DEEPTRACE

THE
STATE
OF
DEEPFAKES

LANDSCAPE, THREATS, AND IMPACT

About Deeptrace

Deeptrace is an Amsterdam-based cybersecurity company providing deep learning and computer vision technologies for the detection and online monitoring of synthetic media. Our mission is to protect individuals and organizations from the damaging impacts of AI-generated synthetic media.

Deeptrace also publishes Tracer, a curated weekly newsletter covering key developments with deepfakes, synthetic media, and emerging cyber threats.

Acknowledgements

Table of Contents

Foreword

The rise of synthetic media and deepfakes is forcing us towards an important and unsettling realization: our historical belief that video and audio are reliable records of reality is no longer tenable.

I choose the word "historical" as this belief is a coincidence of how technology has evolved. Still today, we trust a phone call from a friend or a video clip featuring a known politician, simply based on the recognition of their voices and faces. Previously, no commonly available technology could have synthetically created this media with comparable realism, so we treated it as authentic by definition. With the development of synthetic media and deepfakes, this is no longer the case. Every digital communication channel our society is built upon, whether that be audio, video, or even text, is at risk of being subverted.

Since its foundation in 2018, Deeptrace has been dedicated to researching deepfakes' evolving capabilities and threats, providing crucial intelligence for enhancing our detection technology. In this report we share the insights from our most comprehensive mapping of the deepfake landscape to date, revealing deepfakes' real-world impact. In doing so, we provide an expert overview of the current state of deepfakes, cutting through some of the hyperbole surrounding the topic. The findings we present here are grounded in independently sourced data, accompanied by insights from leading experts in this area.

Our research revealed that the deepfake phenomenon is growing rapidly online, with the number of deepfake videos almost doubling over the last seven months to 14,678. This increase is supported by the growing commodification of tools and services that lower the barrier for non-experts to create deepfakes. Perhaps unsurprisingly, we observed a significant contribution to the creation and use of synthetic media tools from web users in China and South Korea, despite the totality of our sources coming from the English-speaking Internet.

Another key trend we identified is the prominence of non-consensual deepfake pornography, which accounted for 96% of the total deepfake videos online. We also found that the top four websites dedicated to deepfake pornography received more than 134 million views on videos targeting hundreds of female celebrities worldwide. This significant viewership demonstrates a market for websites creating and hosting deepfake pornography, a trend that will continue to grow unless decisive action is taken.

Deepfakes are also making a significant impact on the political sphere. Two landmark cases from Gabon and Malaysia that received minimal Western media coverage saw deepfakes linked to an alleged government cover-up and a political smear campaign. One of these cases was related to an attempted military coup, while the other continues to threaten a high-profile politician with imprisonment. Seen together, these examples are possibly the most powerful indications of how deepfakes are already destabilizing political processes. Without defensive countermeasures, the integrity of democracies around the world are at risk.

Outside of politics, the weaponization of deepfakes and synthetic media is influencing the cybersecurity landscape, enhancing traditional cyber threats and enabling entirely new attack vectors. Notably, 2019 saw reports of cases where synthetic voice audio and images of non-existent, synthetic people were used to enhance social engineering against businesses and governments.

Deepfakes are here to stay, and their impact is already being felt on a global scale. We hope this report stimulates further discussion on the topic, and emphasizes the importance of developing a range of countermeasures to protect individuals and organizations from the harmful applications of deepfakes.

Giorgio Patrini
Founder, CEO, and Chief Scientist

The State of Deepfakes: An Overview

Since emerging in late 2017 the phenomenon of deepfakes has developed rapidly, both in terms of technological sophistication and societal impact. In this section, we provide an overview of the current state of deepfakes by analyzing their proliferation and content.[1]

Total number of deepfake videos online

14,678

Our findings revealed that the total number of deepfake videos online is rapidly increasing, with this measurement representing an almost 100% increase based on our previous measurement (7,964) taken in December 2018.

percentage of deepfake videos online by **pornographic** and **non-pornographic** content

96%

4%
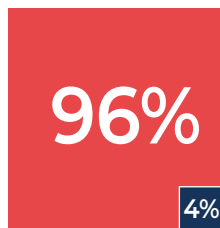
Deepfake pornography accounts for a significant majority of deepfake videos online, even as other forms of non-pornographic deepfakes have gained popularity.

Total number of video views across top four dedicated deepfake pornography websites

134,364,438

Despite being a relatively new phenomenon (the earliest of these websites was registered in February 2018), deepfake pornography has already attracted a large viewership on the top four dedicated deepfake pornography websites alone.

---

1   The appendix details how we collected the data presented in this report

Demographic breakdown of deepfake videos from top five deepfake pornography websites and top 14 deepfake YouTube channels

We analyzed the gender, nationality, and profession of subjects in deepfake videos from the top 5 deepfake pornography websites, as well as the top 14 deepfake YouTube channels that host non-pornographic deepfake videos.

Gender

female
100%
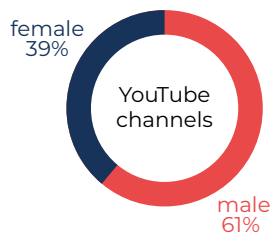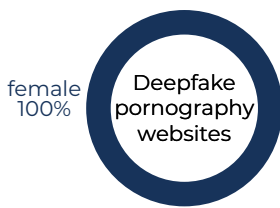
Deepfake pornography websites

female
39%

YouTube channels

male
61%

Deepfake pornography is a phenomenon that exclusively targets and harms women. In contrast, the non-pornographic deepfake videos we analyzed on YouTube contained a majority of male subjects.

