

The background of the entire page is a close-up photograph of the Selfridges building's iconic perforated metal facade, featuring a grid of circular holes.

REPORT 2018

**THE STATE OF
DEEPFAKES:
REALITY UNDER
ATTACK**

DEEPTRACE

ABOUT THIS REPORT

In recent years, rapid advances in AI have led to the technology's widespread commoditisation, with a variety of positive applications. However, these technologies are similarly being exploited by bad actors for malicious uses. This report focuses on one specific malicious application: the use of AI to generate harmful synthetic video, images, or audio, popularized under the broader name of 'deepfakes'.



This report is the first to trace the evolution of this technological and social phenomenon. The reference timeline is 2017-2018, with an outlook to 2019. Data is collected up to 20/12/2018.

About the authors DeepTrace B.V. is a Private Limited company founded in 2018 in Amsterdam, the Netherlands. DeepTrace's mission is to empower people in understanding and trusting what they see. The company is building technologies for fake videos detection and analytics. This document follows the company goal by studying the new phenomenon and its societal impact, with the aim of increasing public awareness and devising technological solutions to the threat.

GLOSSARY

Artificial Intelligence (AI): A branch of computer science that aims to replicate features of human intelligence in software.

Defense Advanced Research Projects

Agency (DARPA): An R&D agency of the United States Department of Defense.

Deepfake: A portmanteau of "deep learning" and "fake", deepfakes refer to any photo-realistic audiovisual content produced with the aid of deep learning. It also refers to the technology creating it. The term implies its misuse for illicit or unethical purposes.

The term is derived from the Reddit user "deepfakes", who was the first to publicly document attempts to synthetically replace the face of a target person with the face of another person, i.e. a face swap, in pornographic videos. Throughout 2018, the term expanded to include other techniques such as facial expression re-enactment, full body and background manipulation as well as audio synthesis.

In some cases, even videos doctored with elementary zoom, speedup or frames reordering have been referred to as deepfakes by the press. We do not encourage the use of the term in this way.

Deep learning: A class of AI methods based on deep neural networks for tasks such as supervised and unsupervised learning and generation.

Generative Adversarial Networks

(GANs): A kind of deep learning algorithm that trains a neural network to generate realistic imagery. Whilst GANs are not integral to creating synthetic media, they represent the most sophisticated recent development in the how new kinds of synthetic media are created.

KEY FINDINGS

- **2018 summary:** Originating from a Reddit community of developers, the deepfakes phenomenon features in top media outlets worried about its consequences for journalism, gaining world-wide popularity. The first instances of fake videos for political propaganda and public shaming take place. Several websites ban deepfakes but the lack of reliable technology for detection makes implementing this ban futile. As concern about deepfakes grew within the US intelligence community, a Chinese state-owned news broadcaster debuts with a synthetic 'news anchor', generated with AI.
- **Public awareness:** Online publication of articles and videos on deepfakes grew rapidly, with Google searches for deepfakes being 1000 x higher than in 2017.
- **Adult industry:** The first area that saw use of deepfakes at scale. We identified 8K+ fake pornographic videos on adult sites.
- **Tech development:** An active community of developers is growing around open source projects, building free and easy to use software tools for creating deepfakes.
- **Synthesis: quality and quantity:** Choosing GAN as a representative technique, we report on the striking increase of research papers published per year since 2014, as well as rapid improvements in synthetically generated image quality.
- **A Lack of Research on anti-forgery:** In 2018 we identified only 25 new research papers focused on detecting fake imagery, compared to 902 papers on GANs.
- **Outlook:** Expert opinion generally agreed that deepfakes are likely to have a high profile, potentially catastrophic impact on key events or individuals in the period 2019-2020.

