These tools use deep learning to identify manipulated content and, although they are far from being foolproof, they can provide additional assistance in spotting deepfakes. Some of these tools are free online such as [DeepWare AI](#), for deepfake video detection and [Maybe's AI Art Detector](#), [Illuminarty](#), [AI or Not](#), [Hive](#) for deepfake image detection. Others require a registration fee and usually a monthly or annual subscription and include [DuckDuckGoose](#), [Sensity AI](#), and [Reality Defender](#), which cover both image and video. Each tool uses its own method to detect deepfakes, which means there may be limitations on the types of manipulations that can be detected, and great variation in the success rates. Below we illustrate two examples: analysis by all mentioned free tools of an AI-generated image of Pope Francis wearing a Balenciaga puffer jacket generated by Midjourney (a corresponding study has been published in an [article](#) published by The New York Times) and analysis of a deepfake video of Vladimir Putin using [Deepware AI](#).

All tools correctly detect the actual forgery with different level of confidence: Maybe's AI Art Detector: 74%; Illuminarty: 53,4%; Hive: 100% (also noting that it is generated by Midjourney 100%); AI or Not: Notification that the image is AI-generated without a confidence level.



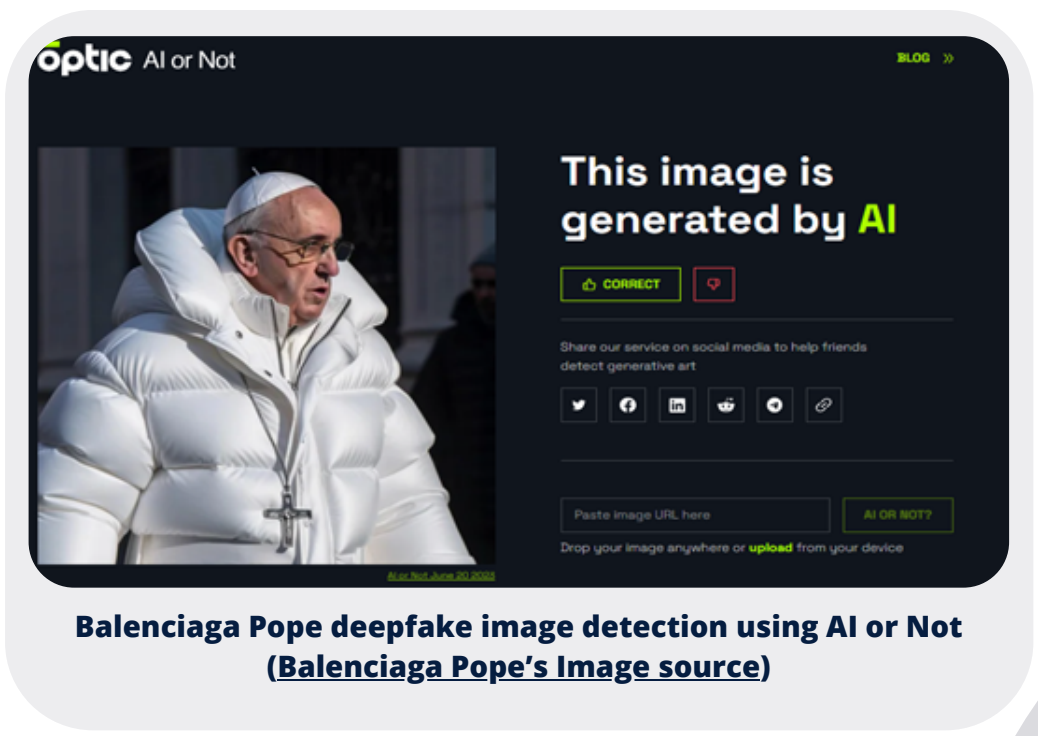
**Balenciaga Pope deepfake image detection using Maybe's AI Art Detector (Balenciaga Pope's Image source)**



**Balenciaga Pope deepfake image detection using Illuminarty (Balenciaga Pope's Image source)**



**Balenciaga Pope deepfake image detection using Hive (Balenciaga Pope's Image source)**



**Balenciaga Pope deepfake image detection using AI or Not (Balenciaga Pope's Image source)**

- **Detect the use of deepfake technology (2)**

Another Deepfake detection tool operating in a similar way is the Deepfake Detector developed by the MeVer team of CERTH, coordinator of MedDMO project, and operates on both image and video. This tool can detect manipulations that are generated by many state-of-the-art deepfake generation approaches and limited access is available upon request by email at [papadop@iti.gr](mailto:papadop@iti.gr).