Nationality

other
13%

India
3%

Canada
6%

UK
12%

Deepfake pornography websites

USA
41%

South Korea
25%

Taiwan
2%

South Korea
2%

other
5%

Canada
4%

YouTube channels

USA
73%

UK
14%
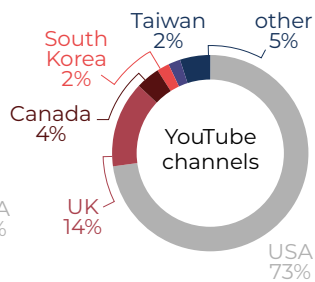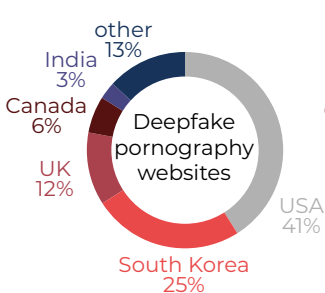
We found that over 90% of deepfake videos on YouTube featured Western subjects. However, non-Western subjects featured in almost a third of videos on deepfake pornography websites, with South Korean K-pop singers making up a quarter of the subjects targeted. This indicates that deepfake pornography is an increasingly global phenomenon.

Profession

news & media
1%

Deepfake pornography websites

entertainment
99%

business owner
2%

politician
12%

news & media
5%

YouTube channels
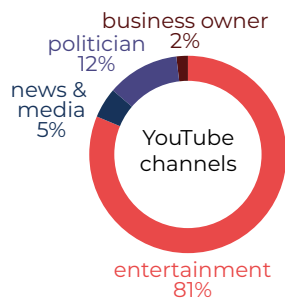
entertainment
81%

All but 1% of the subjects featured in deepfake pornography videos were actresses and musicians working in the entertainment sector. However, subjects featuring in YouTube deepfake videos came from a more diverse range of professions, notably including politicians and corporate figures.

The Commodification of Deepfakes

"Based on the rate of AI progress, we can expect deepfakes to become better, cheaper, and easier to make over a relatively short period of time. Governments should invest in developing technology assessment and measurement capabilities to help them keep pace with broader AI development, and to help them better prepare for the impacts of technologies like this."[2]

Jack Clark, Policy Director of OpenAI, and author of *Import AI*

The term deepfake was first coined by the Reddit user u/deepfakes, who created a Reddit forum of the same name on November 2nd 2017. This forum was dedicated to the creation and use of deep learning software for synthetically faceswapping female celebrities into pornographic videos. Since Reddit's removal of /r/Deepfakes on February 7th 2018, deepfakes have become increasingly commodified as new deepfake forums, tools, and services have emerged.

The growth of GANs

Synthetic audiovisual media can be generated with a variety of deep learning techniques. Currently, the most popular of these techniques is the Generative Adversarial Network (GAN) due to its flexible applications and realistic outputs.

The graph below measures the research output on GANs from their invention in 2014.

This provides an indirect indication of how fast GAN quality is improving and the rate that new capabilities for real-world applications are being developed. In the context of this section, it also indicates how this research is being incorporated into technology accessible to a wider audience outside academia.

Academic papers mentioning GANS in title/abstract published per year on the arXiv[3]

Footnotes:
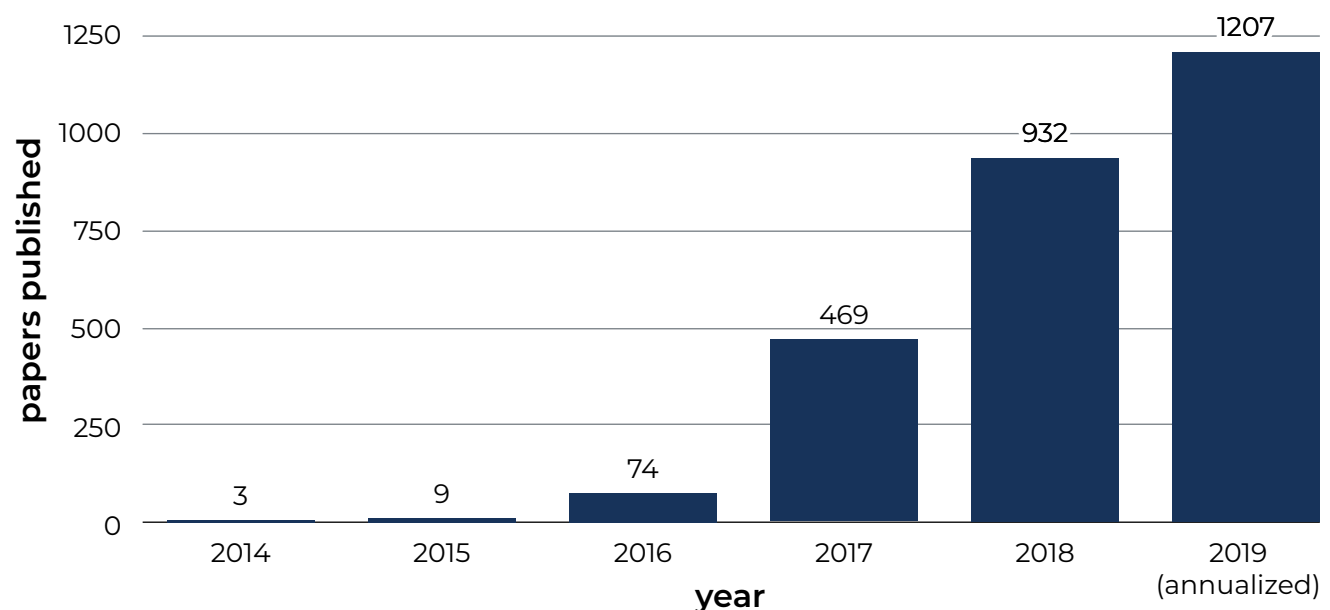
2   Clark, Jack, private correspondence, September 23 2019
3   arXiv is an online archive of scientific and mathematical research papers maintained by Cornell University.

Deepfake Creation Communities & Forums

Deepfake creation communities and forums are a key driving force behind the increasing accessibility of deepfakes and deepfake creation software. Many of these creation communities and forums provide an entry point for people interested in creating deepfakes, and facilitate collaboration between more experienced creators.

20 deepfake creation community websites and forums

We found the majority of creation communities and forums were located on deepfake pornography websites and established forum-based websites including Reddit, 4chan, 8chan, and Voat. Some of these websites such as 4chan and 8chan are notorious for hosting illegal and unethical activity.

non-unique members (from sources that disclosed membership numbers) 95,791

13 deepfake creation communities and forums disclosed their total membership, with these accounting for almost 100,000 members without the other seven whose membership was not accessible.

