HERBERT SMITH FREEHILLS

DEEPFAKES: IS SEEING STILL BELIEVING?

04 November 2019 | London
Legal Briefings

The BBC thriller *The Capture* has captured the public's imagination with its portrayal of the relationship between deepfakes and CCTV evidence, and the serious legal risks associated with this technology.

In a recent report published by non-profit research institute Data & Society, deepfakes were found to be "no new threat to democracy", in that audiovisual media has always been manipulated, for a variety of purposes. What is new is the convincingness of deepfakes; the challenges associated with detecting them; and the risks associated with sharing deepfakes at speed and at scale on social media.

In this post we discuss the issues arising from such fakery and the possible legal counters.

WHAT ARE DEEPFAKES?

Put simply, deepfakes are fake videos created by artificially intelligent systems, predominantly trained using machine learning methods in order to generate or manipulate human bodies, faces and voices.

In the Data & Society report, deepfakes are distinguished from 'cheap fakes': fake videos made using package software or no software at all. Cheap fake techniques may include adjusting the timing of footage, deleting or cutting frames together, re-dubbing sound, or simply re-contextualising footage by changing narration, captions or video titles.

Once confined to academic research groups and Hollywood studios, deepfake technology is now accessible to anyone with enough computational resource to manipulate it. Consumer grade animation software can be used in conjunction with open source programs available on public repositories like GitHub to produce fakes of a similar quality to those created for legitimate purposes by computer scientists.

The first widely-known examples of amateur deepfakes appeared in November 2017, when a Reddit user called “deepfakes” uploaded a number of fake videos depicting celebrities’ faces grafted onto pornography.

FAKE HISTORY

It is important to remember that audiovisual fakery is nothing new. Since the beginning of motion pictures, efforts have been made to create visual and audio effects using methods other than filming or recording them. Historically, these effects were achieved by manually altering existing footage in the cutting room, or later, in special effects departments. More recently, computer-generated imagery has advanced to a point where fakes in the film industry are now not only remarkably convincing, but commonplace – such as the recreated version of the late Carrie Fisher in recent Star Wars movies.

However, what *is* new is the potential for deepfake technology to help audiovisual fakery cross the line from expressive content to informative content.

FAKE NEWS

Fewer people now consume news via traditional methods such as reading newspapers, instead favouring audiovisual media such as 24/7 news channels or news apps on smartphones. In today’s digital age, news is no longer the exclusive remit of mainstream media coverage; many of us today also consume news via alternative resources, such as Twitter, reddit and other social media platforms.

This means that the phenomenon of fake news is not confined to text media alone, and whilst news articles written or published by unreliable sources are easily discredited, video content is much harder to argue with. Consequently, deepfake technology provides an opportunity for malicious actors to try to pass off fiction as fact. Technology now exists which makes it possible to synthesize entirely fake audiovisual performances by almost anyone, providing the “ability to put words and actions in the mouths and bodies of others”. For example, there have been several high-profile examples of deepfake videos featuring Barack Obama, such as the fake speech video created by researchers at Stanford University, and the “public service announcement” video warning against the dangers of deepfakes, created by BuzzFeed and American actor and comedian Jordan Peele.

However, it is not only public figures who are at risk of their images being faked. It only requires a few hundred images of training data to produce a reasonable quality deepfake, and in today’s digital age, there are plenty of images of most people publically available on their social media profiles. As such, it is relatively easy for any one of us to be “faked”.

It is important to note that fake news is a concern for cheap fakes and deepfakes alike – given the speed and scale at which information is transmitted via social media platforms, content can go viral before moderators, fact checkers, journalists and mainstream media are able to identify that content as being faked. The problem with deepfakes, though, is how much harder they are to detect, even with modern deepfake detection algorithms, which are not yet widely available.

This creates a risk that an unsuspecting journalist or unscrupulous mainstream media outlet may pick up a deepfake news story and perpetuate it via legitimate sources, thereby 'validating' the fake content as real news, with potentially harmful consequences.

FAKE EVIDENCE

The possible implications of deepfakes on the reliability and admissibility of CCTV evidence in court were explored to devastating effect in *The Capture*: if seeing is no longer believing, do deepfakes change evidence as we know it?

Historically, courts have always struggled with new forms of evidence. For example, expert witnesses were required to justify the admissibility of photographic evidence to 19th century courts, despite its relative reliability as evidence compared with written witness statements and oral testimony.

That is not to say that audiovisual evidence is the same as 'truth' – such evidence can, and has, been manipulated in past trials, such as in the famous Rodney King case in the US, in which video evidence of a violent arrest was deliberately slowed in such a way as to undermine allegations of police brutality. The jury said that the slow video had "made all the difference", and the police officers were acquitted. Consider the potential effects, then, of video 'evidence' which is not only manipulated, but which has been entirely fabricated by deepfake software, placing people at crime scenes or even depicting them committing crimes.