To detect deepfake images, the tool applies a face detection process and calculates a confidence score for each detected face. The final deepfake confidence score is calculated through a weighted averaging method among the separate faces. The tool applies a video segmentation process and calculates separate confidence scores for every face contained in every segment. The final deepfake score is calculated through a weighted averaging method among the separate faces, within the separate shots. In the following example, we illustrate the detection results on the same deepfake video of Vladimir Putin.



**DEEPPAKE DETECTED** New Scan

Name: <https://twitter.com/izstanur150429091899499...> User: 2022-03-16 16:35:56 UTC  
Size: 1.6 MB Source: 1 year(s) ago

**DETAILS**

Deepware aims to give an opinion about the scanned video and is not responsible for the result. As Deepware Scanner is still in beta, the results should not be treated as an absolute truth or evidence.

**Model Results**

Model	Result	Video	Audio
Analyst	DEEPPAKE DETECTED	Duration: 81 sec	Duration: 81 sec
Avatarify	NO DEEPPAKE DETECTED(0%)	Resolution: 640 x 352	Channel: mono
Deepware	NO DEEPPAKE DETECTED(28%)	Frame Rate: 30 fps	Sample Rate: 44 kHz
Seferbekov	SUSPICIOUS(71%)	Codec: H264	Codec: AAC
Ensemble	SUSPICIOUS(55%)		

[Request Expert Review](#) [Request TakeDown](#)

Deepfake video detection using the Deepware AI detection tool (Video source)

**Deepfake Analysis**  
94% Manipulation probability

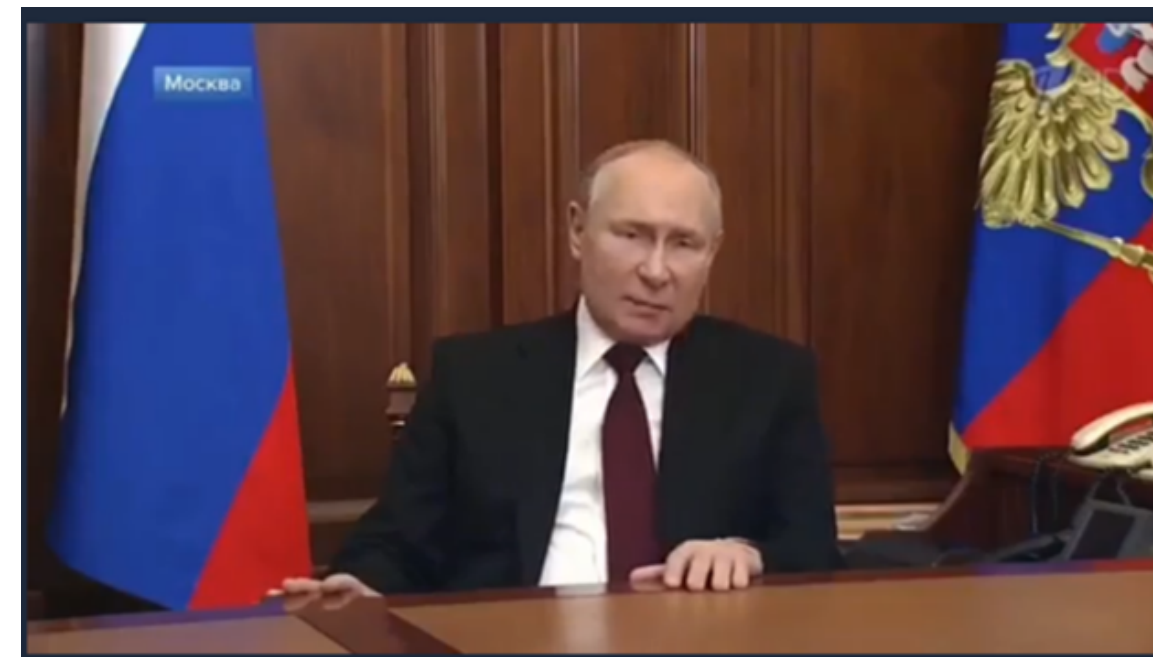
**Global Level Analysis - 94%**  
Manipulation Probability: 94%  
Inference Time: 22s  
Segments: 3  
Frames Analyzed: 70