GitHub Stars

Open-source code repositories are the backbone of modern machine learning projects, and deepfake creation tools are no exception. The faceswap source code from the anonymous /r/Deepfakes creator was donated to the open-source community and uploaded on GitHub. Since then, development has been driven by many project forks, with programmers contributing to improved quality, efficiency, and usability of these new code libraries.

The chart below measures the popularity of the two best maintained deep learning faceswap projects (project names omitted) by "GitHub stars". Both projects show an organic growth in popularity since their initial upload, with faceswap project A's 20,000 stars rivaling the popularity of other industrial-level open-source projects.
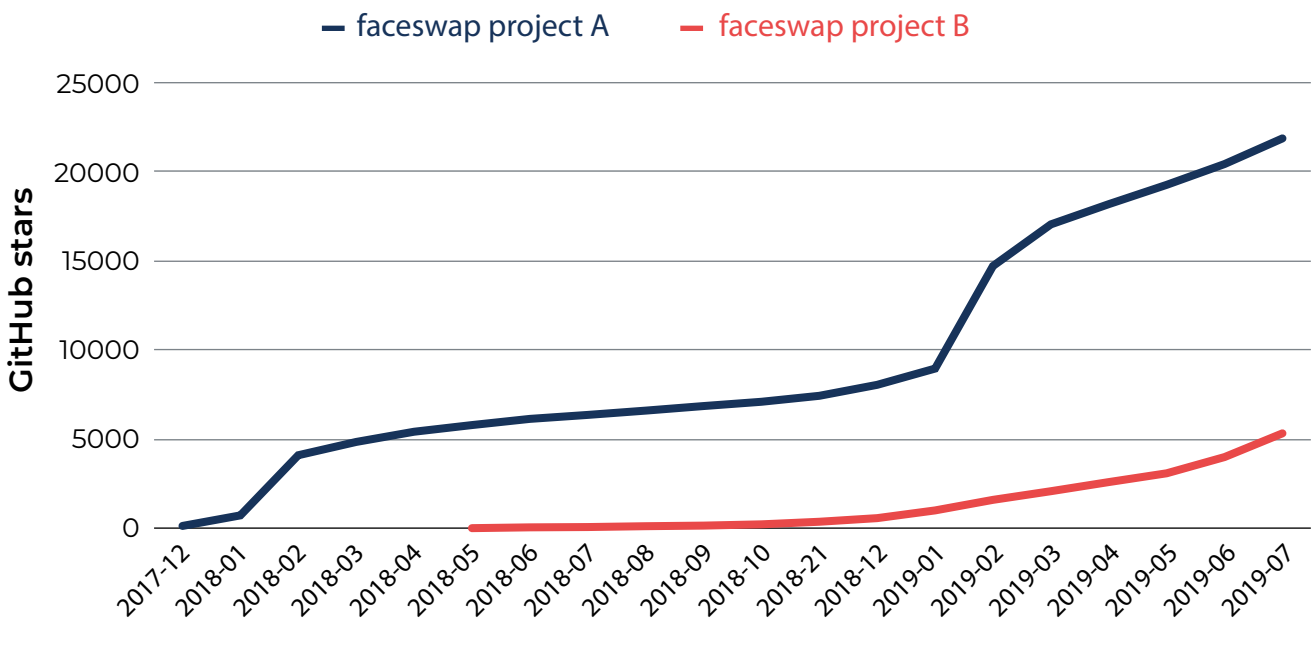
Deepfake Computer Apps & Services

Beyond research and code libraries, commodification is also about the growing accessibility of deepfake creation tools for a non-technical audience. Here we outline the three types of deepfake tools we identified during our research, with each providing a different approach to the creation and distribution of deepfakes.
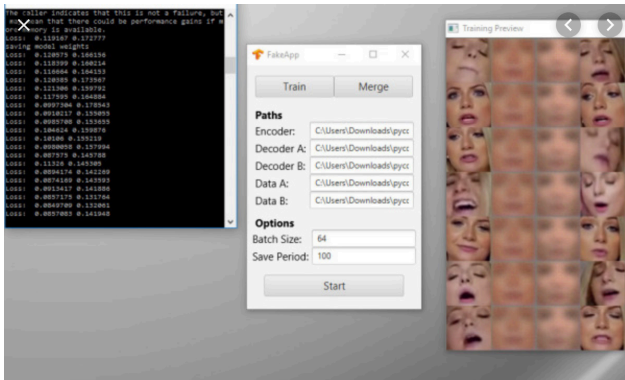
Computer Apps

Deepfake computer apps are downloadable tools or graphical user interfaces (GUIs) that are used for creating deepfakes. Most of these apps provide 'faceswapping' capabilities, but we also identified one app for synthetic voice cloning. Many of these apps require knowledge of programming and a powerful graphics processor to operate effectively, making them unsuitable for amateurs. However, several detailed tutorials have been created that provide step-by-step guides for using the most popular deepfake apps, and recent updates have improved the accessibility of several GUIs.
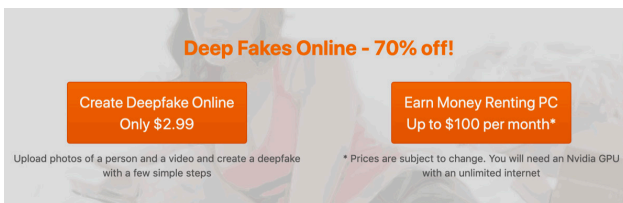
Service Portals

Service portals function as online businesses for generating and selling custom deepfakes. In most cases, this involves an individual uploading photos or videos of their chosen subjects to the online platform and receiving the deepfake once it has been generated. On one service portal we identified, this required 250 photos of the target subject and two days of processing to generate the deepfake. The prices of these services vary depending on the quality and duration of the video requested, but can cost as little as $2.99 per deepfake video generated.
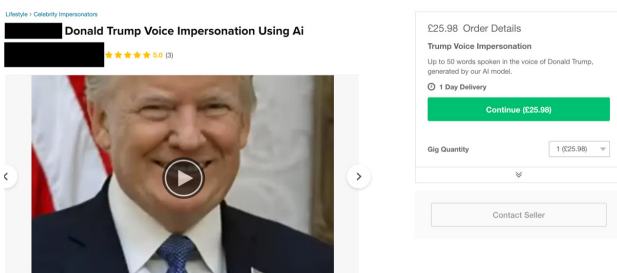
Marketplace Services

Marketplace services refer to individual deepfake creators who advertise their services on forums and online marketplaces. Our research found marketplace services ranging from bespoke faceswap videos for $30 to custom voice cloning for $10 per 50 words generated.

