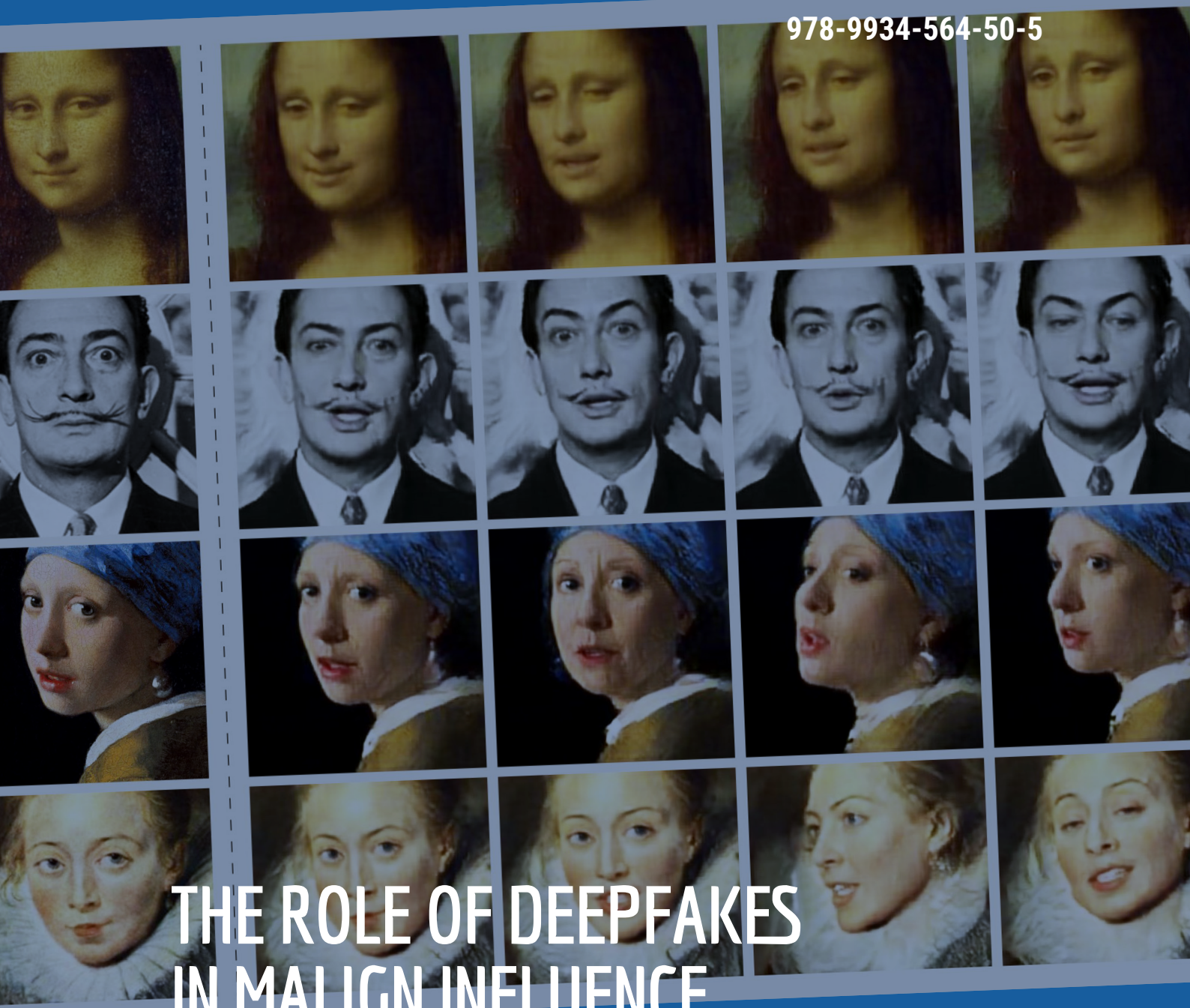
978-9934-564-50-5



THE ROLE OF DEEPFAKES IN MALIGN INFLUENCE CAMPAIGNS

Published by the
NATO Strategic Communications
Centre of Excellence

ISBN: 978-9934-564-50-5
Authors: Keir Giles, Kim Hartmann, and Munira Mustaffa[1]
Contributors to the Project: Berta Jarosova, Nathalie van Raemdonck and Liisa Past
Project manager: Sebastian Bay
Copy-editing: Anna Reynolds
Design: Kārlis Ulmanis

NATO STRATCOM COE
11b Kalciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook/stratcomcoe
Twitter: @stratcomcoe

This monograph was completed in September 2019, and draws on source material that was available by August 2019.