TIMELINE

- 
- 2012** The ImageNet image classification competition is won by a significant margin with a deep neural network for the first time. The deep learning "Renaissance" begins.
 - 2014** Deep learning researchers make breakthroughs in generative modelling, in particular Ian Goodfellow's creation of the Generative Adversarial Network (GANs). Yann Lecun, Chief AI Scientist at Facebook will cite them as the "the coolest idea in deep learning in the last 20 years".
 - 2016** DARPA starts the 4 year Media Forensics (MediFor) program, developing new forensic methods for analysing digital media.
 - 2016** Computer vision researchers publish "Face2Face", a new method for facial re-enactment: it is possible to transfer facial expressions from one person to a realistic digital 'avatar' in real time.
 - 2017** UC Berkeley presents "Cycle GAN" that can transform images and videos into a version with a different style, e.g. different weather.
 - 2017** The University of Washington publishes "Synthesizing Obama: Learning Lip Sync from Audio": a method for synchronising lip movement in video with a speech from a different source.
 - 2017 Nov** The anonymous reddit user "deepfakes" launches a dedicated subreddit r/deepfakes for sharing porn videos where celebrities faces are swapped with the original ones. The code is built on popular open-source deep learning libraries. Many videos begin appearing on popular adult video sharing websites.
 - 2017 Dec** Motherboard (Vice) is the first to report about r/deepfakes. Throughout 2018, deepfakes are featured in hundreds of press articles, including mainstream outlets such as The New York Times, Washington Post, The Guardian, The Economist, The Times and The BBC.

TIMELINE

2018 Jan r/deepfakes gathers over 2500 subscribers

2018 Jan The FakeApp desktop application, built with Google's open-source library Tensorflow, is released along with tutorials for non-experts.

2018 Jan The first of many website offering a deepfake creation service is launched, funded by user donations.

2018 Jan-Feb Several websites, including Discord, Gfycat, Pornhub and Twitter, ban deepfakes with varying degrees of success.

2018 Feb r/deepfakes reaches 90K subscribers and is banned from Reddit for violating the policy "against involuntary pornography".

2018 Feb The first dedicated deepfake adult website launches.

2018 Mar Researchers from the Technical University of Munich and others introduce "FaceForensic", a large-scale video dataset for training media forensic and deepfake detection tools.

2018 Apr BuzzFeed tweets a fake video of Barack Obama played by Get Out-director Jordan Peele. The post receives more than 13K retweets, 29K likes and 100K views on Youtube in one day.



Credits: Buzzfeed 2018

TIMELINE

2018 Apr Indian journalist Rana Ayyub is targeted by deepfake pornography using her image and fake impersonation accounts on social media, with the intent to humiliate and defame.

2018 Apr A domestic violence advisory is published by US law professionals titled "Using fake video technology to perpetrate intimate partner abuse".

2018 May Researchers at Stanford University publish "Deep video portraits": the method enables photo-realistic re-animation of portrait videos, using only an input video.

2018 May The Belgium Flemish Socialist Party posts a fake video of Donald Trump in which he calls on the country to exit the Paris climate agreement. Officially, the video was made "to draw attention to the necessity to act on climate change". Many commenters on Facebook do not realise that the video was fake.

2018 May US Senator Marco Rubio voices his concerns about deepfakes at the Senate Intelligence Committee nomination hearing.

2018 Jun The University of Albany SUNY presents "In Ictu Oculi": a paper proposing to spot deepfakes based on eyes blinking patterns.

2018 Aug UC Berkeley publishes "Everybody dance now": a method for transferring body movements of a person in a source video to another person in a target video.

2018 Sep Three US lawmakers send a letter to the Director of National Intelligence demanding a report on deepfakes by mid-December 2018. The letter's authors argue deepfakes pose a national security threat, highly accessible to hostile foreign nations.

TIMELINE



Credits: Xinhua News Agency 2018

2018 Sep NGO AI Foundation raises \$10M of capital to develop a tool that combines human moderation and machine learning to identify malicious content meant to deceive people such as deepfakes.