Deepfake Pornography

"Deepfake technology is being weaponized against women by inserting their faces into porn. It is terrifying, embarrassing, demeaning, and silencing. Deepfake sex videos say to individuals that their bodies are not their own and can make it difficult to stay online, get or keep a job, and feel safe."[4]

Danielle Citron, Professor of Law, Boston University, and author of *Hate Crimes in Cyberspace*

Deepfake pornography is by far the most prevalent kind of deepfake currently being created and circulated. The damaging effects of deepfake pornography have already affected a significant number of women, including celebrities and private individuals. In this section, we outline the deepfake pornography ecosystem and present a case study on the development of the deepfake pornography creation app DeepNude.

The rise of deepfake pornography

8 of the top 10 pornography websites host deepfakes

9 dedicated deepfake pornography websites

mainsteam pornography websites 6%

deepfake pornography videos by host website type

dedicated deepfake pornography websites 94%
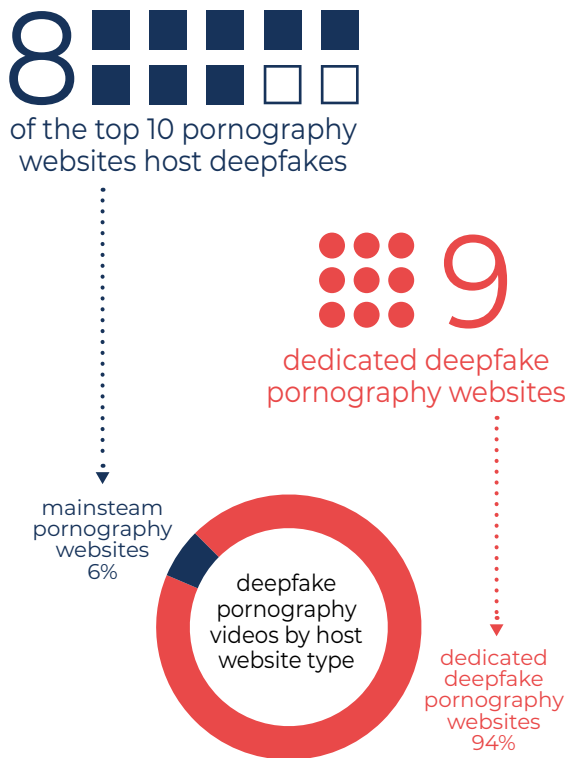
We measured the online presence of deepfake pornography by tracking the number of videos that have been uploaded, and the number of websites that hosted these videos.

Our research revealed that deepfake pornography has a significant online presence across several different websites. These websites fell into two categories: dedicated deepfake pornography websites and mainstream pornography websites.

We found that the deepfake pornography ecosystem is almost entirely supported by dedicated deepfake pornography websites, which host 13,254 of the total videos we discovered. By contrast, mainstream pornography websites only hosted 802 videos.

The number of these dedicated deepfake pornography websites suggests that deepfake pornography could represent a growing business opportunity, with all these websites featuring some form of advertising.

---

4   Citron, Danielle, private correspondence, August 16 2019.

10 individuals most frequently targeted by deepfake pornography

We chose not to publish the names of these women given the sensitive nature of the data, but rather specify their nationality and profession. Of the 10 most frequently targeted individuals we identified, the majority were actresses from Western countries. However, the second and third most frequently targeted individuals, as well as the most frequently viewed individual, were South Korean K-pop singers.

millions of views / **video features**

millions of views		video features
5.8	British actress	257
3.1	South Korean musician	167
5.9	South Korean musician	134
4.3	American actress	126
1.6	British actress	118
1.4	Australian actress	113
2.4	American actress	106
1.9	Israeli actress	105
1.1	American actress	104
4.4	South Korean musician	103
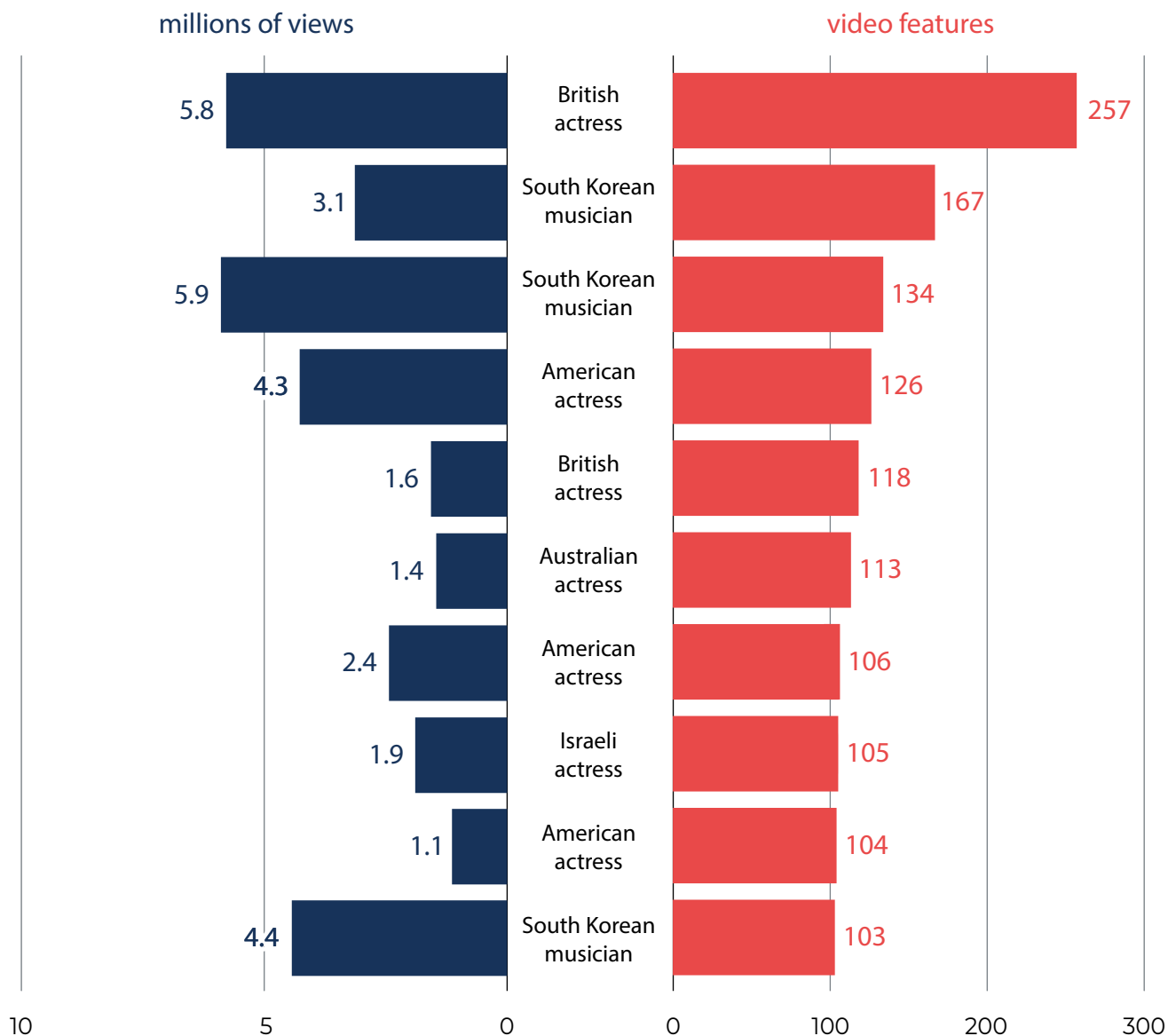
A Case Study: DeepNude