Introduction

The Ballad of Katie Jones

In early 2019 Katie Jones, a young researcher based in Washington DC, set up a new profile for herself on LinkedIn. She filled in her biographical details—degrees from the University of Michigan, a present position at the prestigious CSIS think tank—and set about building her network of contacts.

There was nothing unusual about a new profile being set up on LinkedIn; thousands are created every day. It was not even unusual that Katie Jones was entirely fictitious; with no checks on the identity of individuals creating these profiles, there is nothing to prevent someone from assuming any identity they choose.

Instead, what distinguished Katie Jones from thousands of other false online personae was the image she chose as a profile picture. Ordinarily, fake social media identities steal a picture from another profile, use a stock image, or copy a photograph of a public figure from websites. Instead, Katie was a chimera; her profile used a unique image that was not a photograph of a human being, but was an entirely computer-generated artefact made using machine learning algorithms. Katie's profile, and the use to which it was put, may be the first instance documented in open sources of a malign influence campaign making use of human image synthesis based on machine learning—commonly referred to as a deepfake.

It is not known who created Katie Jones, or for what purpose; whether that purpose was achieved; and if so, what impact it had. But the integration of a computer-generated image sufficiently convincing for a number of highly intelligent individuals to accept Katie Jones's invitations to connect raises a wide range of implications, not only for deepfakes and other deception technologies, but for malign influence campaigns overall.

This paper describes development paths in both technology and deception, considers possible future developments, and discusses policy implications for governments, corporations, and private individuals.

Why Was Katie Jones Created?

Katie Jones connected with current and former think-tankers, academics, military officers and government officials, but the purpose for which she was created is not immediately apparent from the categories of real individuals she sought out. Although the majority were located in the United States and engaged in work on Russia, outliers include specialists on China, energy, and cyber security; a PR specialist in Las Vegas; and a Moscow-based director of an American refrigeration company.

In total, 52 people accepted Jones's invitations to connect. Among the most senior professionals were a former one-star general and US Defence Attaché to Moscow, a top-ranking US State Department official, and a potential nominee to the Federal Reserve. Over 40 of these individuals were interviewed by Raphael Satter of the AP news agency during research for his news story on Katie Jones. None of these people admitted to having received any direct communication from Katie. And Katie Jones did not respond to messages and invitations sent to her via LinkedIn, or to an AP request sent to her AoL address 'asking her to comment on the fact that she did not exist'.[2]

Becoming active in March 2019, the profile was identified as a fake in early April and publicly exposed in June.[3] It is possible that Katie was interrupted before she fulfilled the purpose she was designed for; but given that she appeared to cease making connection requests before becoming the object of suspicion, it is equally possible that she accomplished her mission before being detected.

Possible purposes for which Katie was created include, but are not limited to, the following:

- Advanced spearphishing: establishing an account as a trusted source for corresponding with high-value target individuals to deliver malware or spyware to their device(s) by e-mail or some other messaging system;
- Gaining online or physical access to events by signing up to lists and receiving notifications and credentials;[4]
- Mapping networks: gathering information on who is connected to whom in specific fields of policy research;
- A test run: assessing the plausibility of a profile of this kind, its success at penetration and network-building, and testing its detectability to inform future targeted deception campaigns;
- A joke, perpetrated simply to entertain the creator(s)—in which case they will find the rest of this analysis even more amusing.

After exhaustive attempts by AP to track down and identify all Katie's contacts, one objective that can probably be ruled out is building *bona fides* and influence for another fake profile within her network. But the question remains: how was a fake Katie Jones able to convince a series of intelligent, well-informed individuals to accept her invitations to connect? The answers lie in the convergence of the new technology used to generate her face, the unchanging principles of deception, and the very specific nature of the platform on which she appeared. Each of these will be examined in turn in this paper.