2018 Sep Deepmind publishes the "Big GAN" paper, showing high resolution synthesis of many ImageNet objects and animals.

2018 Nov The Wall Street Journal shares guidelines on how they are preparing to fight the wave of fake videos in news.

2018 Nov State-owned Xinhua News Agency debuts a synthetic AI news anchor that can read news for 24 hours a day, showing realistic-looking speech, lip movements and facial expressions.

2018 Nov The Washington Times, citing top US officials, warns that America's enemies will use deepfakes to foment 2020 election chaos and "destroy lives".

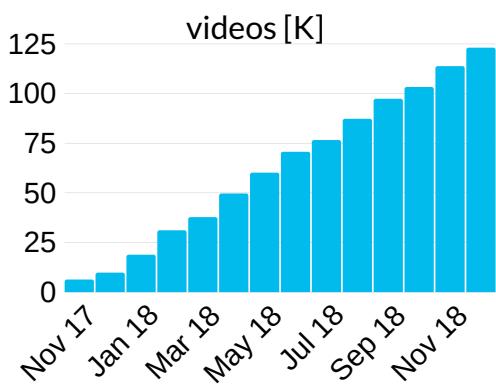
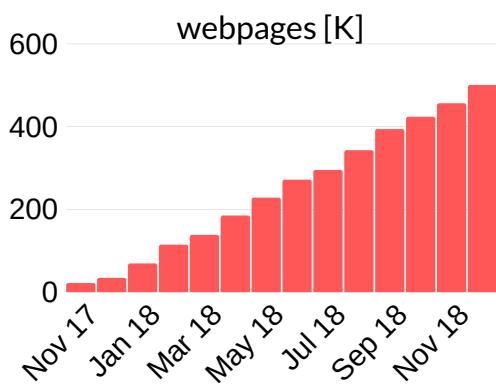
2018 Dec Symantec Corporation presents a demo for a deepfake detector based on face recognition at BlackHat, London.

2018 Dec NVIDIA presents "A Style-Based Generator Architecture for GANs", setting the new standard for synthetic image generation.

INTERNET TRENDS

The widespread publication of articles discussing deepfakes and their impact illustrate increasing public awareness of the threat deepfakes pose. This coincides with the increasing online presence and sophistication of deepfakes.

The number of webpages returned by Google search for "deepfake" grew rapidly from last year, as well as searches for webpages containing related videos. Cumulative statistics:



Until 2017 "deepfake" Google searches globally topped out at 100 per month. With the attention of mainstream media, the keyword attracted between 1M and 10M searches per month, until Jul 2018. The Internet attention stabilised to at most 100K per month as of Dec 2018, a 1000x increase prior to the Reddit phenomenon:

1000x 

ADULT INDUSTRY

The online adult industry is the first to experience large scale impact from the spread of deepfakes, driven by consumer demand for face swapping in a pornographic context

Number of deepfake videos hosted by the top 10 adult websites for traffic. It does not take into account *pornhub.com* which has disabled searches for "deepfakes", although the platform still hosts many despite declaring this ban.

1790+

Number of deepfake videos hosted by adult websites featuring fake video content only:

6174

Number of new websites dedicated to hosting deepfake pornography (all operative since 2018):

3

GITHUB DEVS

The commoditisation of tools for video editing and synthesis is contributing to the spread of deepfakes. A community of developers is growing around open source projects for creating these tools.

Cumulative number of code commits to the most popular deep learning Github repository for performing face swaps in videos, *deepfakes/faceswap*:



The size of the community involved with *deepfakes/faceswap* is comparable to that of other industrial-level deep learning libraries. Number of stars, forks and contributors:



davidsand/facenet

★ 6667 ⚡ 2813
👤 30

facebookresearch/Detectron

★ 18138 ⚡ 3723
👤 23

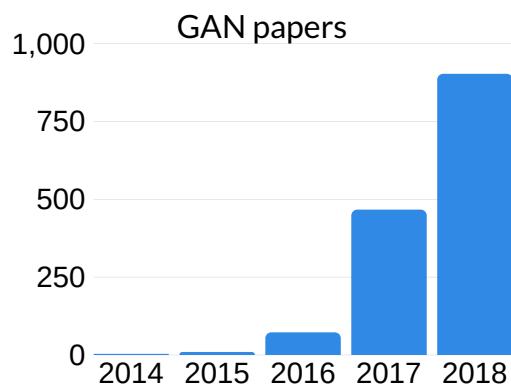
explosion/spaCy

★ 11667 ⚡ 1904
👤 279

PROGRESS OF GANS

Since its inception in 2014, the research community working on improving GANs for generative modelling has been prolific.

Number of papers published on the arXiv (subclasses cs or stat) including "GAN" or "Generative adversarial network" in titles or abstracts has grown to 902 in 2018 only:



This increase in papers is directly related to improvements in generated image quality. A common benchmark used in research papers is the generation of human faces and animals, which we show per year:



ANTI-FORGERY PAPERS

Several research institutes are working on deep learning methods for detecting tampered and synthetic imagery. Many are funded by DARPA's Media Forensics (MediFor) program.

Numbers of papers published by region in 2018. International co-authorships are counted multiple times:



Most prolific in 2018: University Federico II of Naples (Italy), Technical University of Munich (Germany), IDIAP (Switzerland), University at Albany, SUNY and University of Maryland (US)

25

Papers on the topic published in 2018, also including non-peer reviewed ones. 12 are funded by DARPA.

902

In contrast, papers on GANs uploaded to the arXiv in 2018 (data from the previous page)

2019 OUTLOOK: OPINIONS

“ My assumption is that by 2020, this stuff has spread a bit farther and become cheap enough so that we've enlarged the pool of jokers sufficiently so that somebody does this.

Jack Clark, Policy Director at Elon Musk founded OpenAI

[IEEE Spectrum]

“ 2019 will be the year that a malicious ‘deepfake’ video sparks a geopolitical incident. We predict that within the next 12 months, the world will see the release of a highly authentic-looking malicious fake video which could cause substantial damage to diplomatic relations between countries.

Katja Bego, Senior Researcher at NESTA

[NESTA]

“ We are not likely to be awash in deepfakes anytime soon. This technology will remain, for the near-term, a narrow technique likely to be leveraged by states and other well-resourced actors. That's particularly true in a world where there are significantly cheaper and equally effective means of spreading disinformation.

Tim Hwang, Director at Harvard-MIT Ethics and Governance of AI Initiative

[The Poynter Institute]

2019 OUTLOOK: OPINIONS

“

The first wave of deepfakes and maliciously synthesized media is likely to be upon us in 2019. We have the opportunity to be prepared.

Samuel Gregory, Programme Director at WITNESS

[World Economic Forum]

“

I spoke recently with one of the most senior U.S. intelligence officials, who told me that many leaders in his community think we're on the verge of a deepfakes "perfect storm." [...] First, this new technology is staggering in its disruptive potential yet relatively simple and cheap to produce. Second, our enemies are eager to undermine us. [...] China will eventually be incredibly good at this, and we are not ready.

Ben Sasse, US Senator

[Washington Post]

“

The world's largest democracy, India, will hold general election in 2019, while the world's second largest democracy, the US, will begin the lead up to its 2020 presidential campaign. These elections are likely to trigger the production of sophisticated deepfake content made to manipulate and misinform the public. [...] 2019 will see Internet service providers, operators and regulators look seriously into mitigating deepfake content.

Telanor Research

[Telanor]

A close-up photograph of the iconic Selfridges building in Birmingham, featuring its distinctive blue perforated metal facade. The facade consists of numerous small, rounded, protruding elements that catch the light, creating a textured, almost organic pattern. A dark, curved structural element runs diagonally across the frame.

DEEPTRACE

Deeptrace B.V.
Amsterdam, the Netherlands

info@deeptacelabs.com