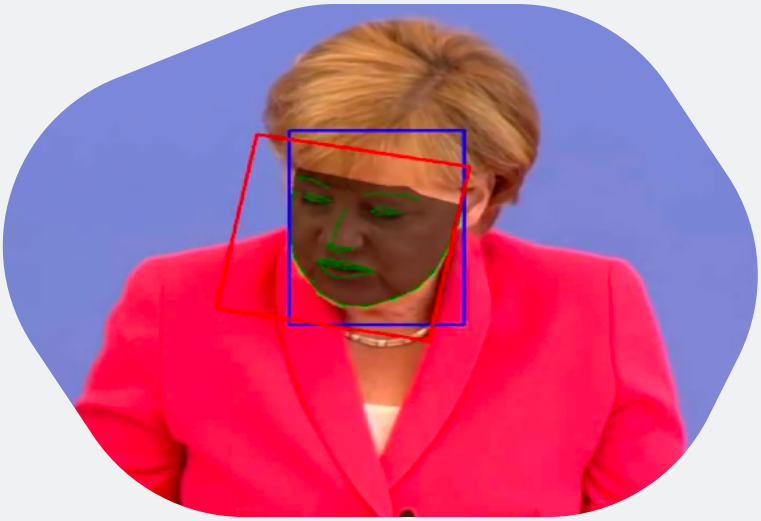
SENTINEL



DEEPFAKES 2020: THE TIPPING POINT

The Current Threat Landscape, its Impact on the U.S 2020 Elections, and the Coming of AI-Generated Events at Scale.

About Sentinel.

Sentinel works with governments, international media outlets and defense agencies to help protect democracies from disinformation campaigns, synthetic media and information operations by developing a state-of-the-art AI detection platform.

Headquartered in Tallinn, Estonia, the company was founded by ex-NATO AI and cybersecurity experts, and is backed by world-class investors including Jaan Tallinn (Co-Founder of Skype & early investor in DeepMind) and Taavet Hinrikus (Co-Founder of TransferWise).

Our vision is to become the trust layer for the Internet by verifying the entire critical information supply chain and safeguard 1 billion people from information warfare.

Acknowledgements

We would like to thank our investors, partners, and advisors who have helped us throughout our journey and share our vision to build a trust layer for the internet.

Special thanks to Mikk Vainik of Republic of Estonia's Ministry of Economic Affairs and Communications, Elis Tootsman of Accelerate Estonia, and Dr. Adrian Venables of TalTech for your feedback and support as well as to Jaan Tallinn, Taavet Hinrikus, Ragnar Sass, United Angels VC, Martin Henk, and everyone else who has made this report possible.

Johannes Tammekänd

CEO & Co-Founder

Contact: info@thesentinel.ai

Authors: Johannes Tammekänd, John Thomas, and Kristjan Peterson

Cite: Deepfakes 2020: The Tipping Point, Johannes Tammekänd, John Thomas, and Kristjan Peterson, October 2020

Executive Summary.

"There are but two powers in the world, the sword and the mind. In the long run the sword is always beaten by the mind." - Napoleon Bonaparte. Psychological warfare has always been an effective asymmetric tool of warfare to deny the adversary or influence a target audience. In the last century, we saw how Adolf Hitler and Joseph Goebbels, Reich Minister of Propaganda, effectively used radio as the means of influencing their population towards a conflict that led to the Second World War. In the 1960s, the Soviet Union understood that it could not compete equally with the U.S. and the West as its economy was smaller. To counter this, the Directorate D of KGB developed asymmetric information warfare measures comprising of ideological subversion and active measures programs to subvert the U.S. and the West. It is estimated by some accounts that the KGB allocated up to 85% of its resources to these programs with only 15% into the areas of intelligence and espionage.

Until recently, Information Operations was a discipline of warfare conducted by human specialists crafting manipulative content for dissemination across a range of information channels. However, with the advancement of AI we are now at the tipping point of AI-generated information warfare campaigns that can be carried out at near zero-cost, hyper-personalized, and at scale. Deepfakes are the primary force behind this next generation of information operations through use of AI-generated synthetic video, audio, images, and text.

Since the inception of deepfakes in 2017, we have witnessed an exponential growth in them similar to that seen in the early days of malware in the 1990s. Since 2019, the number of deepfakes online has grown from 14,678 to 145,227, a staggering growth of ~900% YOY. The majority of these are on popular social media platforms and between them have amassed close to 6B views. This exponential growth has been driven by advancements in AI algorithms, lower cost of compute, and exponential growth of data on the Internet. This has led to deepfakes that can be created cheaply, at low barrier and at scale. Simple productized solutions are now available with intuitive user interfaces and can create deepfakes using only a single image.

The malicious use of such technology has Forrester estimating that deepfake fraud will reach $250M already in 2020. There have already been recorded incidents of deepfakes utilized across multiple domains to deceive its victims.

Executive Summary.

Examples include a deepfake audio used to manipulate a CEO into wiring a false payment, the use of ransomfakes targeting women, and a military coup d'état initiated in Gabon as a result of an alleged deepfake released by the president. As a result of this, governments and organizations have now recognized the threat that deepfakes are posing and have started to introduce measures intended to curb its rise and spread.

2020 is the tipping point where deepfakes are beginning to be used at scale to disrupt economies, influence political events, undermine journalism and wider society. As for 2016, we predict that the 2020 U.S. elections will be the target for influence operations and we assign a high likelihood (80-90%) that for the first-time next generation AI-based information operations will be deployed.

During the 2020 U.S. elections, deepfakes will most likely be used against the democratic process by suppressing votes. Furthermore, it will enable individuals and groups to sow doubt on any claims made against them by claiming that deepfakes have been used — a technique known as the liar's dividend.

As the exponential trend continues, the majority of the world's digital information will be eventually produced by AI. This in turn necessitates a trust layer for the Internet where people can be sure the information they receive is coming from a real person and not malicious.

Sentinel is executing towards this vision for a trust layer together in partnership with the Estonian Government, which is a world leader in e-governance, digital identity, and cybersecurity.

Governments, civil society and organizations should invest immediately into education to prevent society's susceptibility to deepfakes, fund technological countermeasures and draft required legislation. The societies who will adapt the fastest to these changes will dominate the 21st century information environment. The ones who do not will find themselves struggling to catch up as the exponentially evolving information environment does not stop.

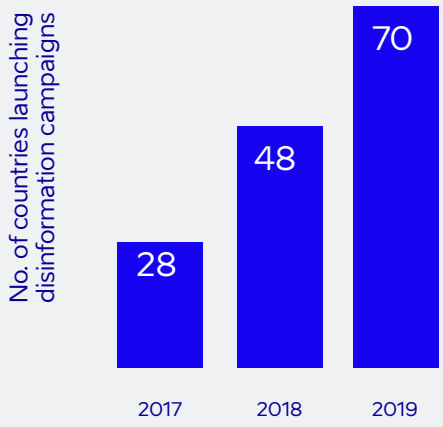
J. Tammekänd

Johannes Tammekänd

CEO & Co-Founder

Table of Contents

Disinformation Campaigns are Growing Quickly and Becoming More Potent.



+150%

Increase in countries launching disinformation campaigns (2017 - 2019)[1].

$78B

Cost of disinformation annually (2019) [2].

False stories are

6x

faster in reaching an audience of 1,500 than true stories[3].

8M

Estimated Twitter accounts a week that spread disinformation[4].

Source: [1] University of Oxford 2019; [2] University of Baltimore, 2019; [3] The spread of true and false news online – MIT, 2018; [4] Washington Post, 2019

The Exponential Growth of Deepfakes: ~6820X since 2019.

100M+

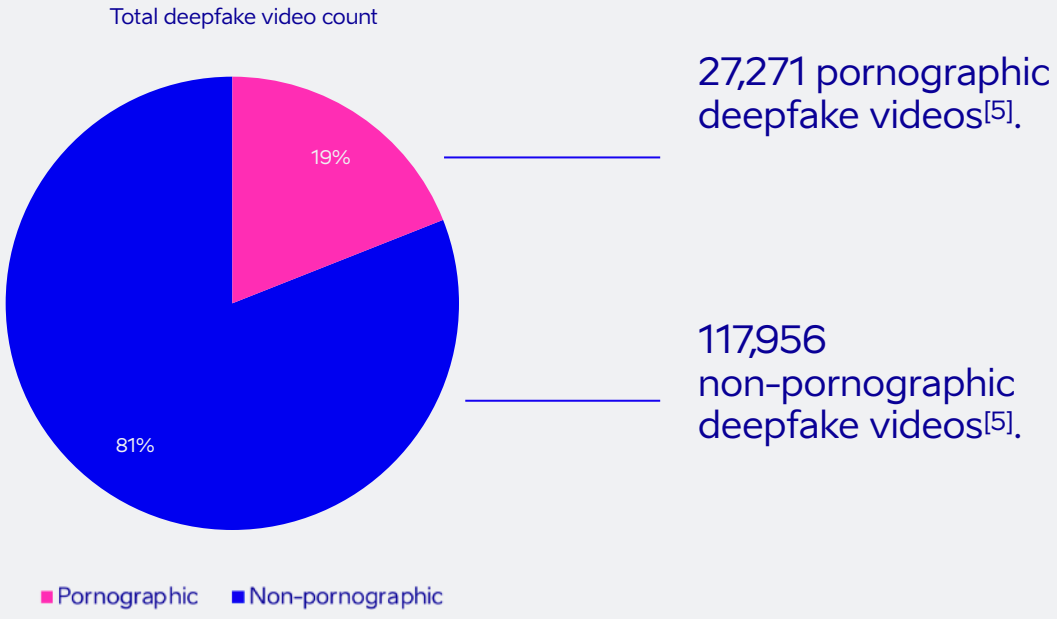
total deepfake videos online (2020)[7]

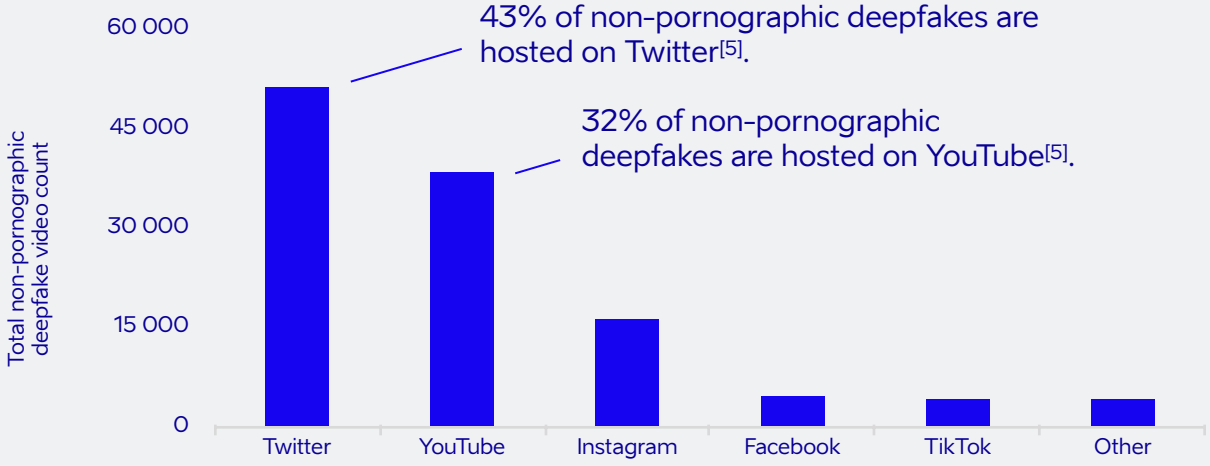
~6820X

YOY growth since 2019[5].

Source: [5] Sentinel Analysis, 2020 [6] DeepTrace, 2019, [97] Business Insider 2020

The Majority of Deepfakes Online are non-Pornographic and Hosted on Twitter and YouTube.





Source: [5] Sentinel Analysis, 2020

Since 2017, Deepfakes have Amassed 5.4 Billion Views on YouTube Alone.

Total views of deepfake videos on YouTube[7].

Total views of deepfake videos on TikTok[7].

Forrester Estimate of Fraud Costs from Deepfakes in 2020[8].

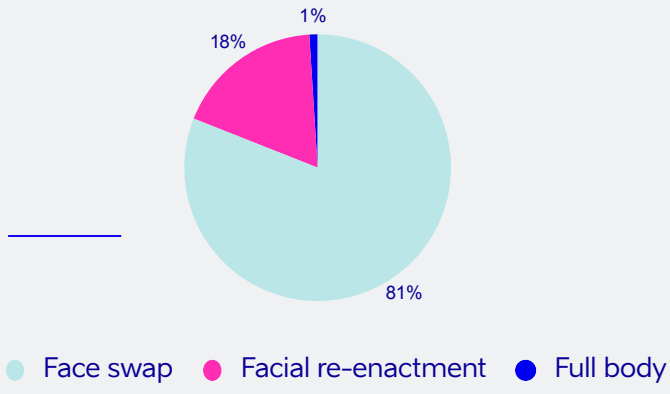
Source: [7] Sentinel Analysis, 2020; [8] Forrester, 2019;

The Majority of Non-Pornographic Deepfakes are Face Swaps and Created to Entertain Viewers.

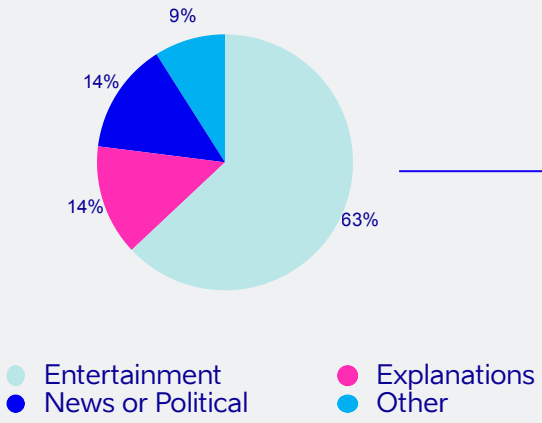
7000

We have the largest verified database of deepfakes which have been independently verified using our technology and labeled using a team of verification specialists. In this public report, we have sampled 7000 videos from our database for the analysis[9].

81% of these deepfakes are Face swaps[9].



*% of deepfake video counts by deepfake type

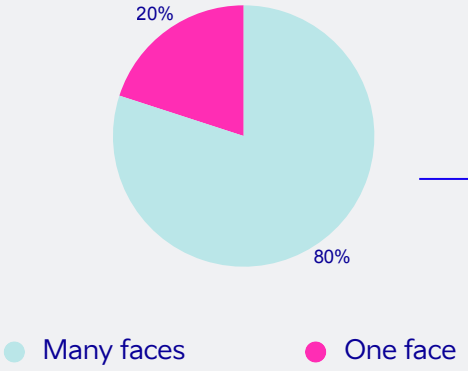


63% of these deepfakes were made for entertainment[10].

*% of deepfake video counts by category

Source: [9, 10] Sentinel Analysis, 2020
Note: Based on analysis of N=7000 deepfake videos from Sentinel's database. Videos span time range of Nov. 2017 to 2020.

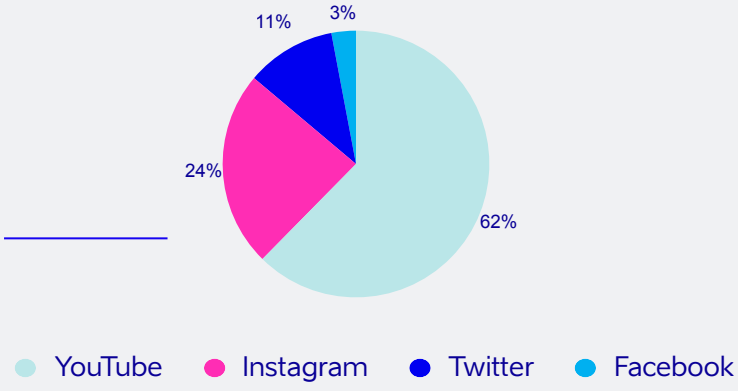
Most Deepfakes Involve More than One Subject in the Video.



80% of these deepfakes had more than one unique subject in the video[11].

*% of deepfake video counts by face count

63% of these deepfakes were found on YouTube[12].



*% of deepfake video counts by deepfake source

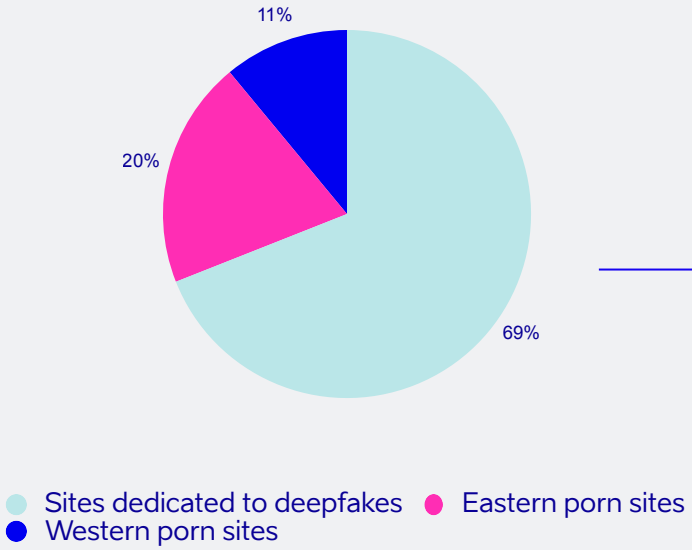
Source: [11, 12] Sentinel Analysis, 2020
Note: Based on analysis of N=7000 deepfake videos from Sentinel's database. Videos span time range of Nov. 2017 to 2020.

27,271 Pornographic Deepfakes Online.

Our analysis of over 30 porn websites revealed a vast trove of pornographic deepfakes, with the majority dedicated to celebrities or people of influence.