DeepNude is a computer app that enables users to 'strip' photos of clothed women. The app uses deep learning image translation algorithms that have been tuned to synthetically remove clothes from images of women, and generate naked parts of their body that were previously covered. These algorithms cannot perform similar translations on images of men, having been specifically trained on images of women. DeepNude's release illustrates the dangers of automated tools for creating deepfake pornography becoming increasingly accessible.

Launch

DeepNude's website launched on June 23rd 2019, featuring a demo and download options for Windows and Linux computer apps. The app could be downloaded for free, but generated images were obscured by a large watermark. To remove this watermark, users had to pay $50 for the licensed version. Images generated with this licensed version still contained a small watermark designating the image as "fake". Each image took the app 30 seconds to generate.

Takedown

DeepNude saw a surge in interest following critical coverage by journalist Samantha Cole for the online publication, Motherboard on June 27th[5]. This overwhelmed DeepNude's servers with traffic and download requests, causing the website to go offline less than 24 hours after Cole's article was published. According to DeepNude, the website received 545,162 visits and 95,464 active users in June, with the majority of this activity likely occurring in this 24 hour period. The creators subsequently stated they had "greatly underestimated [the volume of download] requests" and that they would not release any further versions of the app, claiming that "the world is not ready for DeepNude".
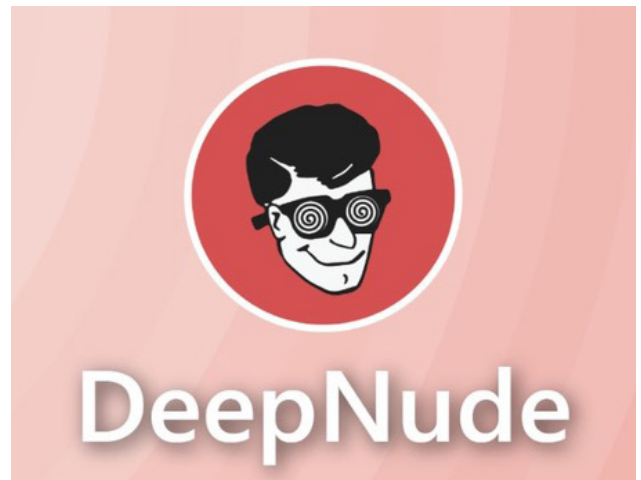
Resurrection

Despite the creators of DeepNude taking the official website offline, the software continues to be independently repackaged and distributed through various online channels, such as open-source repositories and torrenting websites. In addition, two new service portals opened offering an allegedly improved version of DeepNude, with charges ranging from $1 per photo to $20 for a month's unlimited access. Likely having recognized the app's business potential, the original creators put DeepNude up for sale on July 19th for $30,000 via an online business marketplace, where it sold to an anonymous buyer.

Losing control

The moment DeepNude was made available to download it was out of the creators' control, and is now highly difficult to remove from circulation. The software will likely continue to spread and mutate like a virus, making a popular tool for creating non-consensual deepfake pornography of women easily accessible and difficult to counter.

---

5   Cole, Samantha. "This App Uses Deepfake Technology Solely to Objectify Women" Motherboard. June 27 2019.

Politics & Cybersecurity

"We've already seen so-called "shallowfakes" circulated online with the attempt to distort political discourse or delegitimize politicians, and even selective editing has been used to falsely represent how a politically-charged event occurred. High-quality AI-manipulated video ups the stakes considerably. It is only a matter of time before deepfakes are used in an attempt to manipulate elections."[6]

Paul Scharre, <u>Director of Technology & National Security, CNAS</u>

Much of the media attention surrounding deepfakes has focused on their potential to undermine democratic processes and enhance cyberattacks against individuals and businesses. In this section, we outline several cases where deepfakes and the related phenomenon of shallowfakes have had a disruptive impact on the political sphere. We also detail cases where synthetic images and synthetic voice audio have allegedly been used to enhance social engineering and fraud.

You Won't Believe What Obama Says In This Video! 😉

6,719,143 views • Published on Apr 17, 2018   👍 82K   👎 14K   ➔ SHARE   ≡+ SAVE   ...

