36 CYBER BITES

Integrity of Content: Fighting Disinformation, Deepfakes, and a Deadly Virus

It was George Orwell who coined the concept of 'doublespeak' in his dystopian novel "1984". 'Doublespeak' combines two Orwellian themes "doublethink" and "newspeak" and creates a language that deliberately obscures, distorts, disguises, or reverses the meaning of words to manipulate public opinion as a method of controlling thought through language - it is the language of 'fake news', of disinformation.

In 2020 disinformation, in particular 'fake news', has become a globally recognised phenomenon, and with the advent of Deepfakes – visual and audio disinformation – it has been observed that we are entering a new chapter in the battle for truth on the Internet.

"1984" is all the more important as a cautionary tale, if you like, precisely because it increases our awareness and understanding of the processes at work today as information is distorted in the fight against Coronavirus. Telling fact from fiction, truth from lie has never been more important.

Deepfakes can be defined as synthetic media, visual and audio content that has been manipulated using advanced AI software to change how a person, object, or environment is presented. Deepfakes are fake videos or audios made to look and sound real. Since 2016 Deepfake technology has exploded onto the market, becoming increasingly sophisticated and able to look and sound more 'real', and therefore harder to detect.

The dangers of Deepfake technology were powerfully demonstrated by the Buzzfeed's widely viewed Obama deepfake video (https://www.youtube.com/watch?v=cQ54GDm1eL0) using comedian Jordan Peele to give a stark warning about the use of the technology and the need to verify content before holding it as truth: seeing is no longer believing.

One thing is clear though: the threat of Deepfakes shouldn't be downplayed. In criminal, hands, Deepfake technology can be a powerful cyber weapon. It can be deployed to propagate disinformation, 'fake news' for political reasons and propaganda, to sow division between countries, between political rivals, to discredit journalists and others.

The technology can be used for fraud, in the form of voice cloning Deepfake. There have already been cases where companies were victims of such an attack by a simple phone call using a voice cloning Deepfake the attacker convincingly copied the voice of the company's CEO instructing his financial manager to make an urgent transfer of funds to a compromised account.

To this list, there is now the use of Coronavirus disinformation online. Internet users in developing countries, where digital literacy remains relatively low, could be more susceptible to believing Deepfakes, such as the ones pretending to be published by the World Health Organisation claiming that gargling with salt water prevents the virus from penetrating cells in the throat.

All major social networks are now under pressure to combat disinformation surrounding the Coronavirus pandemic. Both Facebook and Twitter have deleted posts spreading disinformation about the virus. So has the BBC in a campaign to fight 'fake news', raise awareness, and protect the public from harm.

Facebook deleted a video from Covid-denier Brazilian President Jair Bolsonaro claiming that hydroxychloroquine was effective in treating the virus. It follows Twitter's deletion of a homemade treatment tweeted by Venezuelan President Nicolás Maduro.

As of 2 April 2020 Twitter has deleted 20,000 fake accounts linked to the governments of Serbia, Saudi Arabia, Egypt, Honduras and

Indonesia, saying they violated company policy and were a *"targeted attempt to undermine the public conversation".*

Twitter also took down a bot network of more than 9,000 Twitter bots that published fake posts promoting the political interests of UAE and Saudi Arabia. Researchers uncovered the network by searching for coronavirus -related hashtags. Under the guise of coronavirus related posts, a closer inspection of the accounts revealed they were used for broader political messaging. Many of the accounts were batch-created meaning they were all first created on the same day within a short timeframe, and the primary purpose of the accounts was to amplify other content rather than engaging with the Twitter community.

There is a growing and darker aspect to this bot-driven political rhetoric; Deepfake, manipulated content online is being designed specifically as a tool for control by totalitarian regimes, as a way for extremist groups to stir up social division exploiting the pandemic for political purposes worldwide.

As well as being disruptive in their own right, there is another surprising and troubling twist about Deepfakes: they could give bad actors an excuse to deny involvement in questionable/criminal activity. If every video has the potential to be a fake, it offers people the opportunity to challenge the veracity of genuine footage.

This, in turn, could have far reaching consequences for how evidence is used in criminal investigations for example, with some experts being of the view that images, stills or moving, will soon become obsolete as physical evidence in criminal trials because of the threat to their veracity by digital alterations. The rise and sophistication of Deepfakes could potentially have the effect of forcing law-enforcement in the future to prove not just what is false but also what is true.

Regulatory/ Legal framework

This world pandemic has brought into sharp focus the requisite to protect and safeguard content and its integrity online in an increasingly fragmented world. There is a greater need now for legislation to encourage social media platforms to pay closer attention to the content being posted on their platforms.

The Government's new Online Harms White Paper sets out an ambitious vision in this regard, with a proposal to establish a new duty of care that would hold tech companies accountable for addressing a range of online harms, one of which being the spread of Deepfake material and disinformation.

Of note is the proposal for an independent regulator to oversee this framework with enforcement powers such as the power to issue warnings, notices, and substantial fines. The White Paper also included proposals for business disruption measures - including Internet Service provider (ISP) blocking, and senior management liability.

The White Paper states that the regulatory framework will apply to online providers that supply services/ tools, which enable or facilitate users to share user-generated content, or to interact with each other online. The regulatory approach will focus on safeguarding freedom of expression, and be based on proportionality and on evidence of risk of harm, with the duty of care designed to ensure companies have appropriate systems and processes in place to improve the safety of their users, and to protect them from harm.

To ensure clarity about how the duty of care could be fulfilled, the Government proposes to provide codes of practice about the applicable expectations on businesses, including where businesses are exempt from certain requirements due to their size or risk.

For SMEs and start-up companies there is now a pressing need to prepare themselves ahead of the new regulatory framework coming into force. Companies will require new systems and measures in place capable of protecting themselves and their users from the risk of harm online. For example, investing in Deepfake detection technology would be a way of ensuring a system able to certify the veracity of audio/visual content. There have been software developments in the filed of provenance checking, which is another way of authenticating third party content. The BBC has long been running a very successful service called 'reality check' which authenticates factual content. This will become a standard requirement for all companies in the future ensuring the safety of their online users, in compliance with their duty of care.

Seeking expert legal advice ahead of the Government's implementation is paramount, and will give companies peace of mind and ensure their senior management are not left vulnerable to potential complaints and liability.

There is also an urgent need for updating existing legislation to factor in Deepfakes. **The Computer Misuse Act 1990,** the main act which deals with cyber criminal offences described as 'hacking' offences is barely adequate to meet the rapid advancement in technology.

Section 3(2)(c) of the Act deals with an unauthorised act in relation to a computer by a person intending to do that act to *'impair the reliability of any such data".* The meaning of *'impairing the reliability of any such data'* was interpreted for the first time in **Zezev and Yarimaka-v-Governor of H.M. Prison Brixton [2002] 2 Cr App R 33**. The Lordships found that *"If a computer is caused to record information which shows that it came from one person, when it, in fact, came from someone else, that manifestly affects its reliability. This information is undoubtedly data."* The Court further found that information that would tell a lie about itself was sufficient to impair the operation of a computer, a forward thinking and forward- looking interpretation at a time when Deepfake technology had yet to be invented.

FLAVIA KENYON