27,271

Total count of pornographic deepfakes[13].



69% of pornographic deepfakes were hosted on dedicated deepfake porn sites[13].

*% of pornographic deepfake video counts by source category

Source:[13] Sentinel Analysis, 2020

In Addition to Dedicated Deepfake Pornographic Sites, there are Underground Communities of Deepfake Porn Creators.

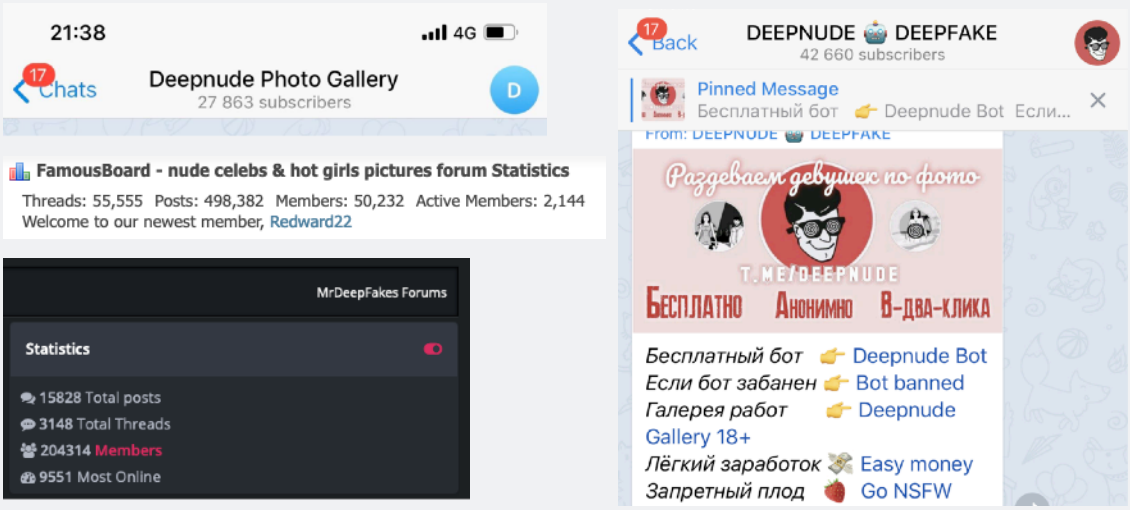


Figure 1. Example screenshots of Forums and Telegram Chat groups where pornographic deepfakes can be purchased or shared, and requests to have them made can be put through.

These communities total over 100,000 members and engage in the development of pornographic deepfakes.

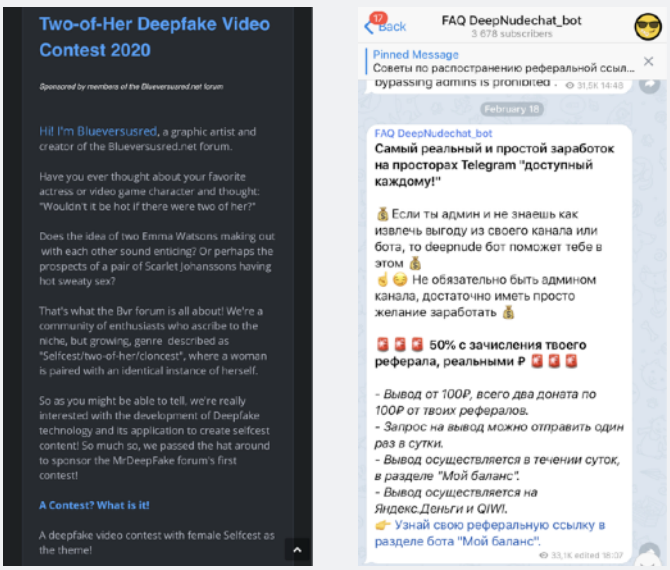


Figure 2. Above you'll see a forum running a deepfake pornographic contest (LHS image) and a screenshot for a pornographic deepfake Telegram chat group advertising a referral program with a monetary reward to grow their community base (RHS image).

Encrypted Telegram chat groups also engage in commerce. Anyone can pay to have custom pornographic deepfakes developed, and one group even had a monetized referral program to grow their community. Moreover, forums actively engage in prized contests to determine who can create the best pornographic deepfake[14].

Source:[14] Sentinel Analysis, 2020

Blackmail and Ransomfake Attacks Against Women using Pornographic Deepfakes are on the Rise.

One can draw parallels to the early days of darkweb malware markets where sites dedicated to creating pornographic deepfakes are emerging. Long-term, this can result in pornographic deepfakes created of nearly every person who has their face pictures on the web.

Figure 3. Journalist Rana Ayyub, a victim of a deepfake pornographic video using Twitter to fight back against the virality of the deepfake video.

Figure 4. Noelle Martin, a victim of photoshopped pornographic images of herself is now an advocate for fighting deepfake pornographic videos and broader image-based abuse.

Women are the most vulnerable as studies have shown pornographic deepfakes have lasting negative effects on its victims. Victims can experience trauma despite knowing that the videos are fake and in cases such as that of Rana Ayyub and Noelle Martin, it affected their job prospects and personal safety[15, 16, 17].

We are already seeing automated bots that crawl the web for images of women from their social accounts, create deepfake porn, and then carry out ransomfake attacks towards them.

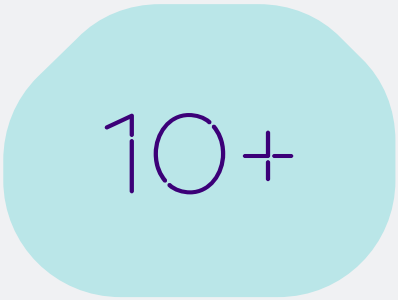
Such blackmail and ransomfake attacks are on the rise with perpetrators asking for payment in cryptocurrencies with the threat to release their victim's pornographic deepfakes to the public.

Source:[15] Loyola Law School, 2019; [16] India Today, 2018 [17] ABC Australia, 2018

Consumers, Industry, and Government are Seeking Solutions Against the Threat of Deepfakes.

Of Americans believe deepfakes could cause more harm than good[18].

Believe fake news impacts their confidence in government institutions[19].



Initiatives and Acts have been issued by the U.S government to tackle deepfakes in 2019[20].

Deepfake detection challenge launched by Facebook, Amazon, and Microsoft[21].

Source: [18] WholsHosting Dodging Deception & Seeking Truth Online Study, 2019; [19] Pew Research Center Survey, 2019; [20] Sentinel Analysis, 2020; [21] Facebook, 2019

What are the Different Flavours of Deepfakes?

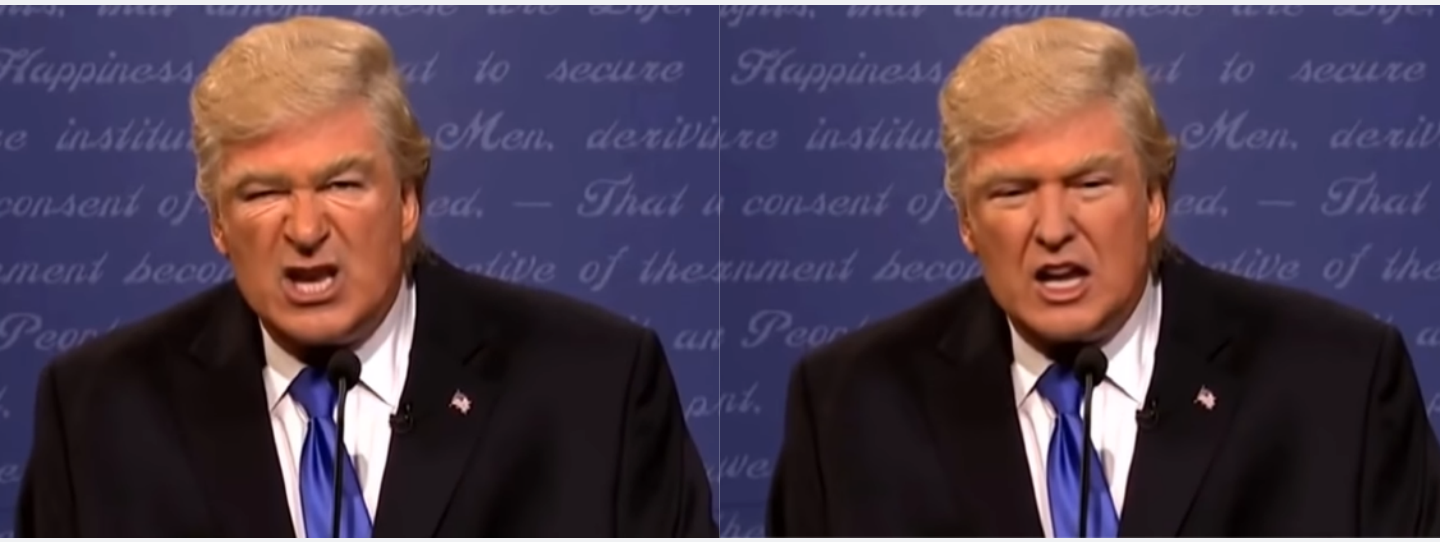


Figure 5: Face swap deepfake using actor Alec Baldwin (original on left) and deepfake on right of President Donald Trump

Deepfake(s) is a term used to define when artificial intelligence techniques are applied to alter digital media, mainly video, audio or images.

Its popular application involves superimposing a source individual's face onto the target individual. This allows the source individual to carry out actions that the subject of the video will appear to do.

Through the use of deep learning algorithms, advanced deepfakes are already indistinguishable to the naked eye and ear. This means only through the use of detection technology can advanced forms of manipulated media be detected and labeled as inauthentic[22].

There are five types of deepfakes today:

1. Facial re-enactment
2. Face swapping
3. Full-body deepfake
4. Audio deepfake
5. Image-based deepfake

[22] Source: FaceForensics++: Learning to Detect Manipulated Facial Images 2016

Facial Re-Enactment Deepfakes.

Facial re-enactment involves the manipulation of facial expressions of a target subject in a video based on input from a source actor. In the figure below, a video of President Bush has his facial expressions from a previously recorded interview altered using an actor's expression[23].

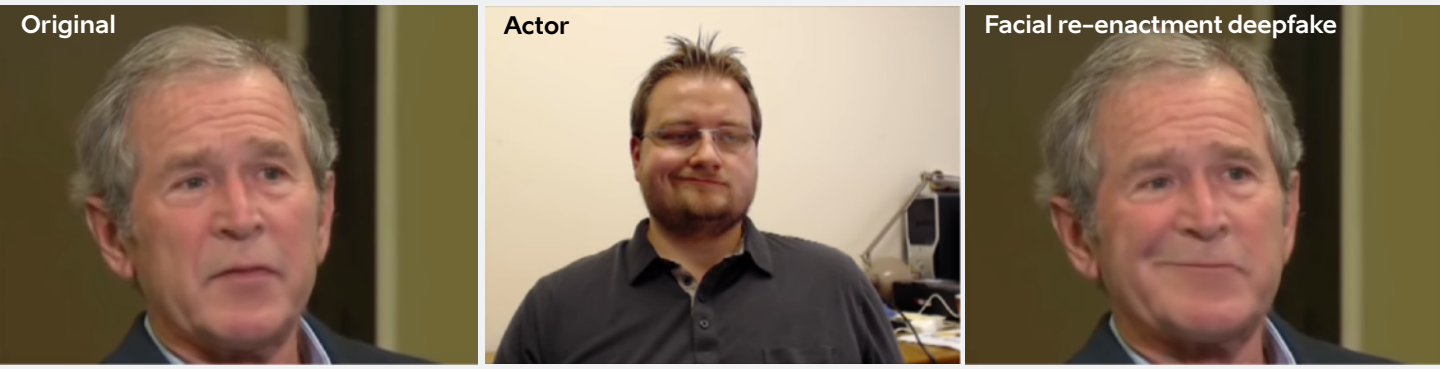


Figure 6: Facial re-enactment deepfake applied to former U.S president George W Bush to mimic the lower mouth expression by an actor.

1 The target's facial features are learned by the AI model

2 The actor provides input of facial expressions

3 The actor's input is now used to generate synthetic movements on the target's face

Figure 7: Neural Voice Puppetry: Audio-driven Facial Reenactment enables applications like facial animation for digital assistants or audio-driven facial reenactment.

Neural Voice Puppetry is a state-of-the-art technique where given an audio of a source person or digital assistant, one can generate a photo-realistic output video of a target person. Their approach generalizes across different people, allowing to synthesize videos of a target actor with the voice of any unknown source actor or synthetic voices that can be generated with text-to-speech[24].

[23] Face2Face: Real-time Face Capture and Reenactment of RGB Videos Stanford University 2018; [24] Neural Voice Puppetry, Technical University of Munich, 2019

Face Swapping Deepfakes.

Face swapping involves having a source subject's face swapped into the target subject face. Through deepfake technology, the subject in a target video can fully act as the source subject.

Figure 8: Face swapping video using politician Elizabeth Warren whose face superimposed onto actress Kate McKinnon

1 The source facial features are learned by the AI model

2 The target's facial features are learned by the AI model

3 The source facial features are superimposed onto the target's and manipulated

The most popular use case example of this is having the face of one individual (the original subject), stitched realistically onto a target individual as seen in the figure above. Development of this type of deepfake requires only a limited amount of training data, consumer level technology, and an actor to help provide the target subject with new actions and speech.

This means a source actor in a video can be made to say and do things as the target subject, with actions that are entirely synthetic and hyper-realistic[25].

[25] Source: Protecting World Leaders Against Deep Fakes, University of California Berkeley, 2019

Full Body Deepfakes.

Full body deepfakes incorporate the techniques of face swapping and facial reenactment but extend AI technology to now synthetically generate the entire body to carry out actions.

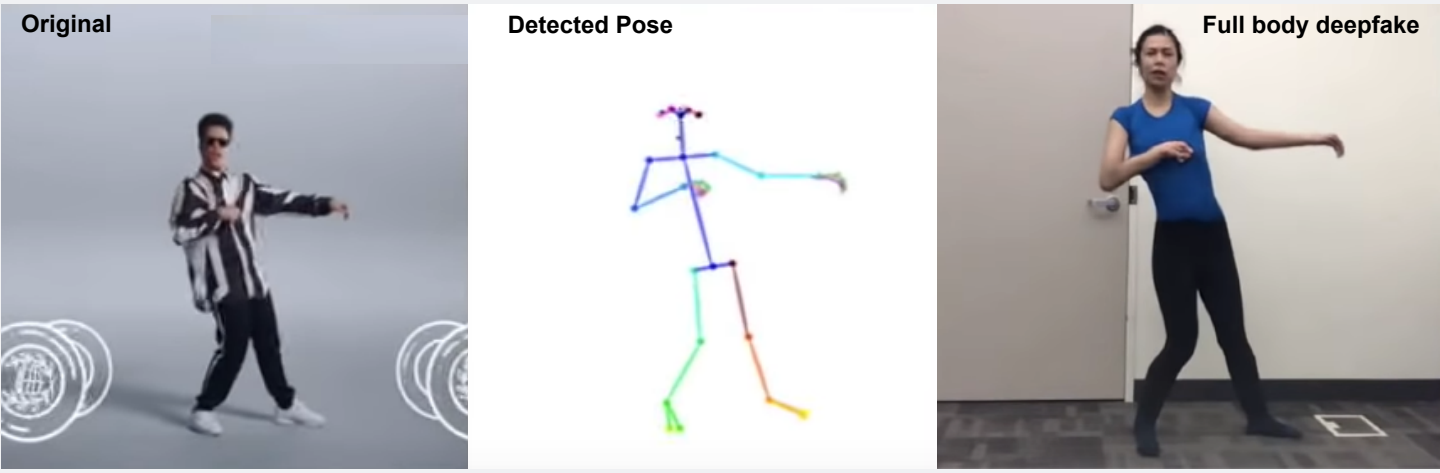


Figure 9: Full body deepfake using publicly available music video from artist Bruno Mars (left), to determine motion (middle), and regenerate body of target (right)

1. The source and target's body features are learned by the AI model
2. The body elements of the target and source are mapped.
3. The target is regenerated by AI with the body elements manipulated

Full body deepfakes use AI to learn the elements that constitute various poses performed by a body. A sequence of these poses form movement which is then applied to the targets synthetically generated body. This means it is possible for people to be synthetically generated in entirety to perform actions beyond just manipulation to the face. Full body deepfakes are harder to generate than facial deepfakes because the underlying technology is not yet as advanced[26].

[26] Source: Everybody Dance Now: Motion Retargeting Video Subjects, University of California Berkeley, 2019

Audio Deepfakes.

Audio deepfakes are when AI technology is used to synthetically generate speech that mimics a target.

This involves AI enabled voice synthesis in order to generate new audio that is identical to the subject's original voice samples (e.g. press interviews, speeches, etc.)

Companies such as Descript, Resemble.ai and Baidu have created products that can clone and synthesize voices.

Figure 10. Example companies creating audio deepfakes

Descript has developed a free productized service that allows anyone to manipulate existing audio by way of editing a simple transcript to create a new version. Baidu's Deep Voice 3 service can clone a voice based on an audio sample in just three seconds[27].

$243,000 stolen

An audio deepfake of a UK energy company's CEO was used by fraudsters to mimic a wire transfer order via phone call to a Hungarian supplier. After the wire transfer was completed, the supplier disappeared and was untraceable[28].

3x

Incidents reported by cybersecurity firm Symantec in 2019 involving audio deepfake based fraud resulting in excess of million dollar loses[29].

This means audio deepfakes can result in phone conversation or voice message being synthetically generated and indiscernible to the human ear on whether it's AI or a real human[30].

Source: [27] Vice, 2018; [28] The Verge, 2019; [29] Wall Street Journal, 2019; [30] Venturebeat, 2019

Deepfakes Based on Synthetic Humans.