Buzzfeed News & Monkeypaw Studios (2018)

---

6   Scharre, Paul, private correspondence, August 8 2019.

The destabilizing political impact of deepfakes

We have provided technical consultation on two cases where deepfakes have played a significant role in creating or exacerbating political crises. Both are powerful examples of how growing awareness of deepfakes is already damaging political discourse, and undermining the perceived objectivity of recorded video in politically charged contexts.

Gabon: Accelerating political unrest

In late 2018 there was intense speculation about the health of the Gabonese President Ali Bongo, who had been absent from public life for several months. In an attempt to end the speculation, the government released a video of Bongo delivering a traditional New Year's address.

However, Bongo's unusual appearance in the video led many on social media, including Gabonese politician Bruno Moubamba, to declare that the video was a deepfake, confirming their suspicion that the government was covering up Bongo's ill health or death. A week after the video's release amid growing unrest, members of Gabon's military launched an attempted coup against the government. In a video announcing the coup, the military mentioned the video's odd appearance as proof that something was wrong with the President.[7] Subsequent forensic analysis

Mother Jones (2019)

did not find signs that the video had been manipulated, and Bongo has since made public appearances following what the government stated was a severe stroke in August 2018.

Malaysia: Sex scandal

A political scandal emerged in June 2019 surrounding a sex tape allegedly featuring the Malaysian Minister of Economic Affairs Azmin Ali and a rival minister's male aide.[8] Same-sex sexual activity is illegal in Malaysia, with politicians having previously been controversially imprisoned on sodomy charges. While the aide claimed the video was real and was subsequently arrested, Ali and his supporters, including the Malaysian Prime Minister, argued that the video was a realistic deepfake made to sabotage his political career. However, international experts could not find any signs that the video had been manipulated. As of mid-August 2019, it is still unclear whether Ali will face criminal charges over the video.

The Straits Times (2019)

---

7   Breland, Ali. "The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink" Mother Jones. March 15, 2019.

8   Blakkarly, Jarni. "A gay sex tape is threatening to end the political careers of two men in Malaysia" SBS News. June 17 2019.

Shallowfakes

The term Shallowfakes was first coined by Sam Gregory of the human rights organization WITNESS, and refers to videos that have been manipulated with basic editing tools or intentionally placed out of context.[9] The Washington Post's guide to manipulated video provides three specific categories for these kinds of shallowfake manipulations:[10]

› Missing context – misrepresenting a video or isolating it from its original context;

› Deceptive editing – omitting passages of a video or inserting new ones;

› Malicious transformation – doctoring or manipulating a video's content.

Shallowfakes are typically designed to exploit an individual's cognitive biases which can result in damage to a target person's reputation even if the fake is of a low quality.

Nancy Pelosi: Manipulated voice audio

A recent high profile shallowfake involved a manipulated video of US Speaker of the House and Democrat Congresswoman Nancy Pelosi.[11] In the video that was shared on May 23rd 2019, Pelosi's speech had been slowed down, making it sound like she was slurring her words. The edited version of the video went viral on social media and was retweeted by the official Twitter account of US President Trump, receiving over 6.3m views as of July 31st 2019.[12] On a popular Facebook page, the video received over 2.2m views in the 48 hours following its initial upload, with commenters calling Pelosi "drunk" and a "babbling mess".[13]

New York Times (2019)

The Washington Post (2018)

Jim Acosta: Manipulated body movement

Shallowfakes have also been falsely cited as evidence to justify controversial political actions. In this case, CNN's correspondent Jim Acosta had his White House press pass revoked on November 7th 2018, following an interaction where an intern attempted to take a microphone from his hand after a tense exchange with the President. In edited footage of the event shared

---

9   Gregory, Sam. "Deepfakes will challenge public trust in what's real. Here's how to defuse them." Defusing Disinfo. February 19, 2019.

10   Ajaka, Nadine, Samuels, Elyse, and Kessler, Glen. "Seeing Isn't Believing The Fact Checker's guide to manipulated video" The Washington Post. June 25 2019.

11   Mervosh, Sarah. "Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump" The New York Times.   May 24 2019.

12   Monje Jr, Carlos. "Twitter letter to Chairman Schiff". Twitter July 31 2019.

13   Harwell, Drew. "'Sexist' videos edited to make Nancy Pelosi look drunk go viral, with Trump's help" The Independent. May 24 2019.

by the White House Press Secretary, the video was sped up, making it appear that Acosta had aggressively blocked the intern from taking the microphone from him with his arm.[14] This was cited as "inappropriate behavior" and the sole reason for his press pass being revoked. However, reporting by The Washington Post's Drew Harwell and other news organizations revealed that the video had been manipulated, leading to Acosta's press pass being reinstated roughly two weeks after the incident first occurred.

Political Satire & Art

The impact of deepfakes on the political sphere is not solely restricted to malicious and deceptive disinformation campaigns, but also includes the creation of new forms of political satire and art. One famous early example of this use of deepfakes was seen with Buzzfeed's and director Jordan Peele's "Deepfake Obama PSA", a deepfake of Barack Obama that warned of the dangers posed by the technology.[15] As awareness of deepfakes has grown, so have efforts to use them as part of political and social commentary.

Spectre

Spectre is a project by the art collective Brandalism that involves several deepfake videos of prominent public figures, including Mark Zuckerberg, Kim Kardashian, and Donald Trump.[16] Each deepfake synthetically synchronized the subject's lip movements to match a voice actor imitating their voice, speaking about how the mass collection of personal data and other technologies have brought them power or wealth.