With not only CCTV, but photographs, dashcams, amateur footage, and the results of facial recognition software, all playing a potential role in court cases, there will be serious political, security and criminal implications if audiovisual evidence can no longer be trusted as proof.

TECHNICAL AND LEGAL COUNTERS

Whilst a limited amount of legislation dedicated to deepfakes exists in some jurisdictions (for example, Californian legislation which places restrictions on deepfakes depicting politicians within 60 days of an election), there is no UK legislation specifically designed to address deepfakes.

Existing causes of action which could be relevant include:

- **Copyright and social media takedown requests**

Currently, social media platforms such as Facebook, Twitter and Youtube, and website hosting companies, benefit from the ‘safe harbour’ principle under US and EU law, meaning that the platforms themselves can’t be held responsible for user content which infringes copyright so long as they have a process for notification and takedown of such content (most providers’ processes are governed by the US Digital Millennium Copyright Act, even if they are not US-based). Content creators can submit evidence of copyright infringement to the platforms (usually via standard-web based forms available on the platforms), and the platforms are then responsible for removing the infringing content.

This mechanism was recently used to remove a Kim Kardashian deepfake from YouTube, and thus offers photographers and filmmakers a relatively straightforward method of removing deepfakes which use their copyrighted work. Additionally, blocking injunctions may be available against internet service providers (ISPs) to tackle repeated and concerted attempts to post infringing content on websites.

However, use of copyright material is permitted under the Copyright, Designs and Patents Act 1988 in a considerable number of circumstances, including for purposes of reporting current events, caricature, parody or pastiche, meaning satirical deepfakes may be exempt from copyright claims. In addition, if only a part of the copyrighted work is used, there could be claims that the deepfake would be exempt under the ‘quotation’/extract provisions.

- **Image rights**

Whilst celebrities enjoy the benefit of ‘image rights’ in some jurisdictions, which protect them from misuse of their likeness, these rights do not exist in the UK, so individuals whose images are used in deepfakes must seek other means of redress (often using passing off via false endorsement, as in the *Rihanna v Topshop* case [2013]). Depending on the context of the deepfake (e.g. if the deepfake is used in a commercial context), options may include bringing a claim under passing off/false endorsement or perhaps trade mark infringement if appropriate images are registered as trade marks.

- **Data privacy**

Individuals' images constitute personal data under Art 4(1) of the General Data Protection Regulation ('GDPR'). As such, victims of deepfakes may also be able to enforce their data protection rights against the data controller, including the right to rectification (Art 16) and the right to erasure (Art 17). However, identifying the 'data controller' (or controllers) in the context of a deepfake video shared online via multiple platforms is unlikely to be a straightforward process.

- **Defamation, harassment and misuse of private information**

These offer alternative causes of action in circumstances where the deepfake: would harm an individual's reputation, amounts to harassment, or is pornographic (which would amount to private information, even though the information is false (*McKennit v Ash* [2006]).

Proposed measures to counter the "fakes" problem:

The shortcomings associated with existing options set out above are compounded by the fact that, even if the victim is able to make out a case, under one or more of these grounds, they may face difficulty in actually bringing their claim if, for example, the creator of the deepfake cannot be identified, or if there are jurisdictional obstacles.

Consequently, there are widespread calls for further legal and te counters to deepfakes. Potential counters being considered include:

- Legislation requiring the identification or registration of content creators
- Legislation forcing individuals to label manipulated content and fine those whose manipulative content is deemed harmful.
- An improved regulatory and legislative framework around potential rights of action for victims, such as establishing image rights in the UK.
- Adding invisible 'noise' to digital images which deepfake algorithms struggle to process, impacting the quality of the resulting deepfake content.
- Tracking the provenance of content through use of distributed verification technology (such as Truepic, which uses blockchain technology to verify the authenticity of content).
- Removing net neutrality for all packets carrying video data, in order to trace them to their real-life creators.

- Embedding automated fake detection software on social media platforms, search engines and internet browsers.
- Employing more content moderators and reconsidering platforms' liability for their role in spreading fake content.
- Raising awareness of fake audiovisual content and encouraging the public to interrogate their sources.

SUMMARY

Whilst audiovisual manipulation is nothing new, whether in the context of news media or evidence in court, the advent of deepfake technology makes fake content much harder to identify. In a world where seeing is not necessarily believing, and where audiovisual footage is shared at speed and at scale on social media and other platforms, the risks associated with deepfakes and cheap fakes alike are becoming increasingly important to address.

KEY CONTACTS

If you have any questions, or would like to know how this might affect your business, phone, or email these key contacts.

RACHEL MONTAGNON
PROFESSIONAL SUPPORT CONSULTANT, LONDON
+44 20 7466 2217
Rachel.Montagnon@hsf.com

LEGAL NOTICE