Deepfake Timeline

Lip-synced video manipulation of former president Barack Obama

2014

2017

2018

Jun — First GAN developed

Nov — /r/deepfakes subreddit created

Dec — Term "Deepfake" coined

Jan — FakeApp launched

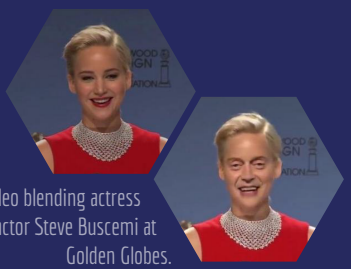
Feb — Reddit bans /r/FakeApp

Apr — Deepfake video of former President Barack Obama

Nov — China debuts world's first AI news anchor

Deepfake video of actress Daisy Ridley's face pasted onto another person's body.

2019



Deepfake video blending actress Jennifer Lawrence with actor Steve Buscemi at Golden Globes.

ThisPersonDoesNotExist.com launched

Feb

Katie Jones created

Mar

Deepfake video of U.S. President Donald Trump advising people of Belgium on climate change.

May

Feb

China unveils first female AI news anchor

Doctored videos of Mark Zuckerberg delivering a modified 2017 video statement on Russian interference, and of Kim Kardashian made with video dialogue replacement (VDR) technology.

May

Distorted video of US House Speaker Nancy Pelosi

Deepfake of actor Keanu Reeves's face in Scarface movie

The Superpersonal app captures a user's face and micro mannerism to create a hyper-realistic moving image of the user's best fashion model self

Deepfakes

Deepfakes Defined

'Deepfake' is a portmanteau of 'deep learning' and 'fake', referring to the method used and the result obtained.

Although most commonly associated in popular consciousness with video, a deepfake is any digital product—audio, video, or still image—created artificially with machine learning methods; and, as discussed below, there is an argument for extending the definition to include output in text form as well. A deepfake is produced by a set of deep learning algorithms known as a GAN, or Generative Adversarial Network. The system consists of two artificial neural networks working against each other—a generator that creates data and a discriminator that judges whether the result is plausible. In other words, a GAN is a machine learning system that determines for itself whether its generated fake output is sufficiently realistic, and if not refines it further.

Where the objective is to mimic a certain known individual, data is collected about the intended target and used to train a neural network. This neural network is able to identify the specific traits of the target that can be used to make picture, audio, and video recordings of this individual unique and identifiable. The information gathered is then used to create new, artificial content that represents the target but is not in fact based on an original image or recording of them. In the case of Katie Jones—an entirely new artificial individual—the challenge was simpler; the content merely had to be convincingly realistic instead of convincingly like a specific person.

A distinctive feature of the deepfake problem is the speed at which the necessary technology is becoming more sophisticated and more widely available. At the Riga StratCom Dialogue conference in June 2019, a presentation focused on *future* threats introduced the concept of GAN-produced images and rhetorically asked: 'Could someone use fake faces for nefarious purposes?' The following day, *AP* published its months-long investigation into Katie Jones, making it plain that rather than being a hypothetical future threat, this was one that was already real and present.

The content created by AI to date is imperfect and contains flaws. These can result from poor specifications provided by humans for their AI algorithms—for example, combining incompatible images can result in blurred ears or mismatched earrings.[5] Such issues are likely to have been of little significance for consumers in what was originally the fastest-growing

application for deepfake video technology, namely pornography, where ears may have been of only tangential concern.[6] But the drive for authenticity in more mainstream applications means such visible flaws will soon be corrected. Creators will continue refining their AI algorithms and input, and the pace of development is such that by the time this paper is published, deepfake images may well have become sufficiently sophisticated that they are undetectable to humans without specialist equipment. In addition, deepfake technology, including dedicated programs and downloadable facesets, is becoming more affordable for the general public. Any notion that deepfakes are exotic or unusual will soon be undermined by their increasing ease of production.[7]

Deepfake Hype

This imminent ubiquity has not made forecasts of the impact of deepfakes any less alarming.[8] Among a proliferation of less high-profile examples, two instances that have brought the potential of deepfakes to broad public attention were a synthesised lip-synced video of Barack Obama[9] and a relatively crude rendering of Mark Zuckerberg apparently giving an unusually frank public statement on Facebook's ambitions for total global domination.[10] In both cases, the demonstrated capability to create a plausible imitation of a public figure by using readily available data has triggered great concern over the potential use of deepfakes for political manipulation.[11] In a 2018 article two American law professors noted that: 'The potential to sway the outcome of an election is quite real, particularly if the attacker is able to time the distribution such that there will be enough window for the fake to circulate but not enough window for the victim to debunk it effectively (assuming it can be debunked at all).'[12] A developing awareness that deepfakes can be deployed to deliver false but believable messages as though they are from people we know has led to warnings that 'violent social unrest' may be triggered by deepfakes,[13] and a concern that 'digital manipulation of video may make the current era of "fake news" seem quaint'.[14]