Utilizing generative adversarial networks, fully synthetic human faces can be generated. This adds a layer of protection for deceivers because each image is unique and it can not be traced back to a source with a reverse image search.

Synthetic generation of images

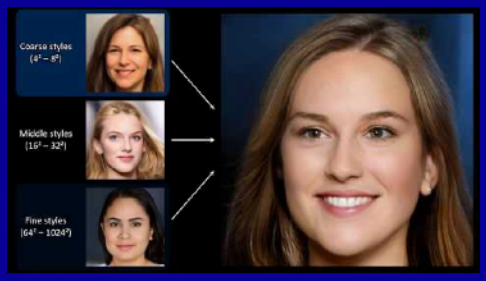


Figure 11. GAN generated synthetic human faces.

Nvidia released latest R&D on a style-based generator based on adversarial networks. This generator would, with limited training data, synthetically generate new human faces[31].

GAN generated synthetic humans and deepfake techniques can be combined to generate deepfakes based on synthetic humans.

Generation of deepfake videos using single image

Figure 12. GAN synthetic deepfake video from still image.

Research from the Imperial College of London was able to demonstrate the synthetic generation of video based on a single image. When paired with deepfake audio, this can be used to create video deepfakes of historical figures where the availability of video content with them in it is limited[32].

This means it is possible to artificially generate images and videos of humans that don't exist. Long-term this will lead to the majority of humans on the Internet being synthetic.

Source: [31] Analyzing and Improving Image Quality of StyleGAN, Nvidia, 2019 [32] Few-Shot Adversarial Training of Realistic Neural Talking Head Models, Samsung AI Center, Moscow, 2019;

Deepfake News: Natural Language Generation.

Advancements in AI and deep learning have also brought about synthetic generation of text. Natural language generation can now be used to create targeted propaganda that can mimic the style of real news.

Natural language AI models such as GPT2, Grover, etc. can generate entire fake news articles. Today, the majority of fake news is written by humans, but early indicators of the success of models such as Grover point to a near-future where disinformation can be generated at near zero cost and at an exponential speed by state sponsored adversaries, criminal groups, etc.

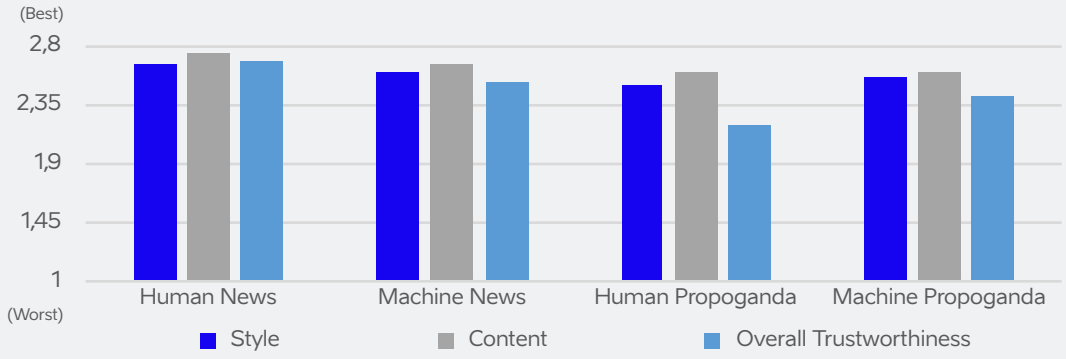


Figure 13. Human evaluation of AI vs. human generated news and propaganda. For each article, three annotators evaluated style, content, and overall trustworthiness (N=100)

The figure above shows the results of human evaluation of AI generated news versus those created by real humans. Results showed that those generated by machines (AI) were rated as more trustworthy and plausible than the original human-written propaganda[33].

In addition to all the visual and audio deepfakes, AI generated textual propaganda can be developed at scale while being highly personalized and inexpensive to produce. It can then be distributed widely through platforms such as social media to conduct information operations.

Source: [33] Defending against neural fake news, University of Washington, 2019

Deepfakes in the Wild

Deepfakes in the Wild.

Figure 14. Deepfake Donald Trump created by Belgian Political Party

A Belgian political party released this deepfake of President Trump asking Belgium to exit the Paris climate agreement. The political party released it assuming the synthetic nature was clear, but many believed it was an authentic video[34].

~20K Views in 48 hours

After a long absence from public appearances, to quell the rumors of ill health or death, a video was released by the Gabon government of President Ali Bongo Ondimba. This video was suspected of being a deepfake due to the odd complexion and eye lid movements of the president. Later, it was found out he had a stroke and the video was real[35].

Figure 15. Alleged deepfake of the president of Gabon that sparked an attempted coup.

Video helped start an attempted coup by the military.

Figure 16. Indian journalist Rana Ayyub, victim of deepfake pornography

Indian Journalist Rana Ayyub was the subject of a deepfake pornographic video that was released after her controversial statements against a gang rape incident in India. The video went viral on Whatsapp and led to threats of hate crime directed at her[36].

~40K shares in 24 hours

Source: [34] Politico 2019; [35] Mother Jones, 2019; [36] India Today 2018;

Deepfakes in the Wild.

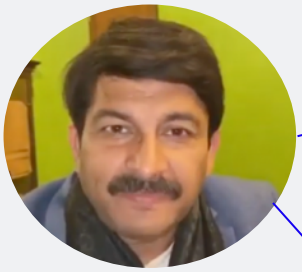


Figure 17. Politician Manoj Tiwari deepfake

A deepfake of Indian politician Manoj Tiwari went viral ahead of legislative elections in Delhi. The deepfake was used to have the politician speak in a variety of languages to resonate with voters better in India[37].

Reached 15M people and 5,800 Whatsapp groups

Figure 18. Video of Uighur poet Heyit where authenticity couldn't be verified

Chinese state media released a video proving a Uighur poet and musician was still alive after Turkey claimed he died in custody. Media organization Reuters was unable to verify the authenticity of the video with activists pointing to odd body language and unnatural speech patterns indicating digital manipulation[38].

Shared on prominent news outlets such as the Guardian with 1M subscribers

Figure 19. Deepfake of Belgium's prime minister

A deepfake of Belgium's prime minister citing the COVID-19 pandemic as an effect of a deeper environmental crisis was shared by the Extinction Rebellion Belgium group on Facebook[39].

700+ likes and 2K shares

Source: [37] Technology Review 2020; [38] Reuters 2019 [39] The Brussels Times, 2020

Influencing and Damaging Diplomacy through Targeted Attacks.

Figure 48. Tweet by Christophe Guilhou, the French ambassador to Cameroon, sharing the debunked deepfake made of him.

The French ambassador to Cameroon found a damaging deepfake of himself that was broadcast on social media saying "The French Republic is the supervisory power that colonized Cameroon"[94].

Source: [94] @ChrisGuilhou, Twitter, 2020

Adversarial State Actors: Operation Secondary Infektion by Russia.

Figure 24. Fake tweet attributed to U.S. Senator Marco Rubio, accusing the UK of interfering in the U.S. midterm elections of November 2018 through the use of deepfakes.

Russian state actor weaponized the concept of deepfakes already in 2018 with a fake Marco Rubio tweet. This was part of a larger influence operation to sow conflict between Western countries. We expect sophisticated adversaries such as China, Russia, Iran, and North Korea to use deepfakes as a "surgical strike" against targets until the 2020 U.S. elections when they will deployed on a larger scale[47].

Source: [47] Sekondary Infection, Graphika, 2020

The Liar's Dividend with George Floyd Death.

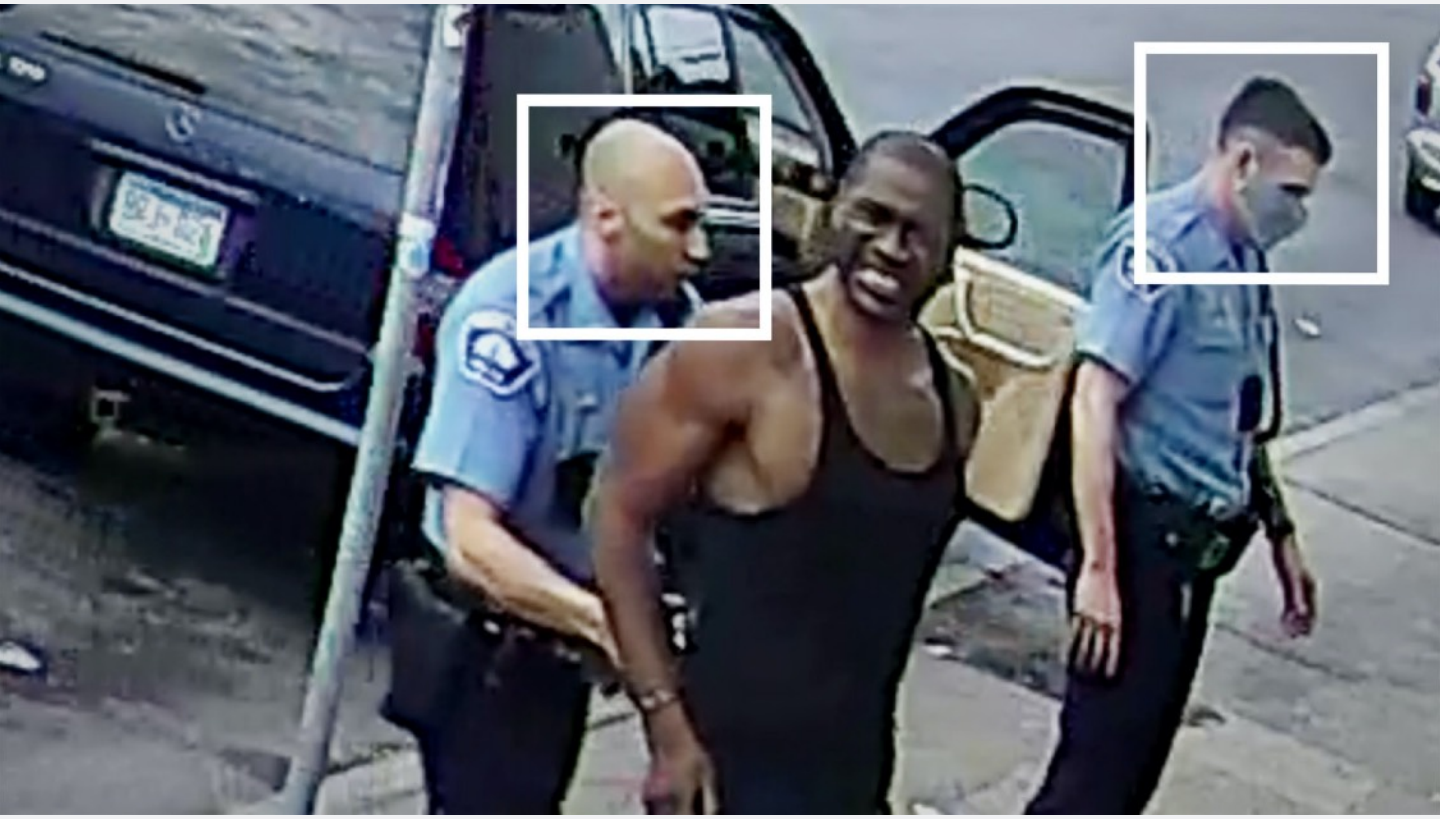


Figure 25. Footage of George Floyd's arrest and time in police custody is disputed as being a deepfake by a GOP candidate in the U.S

A GOP House candidate published a 23-page report claiming George Floyd death was a false flag operation. The document outlines that all individuals in the video are digital composites of two or more real people to form completely new individuals using deepfake technology[48].

When any video can be a deepfake, then instead of being deceived by deepfakes, people may grow to distrust all video and audio recordings as actors deflect and sow doubt on authentic media content.

Source: [48] The Hill, 2020

Next-Generation AI-Enabled Information Operations

Creating Deepfakes has become Easy and Inexpensive.

Cloud computing makes creation inexpensive

- There is a vast amount of publicly available face images and videos on the Internet which are used as the training data.
- The computational power required to run these networks is inexpensive and readily available using cloud computing.
- This low barrier to entry makes deepfakes a potent weapon for anyone with access to the internet and cloud hosted deepfake creation tools.

Various user-friendly tools have democratized access

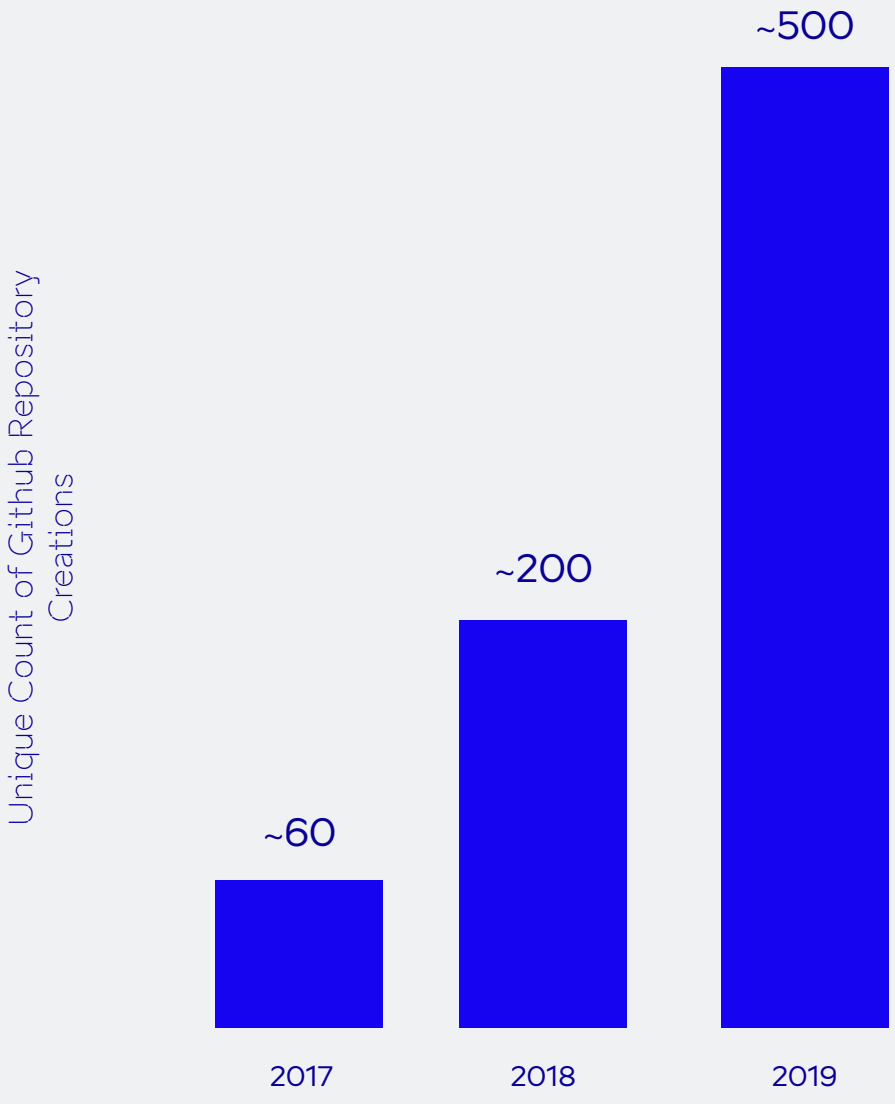
- This has led to a drastic inequality in creators versus those involved in detection of deepfakes - a ratio of around 100:1[4].
- Tools to create deepfakes have been productized and developed into easy-to-use interfaces for any non-technical individual to create a deepfake in minutes.

This is driving rapid innovation

- Easy access and creative capabilities mean that research and innovation in the field is growing.
- Creating deepfakes using different algorithms have led to improved deepfakes that are complex and hard to detect as synthetic/manipulated media.

Source: [4] Washington Post, 2019

Deepfake Attributed Repositories Have Grown 8X Since 2017 on Github.



Source: [44] Sentinel Analysis, 2020

Software as a Service (SaaS) Tools are Further Accelerating the Growth of Deepfakes.

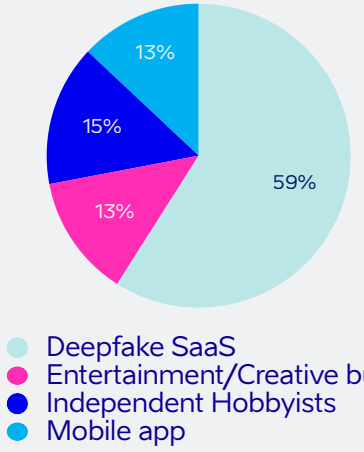
An assessment of the deepfake creation landscape found that four key groups are creating deepfakes today:

1. **Deepfake SaaS** tools that are consumer friendly, inexpensive, and easy to use (e.g. Avatarify)

2. **Entertainment/Creative businesses** that utilize deepfake technology for professional endeavours (e.g. VFX studios)

3. **Independent Hobbyists** such as technical individuals building their own deepfake AI models and publishing them as open-source GitHub repos

4. **Mobile apps** used by consumers purely for entertainment purposes (e.g. Zao App, REFACE App, Impressions App)



~59%

Are easy to use SaaS products for an average consumer.

Any user can create a deepfake in as little as three steps using an intuitive user interface[45].

Source: [45] Sentinel Landscape Analysis.
Note: Landscape analysis focuses on non-pornographic deepfakes.

Example of Democratized Deepfake Tool: Avatarify.