**Segment Level Analysis**

01:02:11

**Segment 2 - 95%**  
Manipulation Probability: 95%  
Start Time: 0:57  
End Time: 1:00  
Start Face: 1  
Frames Analyzed: 24

**Face 1 - 92%**



Deepfake video detection results using the MeVer deepfake detection tool (Video source)

A hand is visible on the left side of the image, pointing towards the right. The background is a light blue and white digital space with a network diagram consisting of nodes and connecting lines. The text is overlaid on a dark blue rectangular box.

## **Next-generation tools:**

Beyond existing methods, services and tools, the research community is quite active experimenting with several methods for deepfake detection. Among the primary concerns for existing methods and tools is their lack of reliability especially in cases where a deepfake item is generated by a method that is completely unknown to the detection tool. Another issue that existing tools face is their high false positive rate, i.e. their tendency to flag authentic videos as deepfake, e.g. in cases of low quality or highly compressed videos. To address these limitations, a few approaches of growing interest include the following:

# Tools



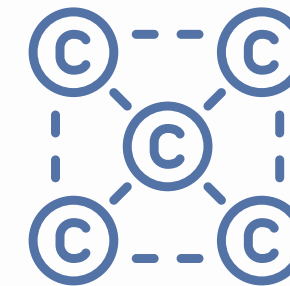
## deepfake detection using multimodal cues:

there are several methods that are trying to detect deepfakes by combining both the visual and audio signals exploiting the intrinsic synchronization and consistency between the visual and auditory modalities (e.g. Zhou and Lim, 2021[1] and Chou et al, 2020[2]).



## real forensics:

instead of training deepfake detection methods on deepfakes, these methods are trying to train on videos of real talking faces, taking advantage of the large numbers of examples and the rich information on natural facial appearance (e.g. Haliassos et al, 2022[1], Cozzolino et al, 2023[2]).



## identity watermarking:

these methods experiment with invisible watermarks that are embedded in the facial areas of authentic videos and are distorted by facial manipulations. Detection of this distortion can then be utilised for identification of manipulations (e.g. Zhao et al, 2023[1] and Huand et al, 2022[2]).



## deepfake detection model transparency:

as every AI-based method has its merits and shortcomings Mitchell et al[1] proposed a standardised way of disclosing machine learning models' performance characteristics and their limitations. Such model cards have been proposed explicitly for deepfake detection by Baxevanakis et al[2].

**As such methods are still experimental, there are still no publicly available tools for testing, but continuous developments are expected and new tools will surely emerge in the near future.**

- <sup>1</sup>Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14800-14809)
- <sup>2</sup>Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020, October). Not made for each other-audio-visual dissonance-based deepfake detection and localization. In Proceedings of the 28th ACM international conference on multimedia (pp. 439-447).
- <sup>3</sup>Haliassos, A., Mira, R., Petridis, S., & Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14950-14962)
- <sup>4</sup>Cozzolino, D., Pianese, A., Nießner, M., & Verdoliva, L. (2023). Audio-visual person-of-interest deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 943-952).
- <sup>5</sup>Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive deepfake defence via identity watermarking. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 4602-4611)
- <sup>6</sup>Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., ... & Ma, K. K. (2022, June). Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 1, pp. 989-997)
- <sup>7</sup>Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).
- <sup>8</sup>Baxevanakis, S., Kordopatis-Zilos, G., Galopoulos, P., Apostolidis, L., Levacher, K., Baris Schlicht, I., ... & Papadopoulos, S. (2022, June). The mever deepfake detection service: Lessons learnt from developing and deploying in the wild. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation (pp. 59-68).



MedDMO is a Digital Europe SME Support Action project co-financed by the EC under Grant Agreement with project ID: 101083756. The content of this document is © the author(s).  
For further information, visit [www.meddmoeu](http://www.meddmoeu)