Upon Spectre's release, the deepfakes had an immediate viral impact across a variety of social media platforms. However, the project did not present the deepfakes as real, but as an artistic device to provoke discussion of political and social issues.

bill_posters_uk
13.3k followers
View Profile

6,647 likes
bill_posters_uk

'Imagine this...' (2019) Mark Zuckerberg reveals the truth about Facebook and who really owns the future... see more @sheffdocfest VDR technology by @cannyai #spectreknows #privacy #democracy #surveillancecapitalism #dataism #deepfake #deepfakes #contemporaryartwork #digitalart #generativeart #newmediaart #codeart

---

14   Harwell, Drew. "White House shares doctored video to support punishment of journalist Jim Acosta" The Washington Post. November 8 2018.
15   Newman, Craig. "How To Spot A Deepfake Like The Barack Obama–Jordan Peele Video" Buzzfeed News. April 17 2019.
16   Posters, Bill. "Gallery: Spectre launches (Press Release)". Bill Posters. May 29 2019.
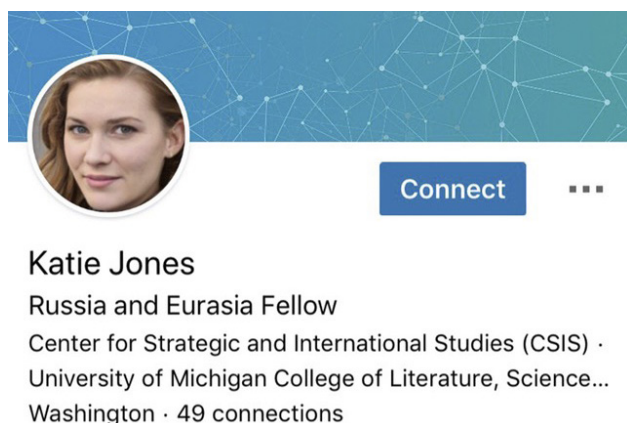
Cybersecurity

Aside from the politics, malicious applications of deepfakes and synthetic media are also changing the cybersecurity landscape by enhancing traditional cyberthreats and enabling entirely new attack vectors. In 2019 we have seen several cases where deepfakes have been deployed to enhance social engineering and fraud.

Enhancing fake digital identities: fraud, infiltration and espionage

We observed two cases where realistic synthetic photos of non-existent people were used on fake social media profiles, in an attempt to deceive other users and extract information.

Deceiving Tesla short sellers

In March 2019, accounts posing as a senior Bloomberg journalist "Maisy Kinsley" connected with 195 people on LinkedIn, and followed several Tesla short sellers on Twitter[17]. Some of these short sellers claimed that the account had contacted them in an attempt to extract personal information. The account's profile picture contained visual anomalies consistent with synthetically generated images. Both accounts were removed by LinkedIn and Twitter. Bloomberg later confirmed they did not employ anyone called Maisy Kinsley.

Maisy Kinsley
@msmaisymade

Senior Journalist @ Bloomberg. I tell data-driven stories about cars, cyber security & web culture through words, graphics & passion. DM open for tips or email

San Francisco, CA

Katie Jones
Russia and Eurasia Fellow
Center for Strategic and International Studies (CSIS) ·
University of Michigan College of Literature, Science...
Washington · 49 connections

Espionage

A LinkedIn account posing as "Katie Jones", a researcher from a US think tank, was believed by experts to be part of a foreign spying operation.[18] The account made 52 connections, including government officials' members of staff. Expert analysis identified several visual anomalies that indicated that the image was synthetically generated. The account was quickly removed by LinkedIn.

---

17   Fleishman, Glenn. "How to spot the realistic fake people creeping into your timelines". Fast Company. April 30 2019.
18   Satter, Raphael."Experts: Spy used AI-generated face to connect with targets" AP News. June 13 2019.

Synthetic voice impersonation and fraud

There have been several reported cases where synthetic voice audio has allegedly been used to defraud companies.[19] While no concrete evidence has been provided to support claims that the audio was synthetic, the cases illustrate how synthetic voice cloning could be used to enhance existing fraud practices against businesses and individuals.

CEO impersonation

A report by French insurance company Euler Hermes stated that one of their clients, an unnamed UK energy firm, was the victim of a fraud attack using synthetic voice audio in March 2019[20]. According to Euler Hermes, the cybercriminals used synthetic voice audio that impersonated the CEO of the firm's German parent company, accurately replicating his accent and speaking style. Recognizing the voice and believing it to be real, the CEO of the British firm complied with a request to wire $243,000 to a Hungarian supplier. These funds were then moved from the Hungarian accounts to several other locations. Euler Hermes covered these losses as part of the company's fraud insurance policy. No further evidence such as recordings of the call have been released to prove the voice audio was synthetic, and at the time of writing the perpetrators have not been identified.

---

19   "Fake Voices 'help cyber-crooks steal cash'" BBC News. July 8 2019.
20   Stupp, Catherine. "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case" The Wall Street Journal. August 30 2019.

Concluding Remarks

Deepfakes pose a range of threats, many of which are no longer theoretical. In this report we have identified five key conclusions on the current state of deepfakes:

> › Deepfake creation technologies and tools are being commodified through a growing number of communities, computer apps, and services;

> › The online presence of deepfake videos is rapidly expanding, with the vast majority of these videos involving pornographic content;

> › Deepfake pornography is a global phenomenon supported by a significant viewership across several dedicated websites, with women exclusively being targeted;

> › Awareness of deepfakes alone is destabilizing political processes by undermining the perceived objectivity of videos featuring politicians and public figures;

> › Deepfakes are providing cybercriminals with new sophisticated capabilities to enhance social engineering and fraud.

These conclusions are drawn from our analysis of the deepfake landscape as it currently lies. However, the speed of the developments surrounding deepfakes means this landscape is constantly shifting, with rapidly materializing threats resulting in increased scale and impact. It is essential that we are prepared to face these new challenges. Now is the time to act.

Appendix

The quantitative data collection detailed here took place between June 1st 2019 and July 31st 2019. Examples and case studies were collected up until August 31st 2019.

Data Sensitivity

Open-source research on sensitive data is a delicate matter that we take very seriously. Providing methodological transparency on some of our sources would involve publicizing tools, services, and websites that are legally ambiguous or have explicitly unethical applications. Accordingly, we have decided to censor the names and URLs of these sources, particularly when related to deepfake pornography.

Mainstream pornography websites

Sources

www.Aelieve.com, 10 most popular mainstream pornography websites in terms of user traffic.

Methodology

We identified the 10 most popular porn websites in terms of web traffic, as indicated by Aelieve.com. We then used the internal site search engine with the keyword "deepfake(s)" and recorded the number of videos returned. One of these websites had disabled searches for the keyword "deepfake", but we were still able to detect several deepfake videos on this website by conducting a manual search.