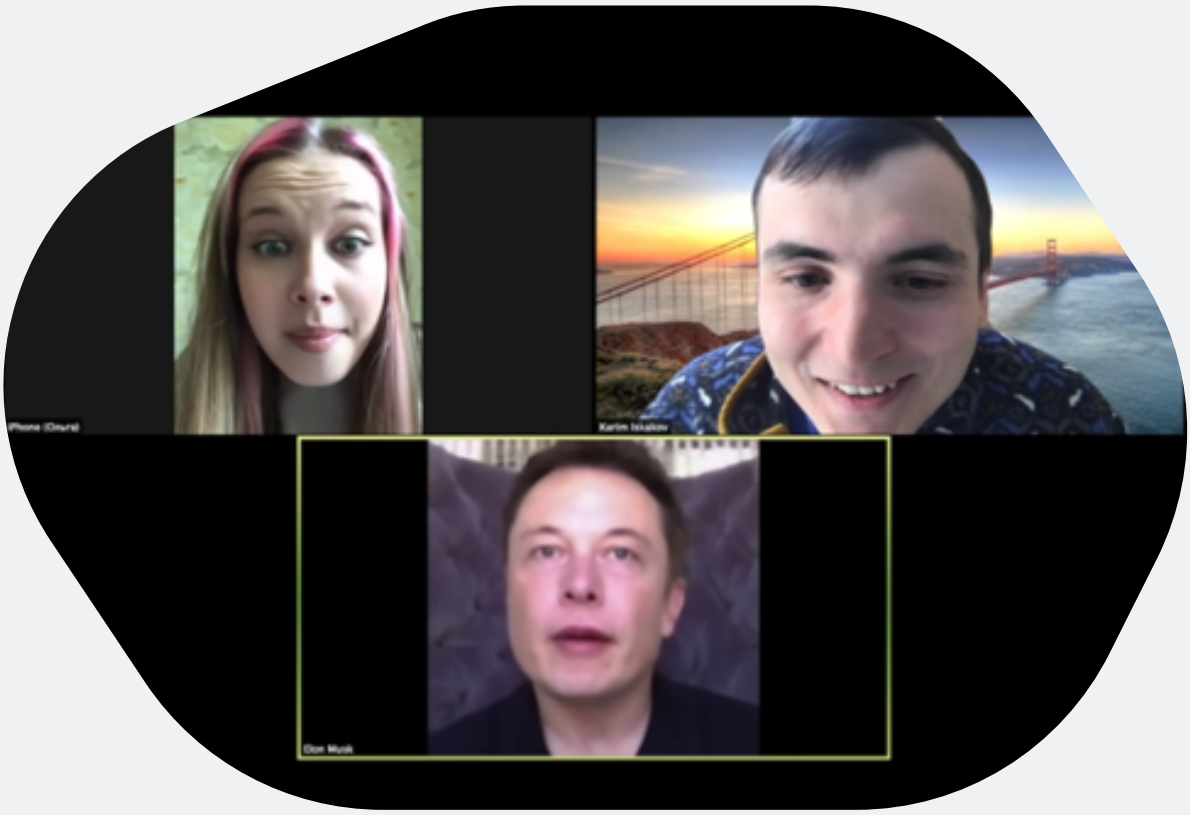


Figure 23. Screenshot of Avatarify where a Elon Musk deepfake has been added into a Zoom video conference in real-time.

Avatarify released a step by step guide to install a Windows application for free that allows anyone to synthetically generate a 3rd party in their videoconferencing sessions and manipulate them to say and do anything real-time.

The tool uses AI technology, specifically a first order motion model for image animation that can synthetically generate a deepfake to be manipulated in real-time using a single image as its training data [46].

Source: [46] Avatarify, 2020

Just one Selfie to Create A Deepfake with the #1 Entertainment App in the US.

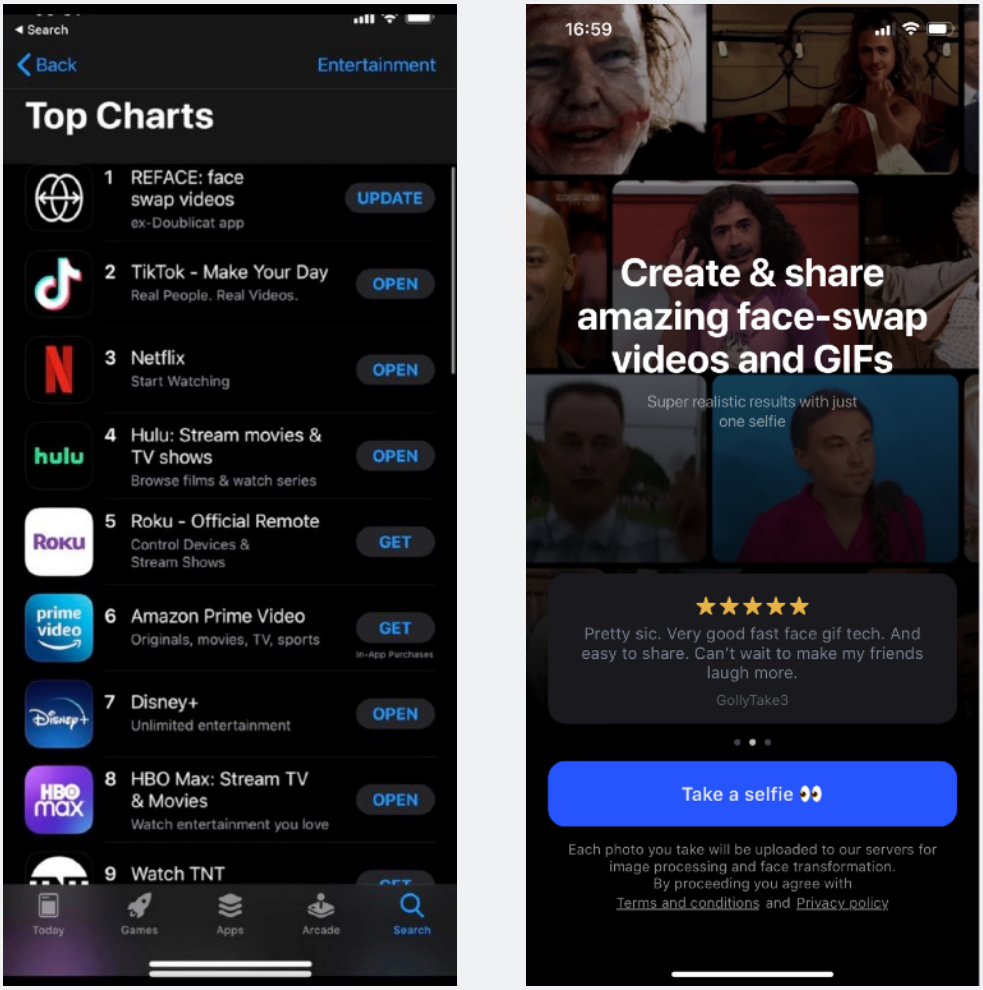


Figure 49. Screenshots of REFACE App shown as the #1 entertainment app in the US App Store.

The REFACE deepfake app hit #1 as the Entertainment app in the US App Store. It's hard to beat TikTok but REFACE app has done it because a single selfie is all that's needed to create a deepfake which can be instantly shared in social media.

Source: [95] REFACE App, 2020

The Three Stages of Deepfake Evolution.

Stage 1: one person can create one deepfake

2018

Creating deepfakes is expensive, resource intensive and technically challenging. Creators were technically savvy users who would spend 2 to 3 days developing a single realistic deepfake.

Stage 2: one million people can create one deepfake

2020

Creating deepfakes is affordable and easy. A single well-made deepfake would take a few minutes and can be developed using productized solutions such as SaaS tools or mobile apps by average consumers.

Stage 3: one person can create one million deepfakes

2021

Creating deepfakes is now scalable. Multiple variations of deepfakes can be created in a few seconds using a single image as training data, and disseminated through various channels automatically by a sophisticated adversary.

Next-Generation Information Operations: AI-Generated Events at Scale.

AI-generated information in its various forms including video, image, audio and text lead to their convergence. This results in whole events being generated at scale which are not based on reality but simulated by the AI with the aim of engaging users or conducting psychological warfare with the goal of influencing the behaviour of the information recipients.

For example, imagine every social media user receiving their own personalized AI-generated Coronavirus information, which is aligned with the user's beliefs and confirms their biases. Recently, Facebook started simulating its users through reinforcement learning by including on its site invisible synthetically-created user accounts to help it understand itself. Thus one can not rule out, that Facebook in the future will generate engaging content for their users through AI[49].

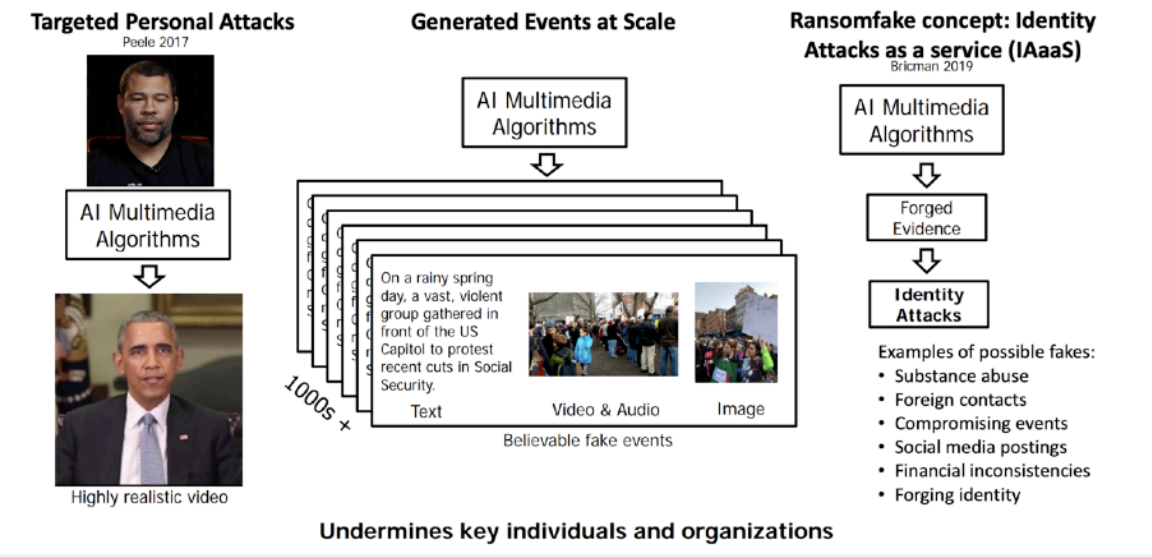


Figure 26. AI Generated events at scale by DARPA[50]

DARPA has a SemaFor program with the goal to develop technologies that determine if multi-modal media assets have been generated or manipulated[50].

Source: [49] Jack Clark, Open AI, 2020, [50] DARPA Media Forensics, 2019

Adversarial Landscape & Threat Actors.

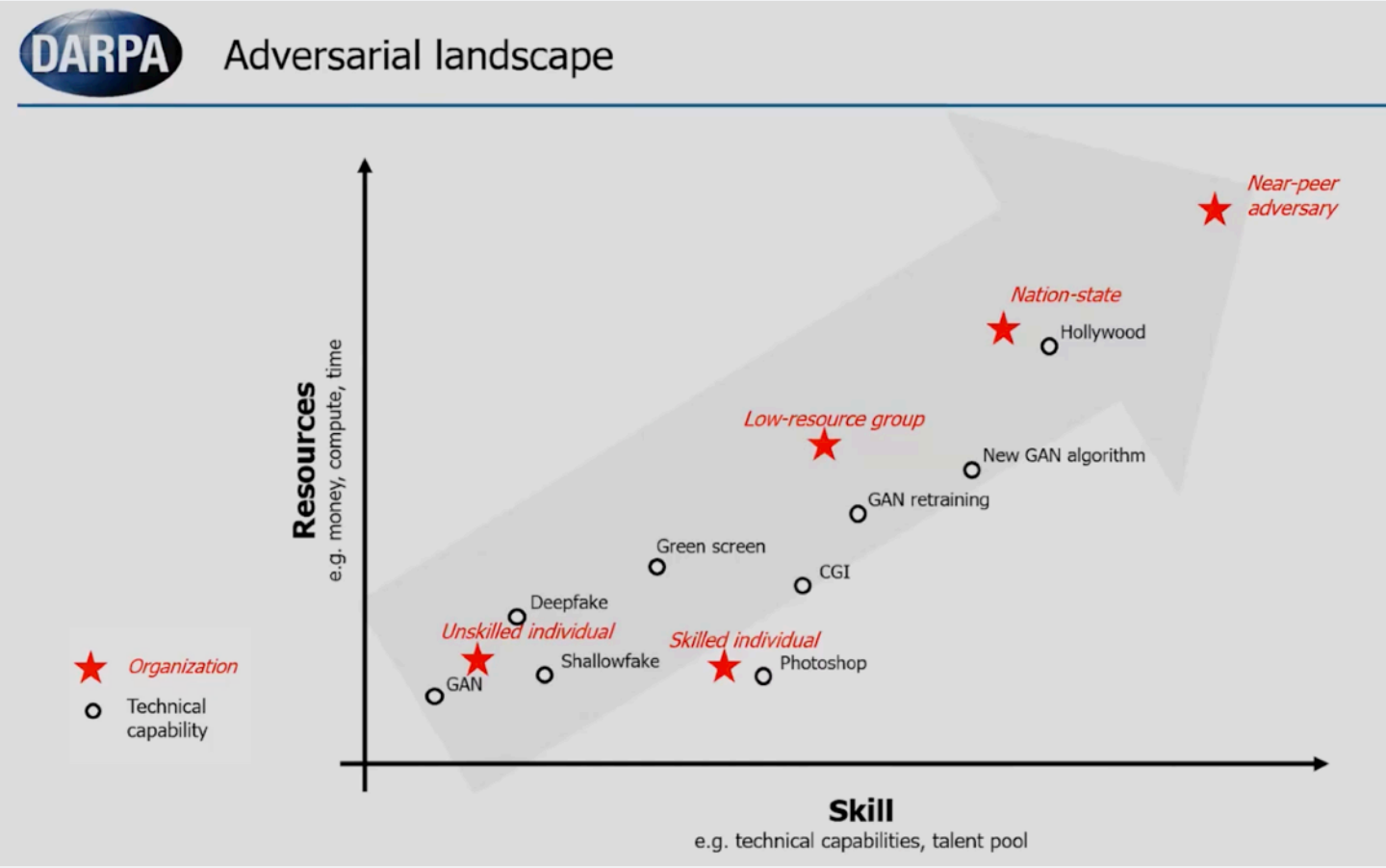


Figure 50. Deepfake adversarial landscape by DARPA[96]

Deepfake adversarial landscape and threat actors can be modelled in analogy to the cybersecurity threat landscape where actors with higher resources, expertise and motivation can develop and deploy more advanced deepfakes and AI-enabled information operations. For example, zero day deepfakes can be developed by advanced threat actors to evade detection methods for a period of time[96].

Source: [96] DARPA Media Forensics, 2020

Cheap Fakes

Before Deepfakes there were Cheap Fakes.

It's easy to focus on deepfakes and disregard their more simplistic version, cheap fakes. However, cheap fakes are equally insidious as they alter facts or remove content, resulting in a completely different action or narrative. Most importantly, they are even easier to create than deepfakes.

Cheap fakes are developed using the following tactics:

Content is removed to change the narrative or context.
E.g. frames or clips are edited out of the video

Content is manipulated to change the narrative or context.
E.g. frames or clips are sped up or slowed down in the video

Content is added to change the narrative or context.
E.g. frames or clips are inserted into the video

Cheap fakes rarely use AI technology and are created using simple video editing applications with very few technical skills required to master their use[40].

Source: [40] MIT Technology Review, 2019

Cheap Fakes: Content Removed.

Figure 20. Joe Biden cheap fake.

Video of Joe Biden suggesting White Nationalist themes was deemed to be edited and stripped of context.

Joe Biden was giving a speech on the culture of violence and its historical origins. But a video edited and stripped of this context was circulated implying Biden was sharing White Nationalist themes[41].

~1.6M views in 48 hours[41].

Source: [41] CNN, 2020.

Cheap Fakes: Content Manipulation.

Figure 21. Nancy Pelosi cheap fake.

Video of U.S House Speaker Nancy Pelosi edited, giving off the impression she was intoxicated.

Slurring of House speaker Nancy Pelosi's speech. The video is Nancy Pelosi but the pace of her speech has been slowed down by 25% and the pitch has been altered - making it sound like she was slurring and intoxicated[42].

3M+ views[42].

Source: [42] Washington Post, 2019

Cheap Fakes: Content Addition.

Figure 22. Screenshot of the news video cheap fake purported to be Beijing Capital Airlines.

A news video was shared on Facebook purported to be a Beijing Capital Airlines plane that did an emergency landing earlier in the day.

The video was spliced together from a computer-generated clip made by a filmmaker and animator a year earlier, into a real news footage[43].

14M+ views and publisher gained 225,000 new page followers[43].

Source: [43] Washington Post, 2018

Creating Deepfakes

The Production of Deepfakes Involves 4 Steps: Gathering, Extraction, Training, and Creation.

1 Gathering

Gathering involves scraping and downloading photos and/or videos of subjects to be deepfaked.

2 Extraction

Extraction involves the process by which input video is analyzed and for each frame in the video, the features of a face are identified and extracted with the use of face recognition.

3 Training

The neural network then begins to look at each of these extracted faces and understands the facial characteristics at a pixel level used to generate the face (e.g. skin tone, position of ears, eyes, nose, etc.). These pixel level characteristics form a basic design specification of a face or object.

4 Creation

With a trained neural network. The design specifications of each face provided to the model allows it to map face A onto face B.

The three most popular methods of creating deepfakes are: Autoencoders, GANs and graphical based implementations[51].

Source: [51] Zucconi Deepfakes Introduction, 2018

Autoencoders.

Neural networks use autoencoders to compress and decompress images. As videos are just frames of images stitched together, an autoencoder can be fed a series of frames to encode and form a compressed version of the image. Using an analogy, it is like zipping your files so they take up less space.

This compressed version contains a set of patterns that help represent the facial characteristics of the original image. This set of patterns are called latent features also known as hidden variables[51].

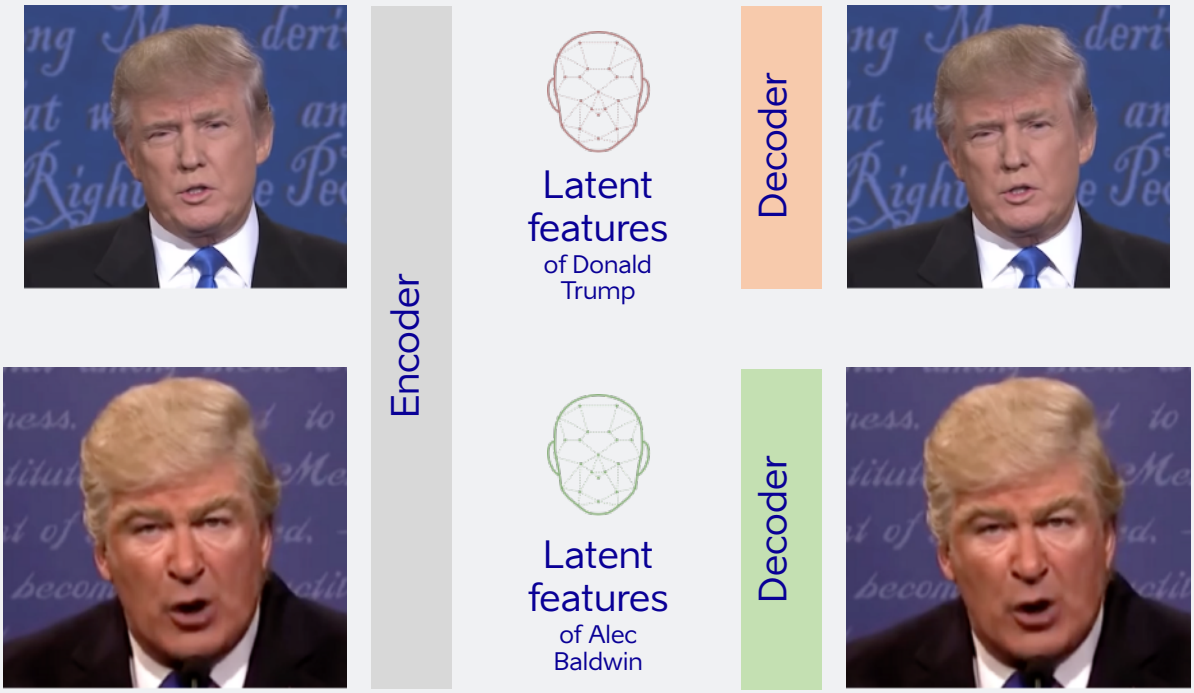


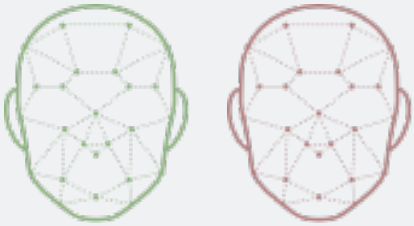
Figure 27. Illustration on how deepfakes are created through use of autoencoders with a single encoder and two decoders.

The same encoder is used to create the latent features from both images.

Separate decoders are then used to rebuild the images from the latent features to best replicate the original.

Source: [51] Zucconi Deepfakes Introduction, 2018

Autoencoders.



Latent features are facial characteristic patterns which dictate face angle, expression, lighting, skin tone, etc. They are decoded by the decoder to rebuild the original image.

In this example, the encoder has now created latent features of both Alec Baldwin and Donald Trump.

Latent features of Donald Trump

Latent features of Alec Baldwin

The decoder is now responsible for regenerating the full image frame of the face based on these latent features.

If you use the decoder that is trained to rebuild Alec Baldwin's face image but feed it the latent features of Donald Trump, the decoder maps characteristics such as expression, skin tone, lighting of Donald Trump onto Alec Baldwin's face.

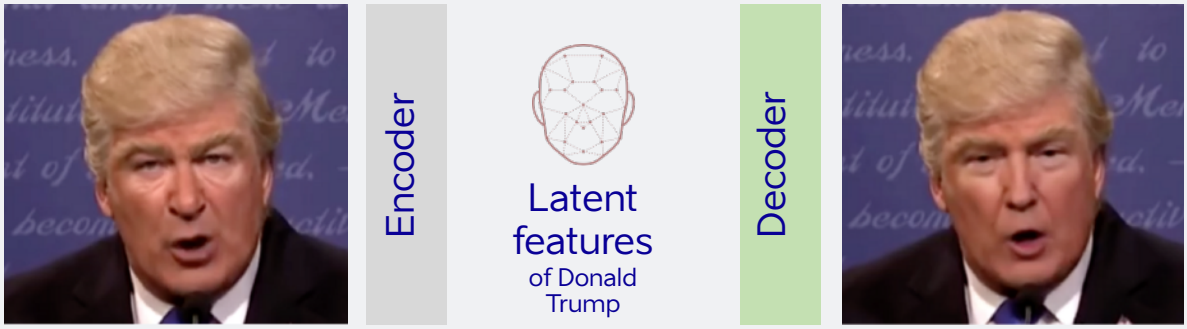


Figure 28. Latent features of Donald Trump used to place face on Alec Baldwin.

The reconstructed image on the right will have the facial characteristics of Trump while maintaining the facial expression of the original on the left.

Take a series of these images together, a set of frames, and it forms a deepfake video[51].

Source: [51] Zucconi Deepfakes Introduction, 2018

Generative Adversarial Networks (GANs).

GANs consist of two neural networks: a generator and a discriminator.

A generator exists to create new instances of data such as an image. A discriminator verifies the authenticity of an image by comparing it to the training to determine whether it was created by the generator or is an authentic real image[52, 53].

If we use an analogy:

1 A criminal (generator) counterfeits money.

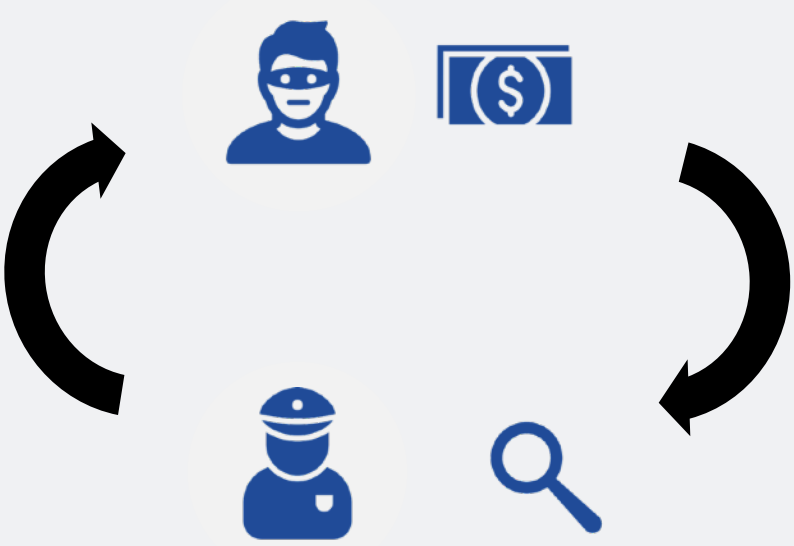


Figure 29. Illustrative example of how GANs work.

2 A policeman (discriminator) checks the money to see if it's real or not.

3

Each time the policeman catches a counterfeit, the criminal improves his technique to try and fool the policemen next time around.

The criminal keeps doing this until the policeman can't distinguish counterfeited money from real money anymore.

Source: [52] The rise of GANs, Use Journal, 2019; [53] AI Deepfake GANs, Vox 2019

Generative Adversarial Networks (GANs).

This adversarial training improves both networks and ultimately leads to the Generator creating high quality synthetic images that are indiscernible to real images.

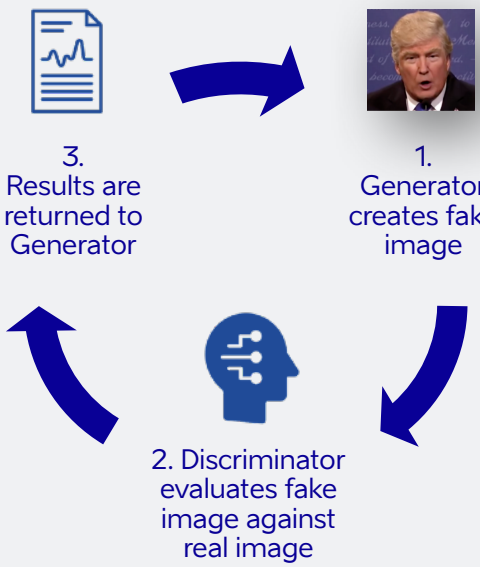


Figure 30. Illustrative example of how GANs work.

In each cycle, the real images of Trump act as the source of 'truth' that the Discriminator compares the fake images against to provide feedback to the Generator.

Initially the Generator has no training on how to synthetically create an image of Donald Trump from scratch, and the Discriminator has no training on how to identify what's defined as a synthetic image of Donald Trump. Over time, they both learn through this feedback loop and compete against each other to get better.

The discriminator predicts whether an image is real or fake and compares the prediction against the truth — whether it was training data or output from the generator. It uses this feedback to learn on a pixel by pixel level the nuances between synthetic and real using the training data and improves.

The generator evaluates how the fake images it is creating performs against the discriminator and uses the feedback to improve. It keeps generating new images by altering pixels every time the Discriminator flags its creation as fake to try and trick the discriminator into passing it as a real image[52, 53].

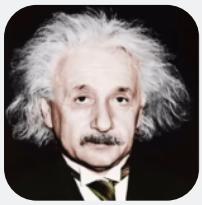
Source: [52] The rise of GANs, Use Journal, 2019; [53] AI Deepfake GANs, Vox 2019

Generative Adversarial Networks (GANs).

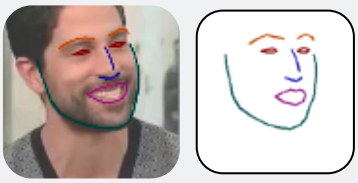
Compared to autoencoders, GANs only need a few training frames to synthetically generate a deepfake video.

For example, the Samsung AI Center in Moscow in partnership with the Skolkovo Institute of Science and Technology was able to generate a video deepfake of Professor Einstein using GANs and a Facial Landmark technique to input movement[54].

1 The GAN is fed the following image as training data:



2 An existing video of an actor is used to develop the underlying facial landmarks and movement to be used in the deepfake.



3 A synthetic version is created by the GAN of Einstein and using the facial landmark movements, a series of frames are created to develop a deepfake video.

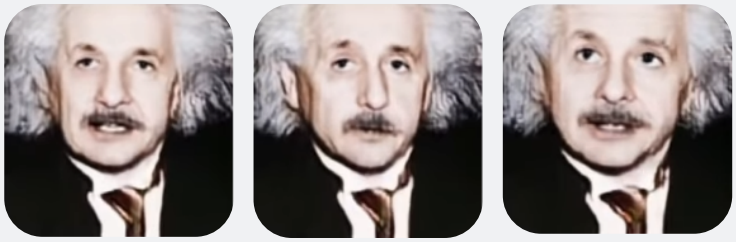


Figure 31. Illustrative example of GANs process results.

Source: [54] Few-Shot Adversarial Training of Realistic Neural Talking Head Models, Samsung AI Center, Moscow, 2019

Computer Graphics-Based Implementations.

Autoencoders and Generative Adversarial Networks (GANs) are deep learning based methods by which deepfakes are created.

An alternative to this is through the use of computer graphics-based implementation such as Face2Face.

These methods don't need learning as in deep learning but instead they rely on traditional graphical / visual effect techniques. This means the required compute for creating such deepfakes is lower but the quality also suffers through that[55, 56].

1. A source image is used to find specific facial landmarks (e.g. chin points, eyes, etc.)

2. These facial landmarks are then used to develop a 3D model of each face.

3. A series of target images (set of frames) are fed in and facial landmarks are located.

4. The 3D model is fitted against this target image's facial landmarks and then rendered to apply skin tone, texture, etc.

Source: [55] FaceSwap GitHub Marek Kowalski 2018; [56] Face2Face: Real-time Face Capture and Reenactment of RGB Videos Stanford University 2018

Face2Face.

Face2Face is an advanced real-time facial reenactment system, capable of altering facial movements in commodity video streams. They combine 3D model reconstruction and image-based rendering techniques to generate their output.

The Face2Face technique works by rendering a 3D model of both the source and target's face by tracking facial landmarks.

Source Actor

Target's face

Over the duration of the video, specific landmarks that modify expression are tracked (e.g. lip movements, eyebrows, far most angles on face, etc.) for both the source and target.

3D template

These tracked expression landmarks are then applied onto the 3D model and used to warp the 3D model to generate expressions.

3D template warped to change expressions

They are then used to reconstruct the target's face on the 3D model, but with the source actor's expressions applied[56].

Face2Face Facial re-enactment

Figure 32: Facial Re-enactment whereby source actors 3D template is then modified and projected onto Target's face to carry out manipulation of expression

Source: [56] Face2Face: Real-time Face Capture and Reenactment of RGB Videos Stanford University 2018

Implications

Deepfakes will Lead to Three Key Forces on Individuals and Organizations.

Disinformation

Video content has the effect of being a potent form of content for audiences to develop a strong reaction towards. Given the phenomenon of 'seeing is believing'. This will allow altered video to spread more quickly, and be less contested as inauthentic.

Exhaustion of thinking

Greater effort will be required by individuals to determine if media is manipulated or authentic. Given the prevalence of even trusted actors sharing inauthentic media by accident (e.g. influential journalists/celebrities), this uncertainty around 'truth' will be exhausting and individuals will either refrain from sharing content leading to truth decay, contesting those they have doubts on or eventually not caring and believing the content if it confirms their biases.

The Liar's Dividend

The Liar's Dividend proposes that the availability of highly realistic synthetic content such as deepfakes offers anyone (individuals or organizations) the ability to deflect accusations on media related to them as inauthentic and not true.

Implications for Individuals and Organizations.

Fraud & Espionage

Fraud, exploitation and espionage can be carried out using deepfakes for various purposes. It enables fake subjects to be generated in order to deceive a recipient or target of the fraud.

- Scams such as authorizing financial transactions purporting to be the authorized decision maker.
- Revenge porn involving synthetically generating the subject and seeking reward to prevent distribution of content.
- Industrial or military/political espionage through a highly personalized social engineering operation that would look like it's coming from a trusted source to extract information[57].

Harm

Deepfakes can be used to incite mental or physical harm on an individual or organization.

- Videos showing any criminal or insidious act can cause others to physically harm the individual.
- Example, a fake presidential announcement about a missile strike in an area can cause mass panic and hysteria[57].
- Pornographic deepfakes cause psychological harm and damage reputation.

Loss of IP Revenue

Deepfakes can be used to dramatically reduce or even avoid the cost of intellectual property and royalty fees.

- Celebrities can now be synthetically generated to carry out brand advertisements. This is lost revenue for the celebrity who is otherwise paid for such an endorsement[57].

Sabotage

Reputational sabotage or sabotage during key events can be carried out by deepfakes.

- Deepfakes can be released at critical events (e.g. elections) where there isn't enough time to debunk the content as inauthentic. This can influence decisions that can sabotage individuals or groups[57].

Source: [57] The Emergence of Deepfake Technology: A Review — Westerlund, 2019

Implications for Society.

Undermine Journalism

Volume and veracity of deepfakes can overpower fact checking amongst journalists. Moreover, video context that is of high value to the public may be withheld due to fears of it being inauthentic.

- A news organization may lose credibility after sharing deepfakes as authentic content.
- A news organization may influence the public mistakenly and cause harm by sharing a deepfake labelled as authentic.
- A news organization may be overwhelmed in their fact checking ability and withhold information due to fear of it being inauthentic[57].

Creating Divisions

Deepfakes can be used to enhance divides between society by using echo chambers to spread falsehoods that entrench opposing views further between groups.

- As an example, the Kremlin could deploy deepfakes to show muslims making derogatory remarks against whites in the U.S and incite further social division[57].

Manipulation of Collective Actions

Collective actions of a group can be manipulated though the use of deepfakes. This can prompt massive immediate changes in sentiment and action.

- Deepfakes on a political candidate up for election can sway voters to the opposing side[57].

Source: [57] The Emergence of Deepfake Technology: A Review — Westerlund, 2019

Implications for Society.

Disrupting Discourse

Public debates, discussions, and policies can be influenced through use of deepfakes that incite polarization, harassment, and exhaustion in individuals from participating further.

- Deepfakes can be used to introduce conflicting ideas and false information during policy debates on gun violence[57].

Eroding Trust

Any institution operating on trust will be vulnerable to the public being influenced by deepfakes that compromise their ethical standards and trust.

- Deepfakes can show institutions and entities such as judges or the IC engaged in corrupt activities[57].

Influencing Judicial Proceedings

Deepfakes can be submitted as real evidence during judicial proceedings. This can be used to influence the outcome of the case, sow doubt or clog the process.

- Deepfakes can be submitted by the defense as evidence to act as alibis or by the prosecution to assist in conviction. In either case, without verification that the evidence is manipulated, it can influence proceedings and the outcome of the court case.

Source: [57] The Emergence of Deepfake Technology: A Review — Westerlund, 2019

Example:
Resurrecting the Dead in Star Wars.

Five years ago, deceased actor Peter Cushing was synthetically generated using a stand-in actor for the movie.

Technology has no doubt improved since then and become inexpensive and democratized for any one to use[58].

Figure 33: Actor Peter Cushing synthetically generated in Star Wars movie.

What if deepfakes were used to rewrite history by unearthing individuals who are deceased and can no longer defend themselves when the deepfake is released?

We have already witnessed this amongst states using propaganda to rewrite political events. For example, Russia is investing significant effort to rewrite history and its involvement in World War II[59].