Dedicated deepfake pornography websites

Sources

Google search, redirect links from deepfake communities and forums.

Methodology

We identified dedicated deepfake pornography websites using Google search with the key phrases "deepfake porn portal" and "deepfake porn website", along with redirect links we found on deepfake forums that were not indexed through Google searches. All dedicated deepfake pornography websites we identified solely contained fake videos (as well as the occasional fake image). Two of these websites also contained some videos that were possibly created without the use of generative algorithms or other AI-based tools, and therefore could be shallowfakes. Since the sites didn't provide any guidance on how this content was created, and in one case the content was not fully visible without buying premium access, we have decided to count these videos.

Excluding these two websites, we selected the top five in terms of content volume for additional data analysis. We used the web scraper Beautiful Soup to identify every hosted video and developed proprietary tools to extract data such as title, views, and likes. The total number of videos we analysed was 7,144. For one of these five websites information on views was not available. Accordingly, the statistics on views are reported for the top four only.

The compiled video metadata gave us results such as total viewership numbers but told us little about the people targeted by deepfake pornography videos. In some cases, we were able to extract the exact names of people present in videos from the page HTML. Unfortunately, in many cases, all we had was a title which usually contained the target's name as part of a broader description of the video. In cases where only a title was available, we used a combination of NLP processing, manual editing, and validation to extract the people named in the video titles. For this, we used the python library SpaCy, taking advantage of its named entity recognition capabilities. We were, therefore, able to determine the names of people targeted in almost every deepfake video we tracked.

Having identified the names of targeted individuals we used the Wikidata API to extract demographic information on each individual, including their gender, nationality, and profession. Wikidata offers a plain text query function that can identify well-known people even with alterations and errors in the spelling of their names as they appear in video titles.

Non-pornographic video hosting websites/channels

Sources

YouTube, Vimeo, LiveLeak, Dailymotion. Reddit was not taken into consideration as almost all deepfakes on Reddit are hosted on YouTube.

Methodology

Searching for the keyword "deepfake" on YouTube leads to inflated results, where search results output content on deepfakes but not actual deepfake videos. To avoid this pitfall we focused our search on outputs from dedicated deepfake channels or creators who exclusively or largely publish deepfake videos. The other sources host a much smaller volume of deepfakes videos, which we were able to find and count by a keyword "deepfake" search. Each video was manually checked to see if it contained deepfakes or deepfake style alterations.

For further analysis of YouTube videos, we downloaded the metadata of the videos hosted on 14 channels dedicated to deepfakes videos. In total, the number of videos was below 500. Post metadata extraction analysis based on videos' titles was identical to the analysis performed on pornographic deepfake metadata as described above.

Forums and creation communities

Sources

Google search, deepfake pornography websites, Reddit, Telegram, 4Chan, 8Chan, Voat

Methodology

We used the terms "forum" and "creation community" to refer to online communities where people can share, discuss, and find information on deepfakes, with a focus on the creation process. We conducted Google searches for "deepfake forums" and "deepfake communities", as well as specific searches on deepfake pornography websites and forum-based websites. From here, we calculated the number of registered users of 13 of the 20 deepfake forums and creation communities we identified, based on publicly accessible data on the related websites and channels.

Deepfakes computer apps

Sources

Google search, GitHub

Methodology

We use the term "deepfake computer app" to refer to downloadable software for creating deepfakes that are independently operated by the user. This can be further broken down into open-source code repositories and graphical user interfaces. We identified deepfake apps and GUIs with Google search using the keywords "deepfake apps" and "deepfake tools", and through GitHub's repository search engine. A voice cloning GUI was found on GitHub by searching for "voice cloning".

Deepfake services portals

Sources

Google search

Methodology

We use the term "deepfake service portal" to refer to an online business that generates and sells custom deepfakes. We identified these deepfake service portals with Google search, using the keywords "deepfake service" and "deepfake tools". Our analysis of each service portal focused on the kinds of deepfake creation services being offered, and the prices that accompanied these services.

Marketplace services

Sources

Google search, website search

Methodology

We used the term marketplace services to refer to individual deepfake creators that advertise their services on forums and online marketplaces. We identified these marketplace services using Google search with the phrase "deepfake services", "deepfake tools" or "synthetic voices", along with the search engines on three of the top online marketplaces and freelance work portals using the keywords "deepfakes" and "deepfake services".

GitHub

Sources

GitHub Archive stored on Google BigQuery

Methodology

The GitHub stars visual in the report shows the number of stars for two faceswap GitHub repositories over time. We used the following SQL query to collect this data:

SELECT project,
YEAR(star_date) as yearly,
MONTH(star_date) as monthly,
SUM(daily_stars) as monthly_stars
FROM (
SELECT repo.name as project, DATE(created_at) as star_date, COUNT(*) as daily_stars
FROM TABLE_DATE_RANGE(
[githubarchive:day.], TIMESTAMP("20171001"), TIMESTAMP("20190731"))
WHERE repo.name IN ("project/project_name") AND type = 'WatchEvent'
GROUP BY project, star_date)
GROUP BY project, yearly, monthly ORDER BY project, yearly, monthly

Papers mentioning GANs in the title/abstract

Sources

www.arXiv.org is an online archive of research articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. arXiv is owned and operated by Cornell University. See more information on arXiv.org.

Methodology

We used arXiv's search engine https://arxiv.org/find with the keyword "GAN" or "generative adversarial network" to identify the total papers containing these keywords in their titles or abstracts for each year. The annualized 2019 figure was derived from 704 papers released between January 1st 2019 and July 29th 2019, by multiplication by 12/7.

DeepNude

Sources

DeepNude websites (original and repackaged versions), articles (names omitted)

Methodology

The DeepNude business sale price and June 2019 website viewing figures/user numbers were directly sourced from a listing on the online business marketplace, which is no longer accessible at the time of this report's publication. We accessed this data on July 18th 2019. The details of the original app's pricing and functionality were sourced from DeepNude's website, which is also no longer accessible at the time of this report's publication. We accessed this data on June 27th 2019.