Deepfakes will now be an additional weapon for malicious actors to utilize in such information warfare initiatives.

Source: [58] The Guardian, Rogue VFX, [59] Foreign Policy, 2020

Implications for Governments.

Undermine Diplomacy

Foreign and domestic actors can leverage deepfakes to disrupt diplomacy between national entities. They can sway public opinion and be released in heated situations to spark rash decisions during critical chokepoints in national affairs.

- A deepfake released during a U.N National Security Council debate between nations may influence the outcome of votes on international security issues.
- Creating purposeful logjams in the state policy making processes can have implications for multilateral engagements and bilateral state relations[57].

Threaten National Integrity and Security

Deepfakes can be used to deny and deceive (D&D) military entities during times of war, and destabilize decision making when time is of the essence. They can be used to influence public opinion domestically and has ramifications for issues such as drafting and enacting of security policies. They will greatly increase the scale and effectiveness of hybrid and cyber attacks, whether by near-peer or asymmetric threats.

- Deepfakes showing U.S military in Iraq carrying out war crimes may impact local sentiment and place U.S personnel in harms way (PsyOps).
- This, in turn, may influence military policy stateside on U.S involvement in Iraq[57].
- Deepfakes allow simulation of friendly and enemy personnel over communications links and broadcast media.

Source: [57] The Emergence of Deepfake Technology: A Review — Westerlund, 2019

Benefits of Deepfakes

Benefits of Deepfake Technology.

While there are many negative implications to deepfake technology, benefits do exist:

Cloak Identity

Figure 34: An anonymous survivor of LGBTQ+ persecution faceswapped with a volunteer's face to anonymize them.

For example, the HBO documentary Welcome to Chechnya utilized deepfake technology to have the faces of volunteers mapped onto the anonymous LGBTQ+ survivors to prevent their prosecution in Chechnya [60].

Similar concepts are being explored by startups such as Alethea AI to develop deepfake—like avatars to digitally cloak users while retaining the original actor's emotion.

Democratize Hollywood

Deepfakes significantly lower the cost of image/video synthesis in the creative field. This democratizes access to small-scale creators who can now utilize deepfake technology to bring their imagination to life without the need for large budgets and expensive VFX technology[61].

Fully Personalized Music

Imagine John Lennon, Michael Jackson and Kanye West singing totally new songs for you in Spotify. Provided with genre, artist, and lyrics as input, Jukebox, a neural network released by Open AI generates music, including rudimentary singing, as raw audio in a variety of genres and artist styles[62].

Source: [60] Vox, 2020 [61] TechCrunch, 2019 [62] Open AI, 2020

Benefits of Deepfake Technology.

Improve Virtual Interactions

Deepfakes enable one to synthetically generate emotional expressions and voices to a greater level of accuracy. Technology such as Samsung's Neon utilize this to bring AI-powered virtual beings, and gaming companies are exploring how a player's selfie can be used to create their own 3D game avatars to improve the in-game experience[61].

Soon virtual interactions can occur in 3D online social spaces such as Facebook Horizon but with life-like avatars that accurately represent the user and their facial characteristics and expressions[63].

Recreate Lost Voices

Deepfake technology can also be used to recreate the voices of musicians and actors who have passed away. Additionally, for those who've lost their voice due to various ailments, voice synthesis can allow them speak again in their own voice using text-to-audio technologies.

Deepfakes are part of a developing industry and as the technology matures we can expect more legitimate use cases to appear.

Source: [61] TechCrunch, 2019 [63] Road to VR, 2019

Countering Deepfakes

Countering Deepfakes: Key Stakeholders.

Multiple stakeholders are required to counter the proliferation and efficacy of deepfakes.

As the volume, velocity, and virality of deepfakes grow, action will be required by all stakeholders affected:

Technology & Detection

Implementing appropriate technical measures and ethics policies for deepfake development, monitoring, control, and usage.

Government & Policy

The drafting of appropriate legislation and regulations governing deepfakes, and their distribution platforms.

Education & Civil Society

Appropriate training and education on deepfakes, disinformation, and media literacy.

Key stakeholders have a critical role to play in making sure harmful deepfakes do not have detrimental effects on democracy. The key areas that this effort should be directed towards are:

1. Identifying harmful deepfakes through the use of technology.
2. Preventing distribution of harmful deepfakes and locating original creators to prevent further releases through policy.
3. Educating society through the use of media literacy which would educate them on deepfakes, disinformation, etc.

Countering Deepfakes: Technology.

Effective technology and its ethical usage and development will underpin any measures to inhibit the distribution of deepfake technology by bad actors.

There is no single silver bullet which would perfectly detect all of the deepfakes all of the time because of their adversarial nature. Instead, the bar for creating undetectable deepfakes must be set extremely high thus taking it off from the hands of most adversaries. Technology today helps detect deepfakes and other manipulated media using the following methods[50]:

1. Digital Integrity
2. Physical Integrity
3. Semantic Integrity
4. Physiological Analysis
5. Biometric Analysis
6. Authentication
7. Audio Analysis

Digital Integrity

Signal:
Meta-data and signal level information such as compression stats, sensor noise, CFA interpolation, etc. can be evaluated to determine if the media has been doctored[50].

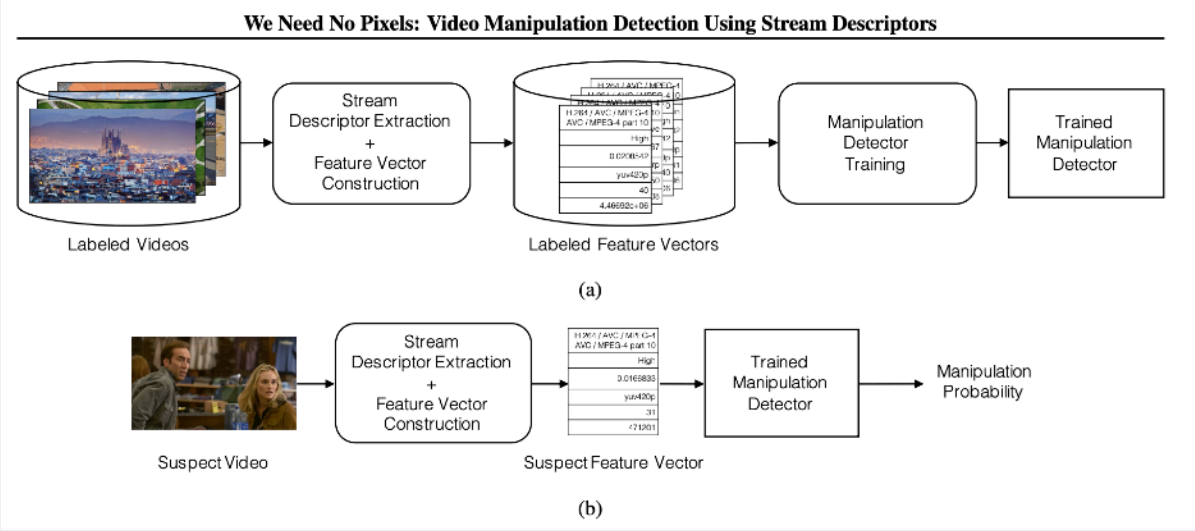


Figure 35: Given a suspect video, a feature vector is generated and processed by the previously selected detector. Finally, a manipulation probability for the suspect video is reported.

Source: [50] DARPA: Media Forensics 2019

Countering Deepfakes: Technology.

Pixel-level:
GANs, autoencoders, etc. make pixel level errors while creating deepfakes thus leaving behind "noise". These errors can be detected with computer vision algorithms in the form of deep learning or feature-based machine learning algorithms.

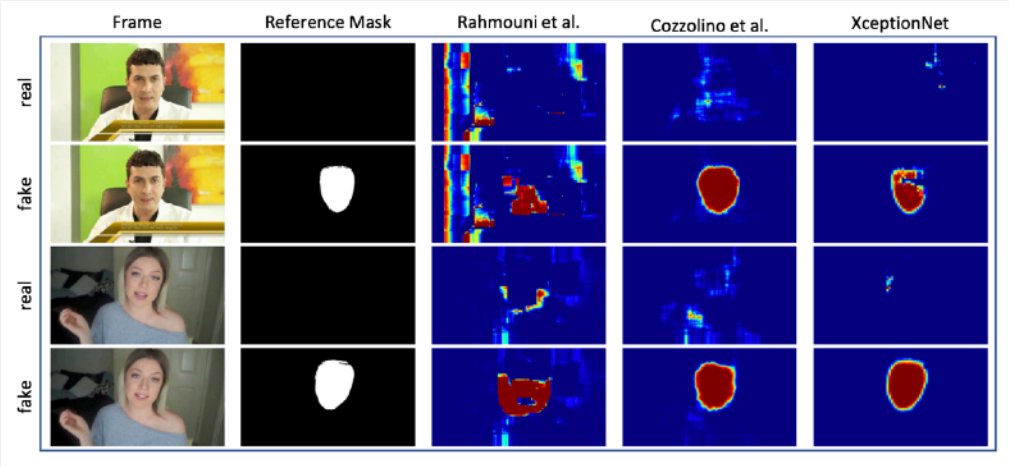


Figure 36: Pixel-level analysis of image through use of heatmaps.

Analyzing heatmap outputs using the XceptionNet segmentation model shows the prevalence of manipulation when looking at the original (first and third row) compared to the manipulated image (second and fourth row)[64].

Source: [64] FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces, 2018

Countering Deepfakes: Technology.

Physical Integrity

Environmental indicators can be used to verify if the laws of nature are followed in a particular piece of content. Looking for signs such as reflections, lighting, shadows, etc. can be used to determine if the media has been manipulated[50].

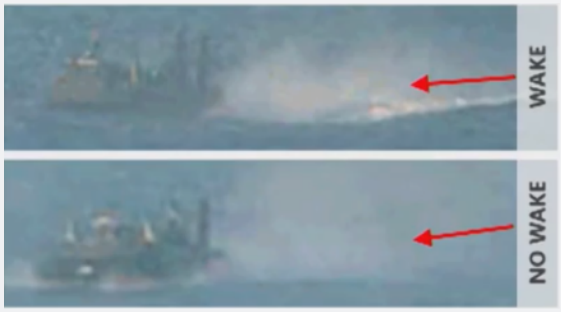


Figure 37: Physical level analysis of an image illustrating laws of nature defied.

Semantic Integrity

Contextual information related to a piece of content contradicts its representation. For example, are there date and time inaccuracies, or has the image been repurposed or placed out of its original context[50].

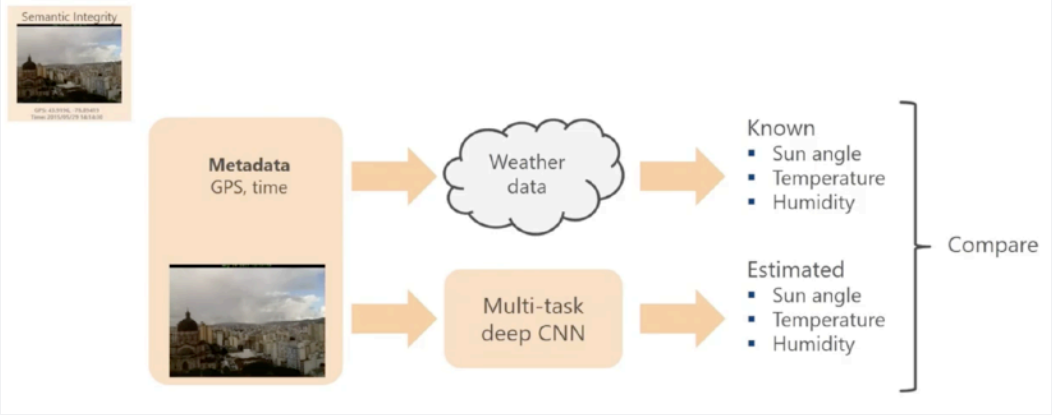


Figure 38: Semantic integrity analysis of an image through context comparison with weather data.

Source: [50] DARPA: Media Forensics 2019

Countering Deepfakes: Technology.

Physiological Analysis

Physiological signs related to the subject in a video can be verified to ensure its authenticity. Examples of such signs are if the individual is breathing or has a normal heart rate/rhythm, etc. [50]

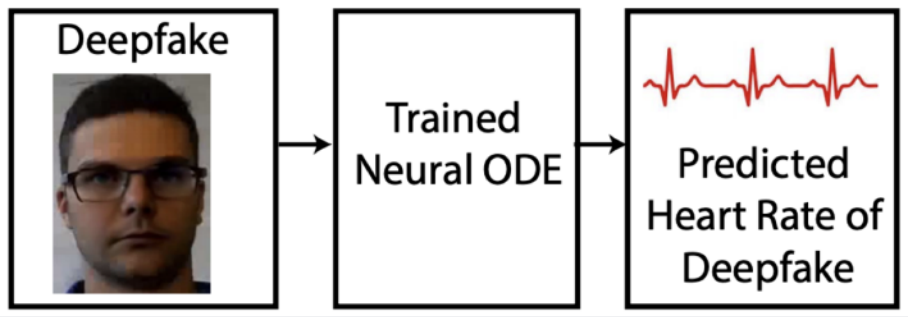


Figure 39: Physiological analysis using a trained neural ODE to predict heart rate and analyze for normality.

Biometric Analysis:

Soft-biometric data can be utilized to identify deepfakes. The technology utilizes machine learning to analyze a specific individual's mannerisms such as style of speech, movement, etc. to create a soft-biometric signature that can then be held up against a deepfake to identify whether a video is real or fake[50].

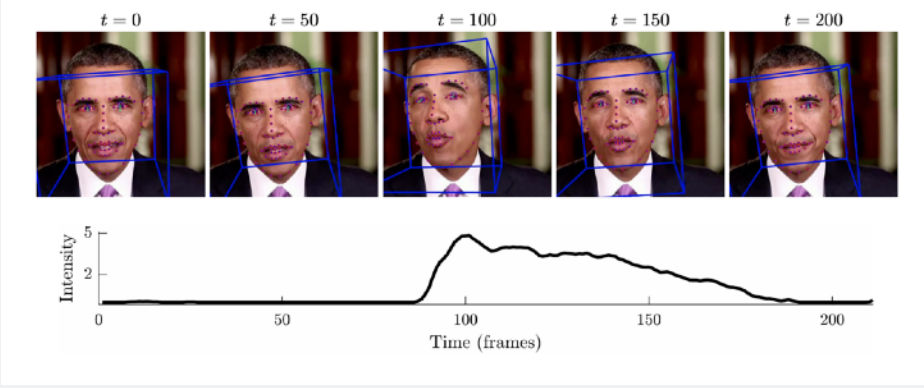


Figure 40: Soft-biometric analysis of deepfake against original to determine if same soft-biometric signature evident.

Source: [50] DARPA: Media Forensics 2019;

Countering Deepfakes: Technology.

Authentication:

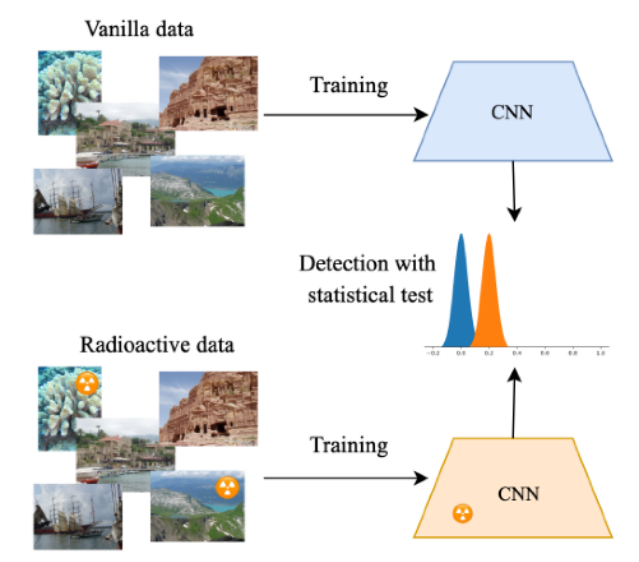


Figure 41: Illustration of radioactive data technique used in inferring AI generated images.

Using technologies such as radioactive data, watermarking and blockchain, authentication of media is carried out to ensure any individual or organization can verify its source, creator, and authenticity.

Radioactive data, for example, can be used to identify when synthetic media has been generated based on data sets that contain this 'radioactive' data[65].

Audio Analysis:

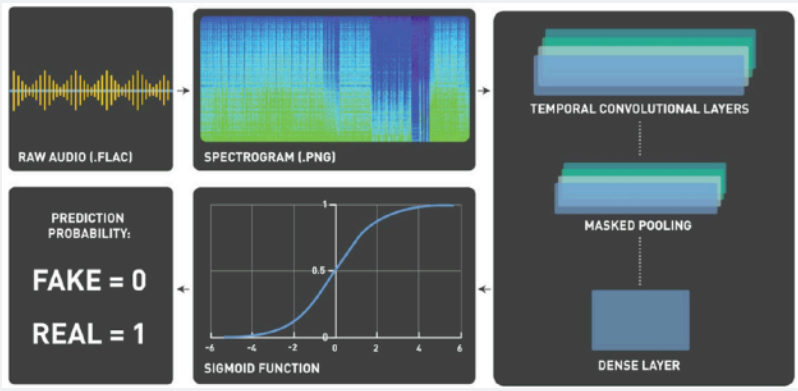


Figure 42: Spectrogram analysis of synthetic audio.

Analysis of audio samples generated by AI compared to human voice show visual differences when analyzed on spectrograms. These spectrograms can be used to train a neural network classifier that detects AI-generated voice.

Since deepfakes are increasingly now paired with synthetic audio, audio analysis can be effective in identifying deepfakes[66].

Source: [65] Radioactive Data: Tracing through Training, 2020, [66] Detecting Audio Deepfakes With AI, Dessa, 2019

Countering Deepfakes: Government & Policy.

In 2019, government entities began to take notice of the growing threats of deepfakes and enacted various pieces of legislation to address the threat.

In the U.S alone, various states and the federal government have initiated over 11 deepfake and broader synthetic media generation bills and policies to help curb the rise of deepfake porn and the use of deepfakes to manipulate democratic processes.

In addition to the U.S., the EU and China have drafted legislations and codes of practice to help tackle disinformation campaigns and revise their rules around handling fake news and disinformation to now include new forms of media such as deepfake videos.

Social media companies have also enacted policies to help detect and ban the use of deepfakes and other forms of manipulated media. Twitter, Reddit, Facebook, and TikTok have adopted content moderation policies to ban the use of deepfake and other manipulated media intended to deceive. Appendix A details all policies that have been implemented to date.

We can anticipate further policies being implemented going forward by both nations and private corporations to address the growing threat of deepfakes and disinformation campaigns in general.

Source: Appendix A

Countering Deepfakes: Education & Civil Society.

In order to effectively combat deepfakes and disinformation, the public must be well educated and informed.

Currently, the best method of prevention of the ill effects of deepfake technologies is teaching journalists, government officials, and the public at large about deepfakes and disinformation.

Media literacy:

Even when looking at a deepfake that may be undetectable to the naked eye, understanding where it originated, the credibility of the party spreading it, and the context behind it can help individuals, journalists, and others decide how much credibility a video should receive.

Case Study:

Reuters has been at the cutting edge of synthetic media detection and has been hired by Facebook to fact-check deepfakes and more. The team at Reuters put together an excellent course in Identifying and Tackling Manipulated Media which includes a Deepfake specific section[67].

Case Study:

Estonia is known as a digital leader in e-governance and being a highly digital society. This means that the nation must be ever vigilant to threats, especially given the history of Soviet Occupation and cyber attacks such as the one that affected a huge portion of the Estonian population in 2007. Estonia has responded to the threats with media literacy education for government employees, ordinary citizens and high school students with a mandatory media literacy course. The country has created a cyber defense volunteer unit, worked to bring the NATO CCDCOE to the capital, and has prioritized cyber hygiene with the President making it a core issue of her work.

Case Study:

Bellingcat is one of the premier global open source fact-checking organizations, Bellingcat has dozens of resources, guides and training to understand and identify disinformation. While they are best known for their work on the downing of MH17 by Russia over Ukraine and being the first to report on the identity of one of the Skripal attackers, Bellingcat has also put together excellent resources pertinent today in the wake of the COVID-19 pandemic that support individuals and institutions investigating Coronavirus disinformation[68].

Source: [67] Reuters 2019; [68] Bellingcat 2020

The U.S. 2020 Elections

What to Expect in 2020.

2020 will be the year deepfakes shift from an emerging threat to one that is mainstream and a staple weapon used by entities in information warfare.

Events

- U.S elections in 2020 will see the use of deepfakes and cheap fakes to influence the public, especially at critical junctures in the election process.
- Political events in other nations (e.g. Iran, Guineau, Ethiopia, Poland, Sri Lanka, etc.) with fewer policies and minimal tracking of disinformation/ propaganda campaigns will see deepfakes emerging and having a strong effect in disrupting political discourse.
- Adjacent events of public interest such as Climate Change, COVID-19 crisis or Immigration will see the use of deepfakes.

Actors

- State actors will use deepfake technology to further influence actions home and abroad.
- Individuals and groups will begin to use deepfake technology more prevalently in malicious civil use cases such as to commit fraud.
- Individuals and groups will begin to use the advent of deepfake technology to dispute any media evidence used against them to sabotage or counter their narratives.

Technology

- Deepfake technology will see rapid improvements alongside the development of detection technology by organizations and governments. This will lead to an on-going arms race and an exponential increase in the advancement of deepfake technology.
- Creation of deepfakes will be paired with automation technologies that rapidly generate and disseminate deepfakes across various channels, making it harder to detect and thwart before it influences audiences.
- Synthetically generated individuals and organizations will supplement the existing growth and prevalence of bots on the internet, elevating the spread of disinformation.

> “Adversaries and strategic competitors probably will attempt to use deep fakes or similar machine-learning technologies to create convincing — but false — image, audio, and video files to augment influence campaigns directed against the United States and our allies and partners...The threat landscape could look very different in 2020 and future elections.”
>
> - Senate Select Intelligence Committee, 2019

Source: [69] Worldwide Threat Assessment - U.S Intelligence Community, 2019

The U.S. 2020 Elections.

Through strategic intelligence indications and warning methodology we assign a high likelihood (80%-90%) that the U.S 2020 elections will be the first-time next generation AI-based information operations are deployed.

Deepfakes will be used against the U.S and will be combined with other active measures practices such as those carried out by Cuccifer 2.0 against the DNC in 2016 and Cambridge Analytica-like psychological operations.

We bring out four possible scenarios of many in which deepfakes can interfere with the U.S. 2020 elections:

1. Suppression of vote

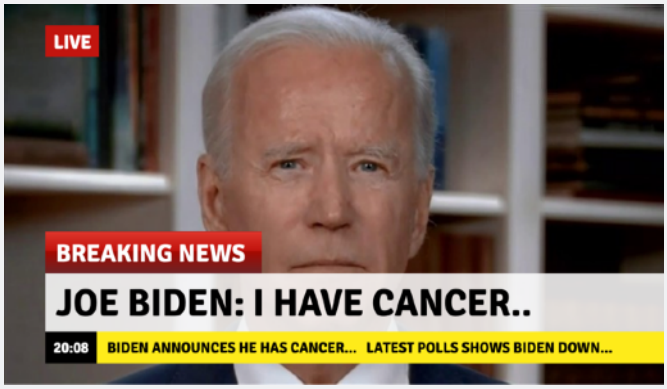


Figure 43. Illustrative example of a Joe Biden deepfake aiding in suppression of votes.

For example, AI-generated audio and videos of Presidential nominee Joe Biden announcing he has cancer could be released.

President Trump could recirculate these deepfakes on his social media amplifying its reach as he already has done[70].

The rapid circulation of these deepfakes in a short amount of time combined with a seeing is believing effect might sway votes.

2. Profiting through liar's dividend

Figure 44. Illustrative example of a Trump tweet showing liar's dividend and how it can seed doubt.

For example, compromising material released on President Trump is instantly refuted by him as a deepfake.

Any accusation towards him can now be dispelled as fake and seed doubt into the general public's minds.

Source: [70] Twitter, 2020

The U.S. 2020 Elections.

3. Inhibiting voter turnout

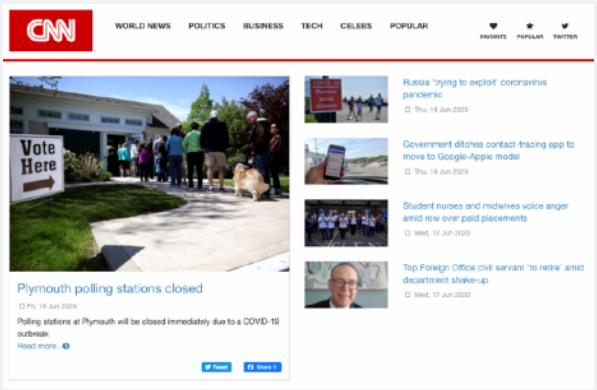


Figure 45. Illustrative example of a fake news article mimicking an authoritative source stating polling stations are closed.

For example, neural fake news mimicking authoritative sources can be generated at near zero cost. This leads to information operations that can be conducted by small groups of actors which makes such operations harder to track. During the 2016 elections, Macedonian private actors earned ad revenue through the use of fake news with pro-Trump websites[71]. Tools for generating plausible AI-propaganda exist online for free[72].

4. Create chaos around election results

Figure 46. Illustrative example of a news clip announcing U.S elections were rigged.

For example, as votes are tallied and the election result is announced, deepfakes can be leaked of election judges who are in charge of overseeing voting results and safeguarding election machines. These deepfakes will state the machines have been tampered with to provide inaccurate final vote counts, and that the election was rigged. This will lead to confusion and doubt on the electoral process, and ultimately fracture the trust U.S citizens have in their democratic system. In 2016, Russia had planned to launch a similar disinformation campaign if Hillary Clinton had won[73].

Source: [71] Wired, 2017; [72] Grover: Allen Institute for AI, 2020 [73] U.S Department of Justice: Report On The Investigation Into Russian Interference In The 2016 Presidential Election, 2019

We are at the Tipping Point.

Throughout human history from the Gutenberg press, modern day radio, television and the Internet, information has become quicker to produce and easier to spread thus reaching nearly every household who uses the Internet.

Figure 47. Illustrative example of the rapid progression in information creation and distribution over time.

However, the production of most information has still been expensive because of the human-in-the-loop, that is until the advent of artificial intelligence. This means videos, images, audio, and text can be synthetically generated at scale.

Imagine a world where AI generated deepfake videos, images, audio, and news articles are created to be highly personalized to one's preferences based on their data. This can be done at close to zero cost and at the scale of hundreds of billions.

Pair this with inexpensive distribution channels such as social media and this leads to the capability to manipulate citizens, damage the economy and disrupt democratic processes at close to zero cost and at exponential speed. We are at this tipping point.

This exponential growth will lead to the majority of the world's digital information being produced by AI, which in turn necessitates a trust layer for the Internet where people know the information they receive is coming from a real person and not malicious.

Sentinel is executing towards this vision for a trust layer together in partnership with the Estonian Government, which is a world leader in e-governance, digital identity, and cybersecurity.

Governments, civil society and organisations should invest immediately into education to prevent society's susceptibility to deepfakes, fund technological countermeasures and draft required legislation. The societies who will adapt the fastest to these changes will dominate the 21st century information environment and the ones who do not will find themselves struggling to catch up as the exponentially evolving information environment does not stop.

J. Tammekänd

Johannes Tammekänd, CEO & Co-Founder

https://thesentinel.ai

Appendix A: Government & Policy Initiatives.

In 2019, government entities began to take notice of the growing threats of deepfakes and enacted various pieces of legislation to address the threat:

U.S Government:

- February, 2019: HB 2678 bill was passed in the state of Virginia that allows residents of Virginia that were victims of deepfake pornography to sue their creators[74].
- April, 2019: SB751 bill was passed in the state of Texas that criminalizes the creation of deceptive video such as deepfakes with the intent to influence election outcomes[75].
- May 2019: The National Defense Authorization Act for 2020 instructed the formation of a deepfakes competition to stimulate research and development of technologies to detect manipulated media, and award a prize of up to $5 million[76].
- September, 2019: HR 4355 — The IOGAN Act (Identifying Outputs of Generative Adversarial Networks Act) was introduced to ensure research and development of technology to identify deepfakes and prevent any malicious harm[77].
- September, 2019: The Deepfake Report Act was introduced by the Senate to direct security establishments in the U.S to analyze and determine the threat of deepfakes to national security[78].
- October, 2019: AB-730 Elections: deceptive audio or visual media bill was passed in California that makes it illegal for doctored media of politicians to be circulated within 60 days of an election[79].
- October, 2019: AB-602 bill was passed in California that allows California residents that were victims of deepfake pornography to sue their creators[80].
- November, 2019: NIST's (National Institute of Standards and Technology) FARSAIT (Fundamental and Applied Research and Standards for AI Technologies) program is supporting several research endeavors related to reducing bias in A.I systems, and track the development of GANs[81].

Source: [74] Viriginia Legislative System, 2019; [75] Texas Gov, 2019; [76] NDAA Congress.gov, 2019; [77] HR4355 Congress.gov, 2019; [78] Deepfake Report Act, Congress.gov, 2019; [79] AB-730 California Legislative Information, 2019; [80] AB-602, California Legislative Information; [81] FARSAIT, NIST, 2019

Appendix A: Government & Policy Initiatives.

U.S Government:

- December, 2019: The Deepfake Accountability Act was introduced to criminalize the use of synthetic media and directed the Department of Homeland Security to create a task force to address the threat[82].
- February 2020: U.S Department of Defense adopts ethical principles for the use of Artificial Intelligence in national security applications[83].
- March 2020: HR6088 - Deepfakes in Federal Elections Prohibition Act to prohibit the distribution of materially deceptive audio or visual media prior to an election for Federal office, and for other purposes[84].

Source: [82] Deepfake Accountability Act, Congress.gov, 2019; [83] Defense.gov, 2020; [84] HR6088 Congress.gov, 2020

Appendix A: Government & Policy Initiatives.

Non U.S Governments:

- April, 2018: Belgium published a strategy on tackling disinformation including guidelines for combatting deepfakes[85].
- December, 2018: European Union releases a Action Plan against Disinformation that will support, where appropriate, information campaigns to raise users' awareness of the most recent technologies (e.g. deepfakes), and mobilize private sector parties to tackle disinformation[86].
- November, 2019: China introduces law that criminalizes the publishing of deepfakes or fake news without appropriate disclosure that it has been manipulated[87].

Corporations:

- November, 2019: Twitter announces a draft policy intended to tackle manipulated media such as deepfakes[88].
- January, 2020: Reddit bans the use of impersonation content such as deepfakes on its platform[89].
- January, 2020: Facebook bans the use of manipulated videos and photos on its platform that were generated using machine learning or artificial intelligence[90].
- January 2020: TikTok updates its Community Guidelines to ban use of manipulated media to spread misinformation[91].
- February 2020: YouTube releases public statement reminding users to not public election-related content such as deepfakes intended to mislead users and spread misinformation[92].
- March, 2020: Twitter adopts a synthetic and manipulated media policy to enforce labeling on any tweet to help readers understand its authenticity[93].

Source: [85] European Commission, 2018; [86] European Commission, 2018; [87] South China Morning Post, 2019; [88] Twitter, 2019; [89] Verge, 2020; [90] Verge, 2020; [91] TikTok, 2020; [92] YouTube, 2020; [93] Twitter, 2020

References.

- [1] University of Oxford 2019: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf
- [2] Economic cost of bad actors on the internet, University of Baltimore, 2019: https://info.cheq.ai/hubfs/Research/THE_ECONOMIC_COST_Fake_News_final.pdf
- [3] The spread of true and false news online — MIT, 2018: https://science.sciencemag.org/content/359/6380/1146
- [4] Washington Post, 2019: https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned
- [5] Sentinel Analysis, 2020: For 2020 deepfake video count, analysis of count of videos uploaded to the following sites: YouTube, Facebook, TikTok, Instagram, Twitter, Yoku, Iqiyi, Tencent Video, Bilibilli, Nicovideo, ok.ru, Vimeo, Dailymotion, and 30+ porn sites included dedicated deepfake porn sites we can disclose in a ROI. These videos were captured through the use of official APIs, keyword search, and scraping techniques and duplicate videos based on URL were removed. For YouTube and Twitter's we used their APIs. Sentinel searched for videos associated with the following terms: deepfake, deepfakes, faceswap, faceswaps, faceapp, fakeapp, zaoapp. Videos were attributed to creation date and we collected videos starting November 2017 until June 2020. Deepfake video count growth metric estimated using DeepTrace 2019 deepfake video count (14,678) value against 2020 Sentinel deepfake video count value.
- [6] DeepTrace, 2019: The State of Deepfakes: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf
- [7] Sentinel Analysis, 2020: Sentinel analysis of YouTube for the following videos with attributed keywords: deepfake, deepfakes, faceswap, faceswaps, faceapp, fakeapp, zaoapp from November 2017 to June 2020. Sum of total views across all videos. (N=38,242). For TikTok view count, Sentinel analysis of unique URL videos associated with the following keywords: deepfake, deepfakes, faceswap, faceswaps, fakeapp, zaoapp, doublicat, impression app. Duplicate videos based on URL were removed. (N=3,898)
- [8] Forrester, 2019: https://www.zdnet.com/article/2020-is-when-cybersecurity-gets-even-weirder-so-get-ready/
- [9] Sentinel Analysis, 2020: Analysis of videos from Sentinel's data pipeline of verified deepfakes using Sentinel's IP and deepfake detection platform alongside human verification by specialists. Subset of N=7000 videos that are verified deepfakes and labeled by source, face count, and category reported in data. Videos analyzed range from November 2017 to May 2020. Facecounts defined as: Face swap: These are videos that involve an individual's face overlaid or mapped onto the face of another person. In essence, the face has been swapped out and another has been stitched in to replace it. Facial Re-enactment: These are videos where the original subject's facial features have been manipulated to carry out fake actions. Full body deepfake: These are videos where the subject's body is visible and is manipulated. Full and/or partial body including body parts are visible and it's clear the body has been synthetically generated as opposed to a faceswap or facial re-enactment on an existing body.

References.

- [10] Sentinel Analysis, 2020: Analysis of videos from Sentinel's data pipeline of verified deepfakes using Sentinel's IP and deepfake detection platform alongside human verification by specialists. Subset of N=7000 videos that are verified deepfakes and labeled by source, face count, and category reported in data. Videos analyzed range from November 2017 to May 2020. News or Political is defined as: Any news reel, or video anchored around political events but not intended to entertain viewers but spark discussion; Entertainment is defined as: To provide amusement, entertainment, etc; Explanations is defined as Providing an explanation on a particular subject, excludes videos if instructions on how to create, or build a deepfake are provided. Other label consists of videos labelled: Harmful; Porn; Controversial; Tutorial.

- [11] Sentinel Analysis, 2020: Sentinel Analysis, 2020: Analysis of videos from Sentinel's data pipeline of verified deepfakes using Sentinel's IP and deepfake detection platform alongside human verification by specialists. Subset of N=7000 videos that are verified deepfakes and labeled by source, face count, and category reported in data. Videos analyzed range from November 2017 to May 2020. One face defined as a singular face in the video. Many face defined as more than one individual's face found in the entirety of the video or in a single given scene/frame.

- [12] Sentinel Analysis, 2020: Analysis of videos from Sentinel's data pipeline of verified deepfakes using Sentinel's IP and deepfake detection platform alongside human verification by specialists. Subset of N=7000 videos that are verified deepfakes and labeled by source, face count, and category reported in data. Videos analyzed range from November 2017 to May 2020. Sources used are YouTube, Instagram, Twitter, Facebook.

- [13] Sentinel Analysis, 2020: Analysis of pornographic deepfake videos from 31 porn sites. Note popular platforms such as Pornhub.com were analyzed but have banned the use of deepfakes. Analysis performed using the following search terms: deepfake, deepfakes, faceswap, faceswaps, fakeapp (N=27,271). Sites dedicated to deepfakes defined as sites where primary brand messaging and value proposition was the access to deepfake pornographic videos. Eastern porn sites defined as porn sites belonging to eastern hemisphere with primary language not English. Western porn sites defined as porn sites targeted to western hemisphere users with primary language as English.

- [14] Sentinel Analysis, 2020: Sentinel performed a qualitative analysis of popular deepfake forums online such as famousboard.com and mrdeepfake.com's forum page. In addition, Sentinel analyzed popular Telegram Chat groups that had a userbase dedicated to the creation and dissemination of deepfake material. Analysis performed June 2020.

References.

- [15] Loyola Law School, 2019: https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5640&context=flr
- [16] India Today, 2018: https://www.indiatoday.in/trending-news/story/journalist-rana-ayyub-deepfake-porn-1393423-2018-11-21
- [17] ABC Australia, 2018: https://www.abc.net.au/radionational/programs/earshot/noelle-martin-personal-story-of-revenge-porn-and-deepfakes/11417940
- [18] WhoIsHostingThis, Dodging Deception & Seeking Truth Online Survey, 2020: https://www.whoishostingthis.com/blog/2019/09/02/seeking-trust-online/
- [19] Pew Research Center Survey, 2019: https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/
- [20] Sentinel Analysis, 2020: An assessment of U.S legislation and bills proposed during 2019 that were attributed to "deepfakes" and "synthetically generated media". See Appendix A for sources.
- [21] Facebook, Deepfake Detection Challenge, 2019: https://deepfakedetectionchallenge.ai/
- [22] FaceForensics++: Learning to Detect Manipulated Facial Images 2016: http://www.graphics.stanford.edu/~niessner/thies2016face.html
- [23] Face2Face: Real-time Face Capture and Reenactment of RGB Videos Stanford University 2018
- [24] Neural Puppetry, Technical University of Munich, 2019: https://arxiv.org/pdf/1912.05566.pdf
- [25] Protecting World Leaders Against Deep Fakes, University of California Berkeley, 2019: http://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf
- [26] Everybody Dance Now: Motion Retargeting Video Subjects, University of California Berkeley, 2019: https://arxiv.org/pdf/1808.07371.pdf
- [27] Vice, 2018: https://www.vice.com/en_us/article/3k7mgn/baidu-deep-voice-software-can-clone-anyones-voice-with-just-37-seconds-of-audio
- [28] The Verge, 2019: https://venturebeat.com/2019/12/17/resemble-ai-launches-voice-synthesis-platform-and-deepfake-detection-tool/
- [29] Wall Street Journal, 2019: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
- [30] Venturebeat, 2019: https://www.theverge.com/2019/9/5/20851248/deepfakes-ai-fake-audio-phone-calls-thieves-trick-companies-stealing-money

References.

- [31] Analyzing and Improving Image Quality of StyleGAN, Nvidia, 2019: https://arxiv.org/abs/1912.04958
- [32] Few-Shot Adverserial Training of Realistic Neural Talking Head Models, Samsung AI Center, Moscow, 2019; Analyzing and Improving Image Quality of StyleGAN, Nvidia, 2019: https://arxiv.org/pdf/1905.08233v1.pdf
- [33] Defending against neural fake news, University of Washington, 2019: https://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf
- [34] Politico, 2019: https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/
- [35] Mother Jones, 2019: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/
- [36] India Today, 2018: https://www.indiatoday.in/trending-news/story/journalist-rana-ayyub-deepfake-porn-1393423-2018-11-21
- [37] MIT Technology Review, 2020: https://www.technologyreview.com/f/615247/an-indian-politician-is-using-deepfakes-to-try-and-win-voters
- [38] Reuters, 2019: https://www.reuters.com/article/us-china-xinjiang-turkey/china-releases-video-it-says-proves-reports-of-uighur-poets-death-untrue-idUSKCN1Q00E7
- [39] The Brussels Times, 2020: https://www.brusselstimes.com/all-news/belgium-all-news/politics/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-covid-19-with-climate-crisis/
- [40] MIT Technology Review, 2019: https://www.technologyreview.com/s/613172/deepfakes-shallowfakes-human-rights/
- [41] CNN, 2020: https://www.cnn.com/2020/01/03/politics/biden-clip-inaccurate-white-nationalism-fact-check/index.html
- [42] Washington Post, 2019: https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/

References.

- [43] Washington Post, 2018: https://www.washingtonpost.com/technology/2018/10/18/i-fell-facebook-fake-news-heres-why-millions-you-did-too/?utm_source=reddit.com
- [44] Sentinel Analysis, 2020: GitHub Repository creation dates; Analysis by Sentinel using keywords: "deepfakes"; "deepfake"; "Faceswap"; "Face2Face" for timeframes Year 2017, 2018, 2019
- [45] Sentinel Landscape Analysis, 2020: Desk qualitative research of creators in non pornographic deepfake space on January 2020 (n=34)
- [46] Avatarify, 2020: https://github.com/alievk/avatarify; https://www.youtube.com/watch?v=lym9ANVb120
- [47] Graphika: Sekondary Infektion, 2020: https://secondaryinfektion.org/downloads/secondary-infektion-report.pdf
- [48] The Hill, 2020: https://thehill.com/homenews/house/504429-gop-house-candidate-publishes-23-page-report-claiming-george-floyd-death-was
- [49] Jack Clark, Import AI, 2020: https://jack-clark.net/2020/04/14/import-ai-193-facebook-simulates-itself-compete-to-make-more-efficient-nlp-face-in-painting-gets-better/
- [50] DARPA Media Forensics, 2019: https://www.youtube.com/watch?v=Crfm3vGoBsM
- [51] Zucconi Deepfakes Introduction, 2018: https://www.alanzucconi.com/2018/03/14/introduction-to-deepfakes/
- [52] The rise of GANs, Use Journal, 2019: https://blog.usejournal.com/the-rise-of-generative-adversarial-networks-be52d424e517
- [53] AI Deepfake GANs, Vox 2019: https://www.vox.com/future-perfect/2019/5/31/18645993/ai-deepfakes-gan-explained-machine-learning
- [54] Few-Shot Adversarial Training of Realistic Neural Talking Head Models, Samsung AI Center, Moscow, 2019: https://arxiv.org/abs/1912.04958
- [55] FaceSwap Github Marek Kowalski 2018: https://github.com/MarekKowalski/FaceSwap/
- [56] Face2Face: Real-time Face Capture and Reenactment of RGB Videos Stanford University 2018: https://niessnerlab.org/papers/2019/8facetoface/thies2018face.pdf
- [57] The Emergence of Deepfake Technology: A Review — Westerlund, 2019: https://timreview.ca/article/1282
- [58] The Guardian, 2017: https://www.theguardian.com/film/2017/jan/16/rogue-one-vfx-jon-knoll-peter-cushing-ethics-of-digital-resurrections

References.

- [59] Foreign Policy, 2020: https://foreignpolicy.com/2020/01/21/vladimir-putin-wants-to-rewrite-the-history-of-world-war-ii/
- [60] Vox, 2020: https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-Chechnya
- [61] Techcrunch, 2019: https://techcrunch.com/2019/07/04/an-optimistic-view-of-deepfakes/
- [62] Open AI, 2020: https://openai.com/blog/jukebox/
- [63] Road to VR, 2019: https://www.roadtovr.com/facebook-expands-on-hyper-realistic-virtual-avatar-research/
- [64]: FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces, 2018: https://arxiv.org/pdf/1803.09179.pdf
- [65] Radioactive data: Tracing through Training: https://arxiv.org/pdf/2002.00937.pdf
- [66] Detecting Audio Deepfakes With AI, Dessa, 2019: https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35
- [67] Reuters, 2019: https://www.reuters.com/manipulatedmedia
- [68] Bellingcat, 2020: https://www.bellingcat.com/resources/2020/03/27/investigating-coronavirus-fakes-and-disinfo-here-are-some-tools-for-you/
- [69] Worldwide Threat Assessment — U.S Intelligence Community, 2019: https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf
- [70] Twitter, 2020: https://twitter.com/davidfrum/status/1254613618911559682
- [71] Wired, 2017: https://www.wired.com/2017/02/veles-macedonia-fake-news/
- [72] Grover: Allen Institute for AI, 2020: https://grover.allenai.org/
- [73] U.S Department of Justice: Report On The Investigation Into Russian Interference In The 2016 Presidential Election, 2019: https://www.justice.gov/storage/report.pdf
- [74] Viriginia Legislative System, 2019: https://lis.virginia.gov/cgi-bin/legp604.exe?191+ful+HB2678S1&191+ful+HB2678S1
- [75] Texas Gov, 2019: https://capitol.texas.gov/tlodocs/86R/billtext/html/SB00751S.htm
- [76] NDAA Congress.gov, 2019: https://www.congress.gov/bill/116th-congress/house-bill/2500
- [77] HR4355 Congress.gov, 2019: https://www.congress.gov/bill/116th-congress/house-bill/4355

References.

- [78] Deepfake Report Act, Congress.gov, 2019: https://www.congress.gov/bill/116th-congress/senate-bill/2065
- [79] AB-730 California Legislative Information, 2019 https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730
- [80] AB-602, California Legislative Information: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602
- [81] FARSAIT, NIST, 2019 https://www.nist.gov/topics/artificial-intelligence/ai-research
- [82] Deepfake Accountability Act, Congress.gov, 2019: https://www.congress.gov/bill/116th-congress/house-bill/3230
- [83] Defense.gov, 2020: https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/
- [84] HR6088 Congress.gov, 2020: https://www.congress.gov/bill/116th-congress/house-bill/6088/text
- [85] European Commission, 2018: https://ec.europa.eu/digital-single-market/en/news/communication-tackling-online-disinformation-european-approach
- [86] European Commission, 2018: https://ec.europa.eu/digital-single-market/en/news/action-plan-against-disinformation
- [87] South China Morning Post, 2019: https://www.scmp.com/tech/apps-social/article/3039978/china-issues-new-rules-clamp-down-deepfake-technologies-used
- [88] Twitter, 2019: https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html
- [89] Verge, 2020: https://www.theverge.com/2020/1/9/21058803/reddit-account-ban-impersonation-policy-deepfakes-satire-rules
- [90] Verge, 2020: https://www.theverge.com/2020/1/7/21054504/facebook-instagram-deepfake-ban-videos-nancy-pelosi-congress
- [91] TikTok, 2020: https://www.tiktok.com/community-guidelines?lang=en
- [92] YouTube, 2020: https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html
- [93] Twitter, 2020: https://help.twitter.com/en/rules-and-policies/manipulated-media
- [94] Twitter, 2020: https://twitter.com/ChrisGuilhou/status/1276914961604849674
- [95] REFACE, 2020: https://www.linkedin.com/posts/activity-6704272701571289088-hOwE/
- [96] DARPA, 2020: https://youtu.be/IhAfHSEuqxk
- [97] Business Insider, 2020: https://www.businessinsider.com/ai-deepfake-apps-memes-misinformation-2020-9

References - Figures.

- [Figure 1, Figure 2]: FamousBoard Forum, MrDeepfakes Forums, Telegram Chat Deepfake communities. To prevent growth we have limited disclosing the group URLs, further information please contact Sentinel.
- [Figure 3]: Tinus Talks, 2019: https://tinius.com/2019/02/05/the-journalist-a-target-in-social-media/
- [Figure 4]: Vide, 2019: https://www.vice.com/en_au/article/gy4p47/how-it-feels-to-find-your-face-photoshopped-onto-internet-porn
- [Figure 5]: Designtaxi, 2018: https://designtaxi.com/news/398378/Watch-Deepfake-AI-Eerily-Recreates-A-Convincing-Version-Of-Donald-Trump/
- [Figure 6]: CVPR 2016 Paper Video, 2016 https://www.youtube.com/watch?v=ohmajJTcpNk
- [Figure 7]: Neural Voice Puppetry, 2019 https://arxiv.org/pdf/1912.05566.pdf
- [Figure 8]: CNN Business, 2019: https://www.youtube.com/watch?v=wCZSMIwOG-o
- [Figure 9]: Caroline Chan, 2018: https://www.youtube.com/watch?v=PCBTZh41Ris
- [Figure 10]: Logos sourced from: https://www.resemble.ai/ ; https://www.descript.com/ ; https://www.baidu.com/
- [Figure 11]: Analyzing and Improving Image Quality of StyleGAN, Nvidia, 2019: https://arxiv.org/abs/1912.04958
- [Figure 12]: Few-Shot Adverserial Training of Realistic Neural Talking Head Models, Samsung AI Center, Moscow, 2019; Analyzing and Improving Image Quality of StyleGAN, Nvidia, 2019: https://arxiv.org/pdf/1905.08233v1.pdf
- [Figure 13]: Defending against neural fake news, University of Washington, 2019: https://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf
- [Figure 14]: Politico, 2019: https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/
- [Figure 15]: Mother Jones, 2019: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/
- [Figure 16]: Quint Neon, 2018: https://www.youtube.com/watch?v=YwCa2Qxjxq8
- [Figure 17]: MOJO STORY, 2020: https://www.youtube.com/watch?v=4vB4VL6pfHk
- [Figure 18]: Guardian News, 2019: https://www.youtube.com/watch?v=VYKbI3eZ3i0

References - Figures.

- [Figure 19]: The Brussels Times, 2020: https://www.brusselstimes.com/all-news/belgium-all-news/politics/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-covid-19-with-climate-crisis/
- [Figure 20]: Twitter, 2020, @mooncult: https://twitter.com/mooncult/status/1212549854016106496?lang=en
- [Figure 21]: CNN, 2019: https://www.youtube.com/watch?v=vm_rjs9fyQk
- [Figure 22]: Aviation24, 2018: https://www.youtube.com/watch?v=RyeTMQn6EIg
- [Figure 23]: Avatarify, 2020: https://www.youtube.com/watch?v=lONuXGNqLO0
- [Figure 24]: Graphika: Sekondary Infektion, 2020: https://secondaryinfektion.org/downloads/secondary-infektion-report.pdf
- [Figure 25]: New York Times, 2020: https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html
- [Figure 26]: DARPA Media Forensics, 2019: https://www.youtube.com/watch?v=Crfm3vGoBsM
- [Figure 27]: Designtaxi, 2018: https://designtaxi.com/news/398378/Watch-Deepfake-AI-Eerily-Recreates-A-Convincing-Version-Of-Donald-Trump/
- [Figure 28]: Rockicon, Galaxicon, Aldric Rodriguez, Gregor Cresnar, thenounproject.com Creative Commons
- [Figure 29]: Designtaxi, 2018: https://designtaxi.com/news/398378/Watch-Deepfake-AI-Eerily-Recreates-A-Convincing-Version-Of-Donald-Trump/ ; Alfredo, Adrien Coquet, thenounproject.com Creative Commons
- [Figure 30]: Few-Shot Adverserial Training of Realistic Neural Talking Head Models, Samsung AI Center, Moscow, 2019: https://arxiv.org/abs/1912.04958
- [Figure 31]: FaceSwap Github Marek Kowalski 2018: https://github.com/MarekKowalski/FaceSwap/
- [Figure 32]: Face2Face: Real-time Face Capture and Reenactment of RGB Videos Stanford University 2018: https://niessnerlab.org/papers/2019/8facetoface/thies2018face.pdf
- [Figure 33]: The Guardian, 2017: https://www.theguardian.com/film/2017/jan/16/rogue-one-vfx-jon-knoll-peter-cushing-ethics-of-digital-resurrections
- [Figure 34]: Vox, 2020: https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-Chechnya

References - Figures.

- [Figure 35]: DARPA Media Forensics, 2019: https://www.youtube.com/watch?v=Crfm3vGoBsM
- [Figure 36]: FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces, 2018: https://arxiv.org/pdf/1803.09179.pdf
- [Figure 37, 38, 39, 40]: DARPA Media Forensics, 2019: https://www.youtube.com/watch?v=Crfm3vGoBsM
- [Figure 41]: Radioactive data: Tracing through Training: https://arxiv.org/pdf/2002.00937.pdf
- [Figure 42]: Detecting Audio Deepfakes With AI, Dessa, 2019: https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35
- [Figure 43]: Generated using https://breakyourownnews.com/
- [Figure 44]: Generated using https://faketrumptweet.com/
- [Figure 45]: Generated using https://www.worldgreynews.com/add-news
- [Figure 46]: Generated using https://breakyourownnews.com/
- [Figure 47]: Shutterstock: https://image.shutterstock.com/image-photo/vintage-radio-isolated-on-white-260nw-36604000.jpg ; SubPNG: https://www.subpng.com/png-ql51k8/ ; Pinterest: https://www.pinterest.com/pin/349451252308008261/
- [Figure 48]: Twitter: https://twitter.com/ChrisGuilhou/status/1276914961604849674
- [Figure 49]: LinkedIn: https://www.linkedin.com/posts/activity-6704272701571289088-hOwE/
- [Figure 50]: Youtube: https://www.youtube.com/watch?v=IhAfHSEuqxk&list=PLx2Zn7hPXT7fiCsXXWltQL8QltU09GVPk

https://thesentinel.ai