However, much of the current alarm over the potential impact of deepfakes overlooks the fact that this technology has already been available for some time, yet was not exploited for malicious purposes when it was still relatively unknown and would have had the greatest power to convince an unwary public. Already by early 2017 it was recognised that: 'the CGI technology for faking video is mature and affordable […] So instead of rumours and gossip about politicians, why we have not seen videos of politicians taking drugs, or taking bribes, or taking liberties, or torturing puppies, or simply making policy statements which are completely incompatible with what they are supposed to say?'[15]

Instead, deepfakes have rapidly become a widely recognised concept among the general public and a common topic in the

media. By going mainstream, deepfake videos in particular have lost their initial shock value. In 2016–17, amid growing recognition of the ambitions of Russia's information warfare, it was predicted that malicious actors would withhold their deepfake capabilities until faced with an event of sufficient importance to justify their use against unsuspecting adversaries. If this is so, they have been withheld too long, since the adversaries are no longer unsuspecting.

Furthermore, it is well recognised that a story does not have to be credible to go viral. But in addition, it does not need to be supported by doctored imagery or video; in fact in some cases, this could be counter-productive, as fakes are easily compared side by side with originals for the purpose of debunking disinformation. If this debunking is successful and achieves the same or greater virality as the original, it delegitimises the messengers that spread the fake among those sectors of their audience that are open to doubt. Ill-judged use of deepfakes, therefore, could burn through the amplifiers and toxic superspreaders that disinformation campaigns rely on for achieving their broad penetration.[16] In addition, recent examples show comparable effects can be achieved using doctored or edited videos—'dumb fakes'—with no need for advanced technology or AI-generated content.[17] Videos appearing to show Nancy Pelosi drunk,[18] or Jim Acosta showing violence to a White House intern,[19] demonstrate that simple editing remains just as effective at discrediting individuals.[20]

Other media are no less suitable for tapping into timeless principles of deception. GANs are also capable of generating authentic voice imitations.[21] As described below, if they impersonate a known individual this raises the possibility of malign actors successfully imitating a CEO to defraud a company, or a CO to deceive a military unit. But there are substantial implications even with the creation of fake voices that are entirely anonymous. Google's attempt at impersonation for its Duplex scheduling assistant sounded sufficiently human that serious ethical concerns were raised. Google responded to the resulting protests by introducing transparency measures—the system now identifies itself as an automated voice when making calls.[22]

The production of deceptive text output by machine-learning systems also has the potential to be highly dangerous. In the early part of this decade, Russia found that automated systems were inadequate to influence debate on domestic or international fora and social media, and human intervention in the form of professional trolls was required.[23] With the development of 'fake text', trolls could finally be replaced by automated systems for all but the most sophisticated interactions with real humans. This new automated astroturfing could result in a massive scaling up of operations, generating an entire fake public to influence political decisions.[24] Once again, malign influence operations could follow a trend set by the commercial sector, where chatbots generating near-lifelike interactions have proven cheaper and more effective for routine transactions than outsourcing customer service operations to India (the only remaining question being whether they provide a more or less infuriating customer experience). There is even a text-based equivalent to 'dumbfake' video, in the form of images that resemble screenshots of tweets. Trivially simple to create, they have already been used to sow confusion and attempt to discredit public figures.[25]

Even so, predictions of deepfake doom from the expert community are beginning to give way to more sceptical and moderate assessments of their potential impact;[26] particularly as 'the public comes to terms with what seems like an inevitability […] that people can and will use AI to create super-realistic fake videos and images'.[27] This inevitability is driven not by the use of deepfakes for political machinations, but by their adoption in marketing. Techniques for achieving virality in malicious influence campaigns are precisely the same as those used to market a product. The rapid normalisation of deepfakes is being driven more by business realities than by adversaries waging information warfare.[28] Consumers engage willingly with 'virtual influencers', marketing tools that are no more real than Katie Jones.[29] Another strong indicator that deepfakes have already lost their power to shock is the virtual news anchor phenomenon. AI anchors have been created using 'machine learning to simulate the voice, facial movements, and gestures of real-life broadcasters' and are followed by millions of television viewers.[30]

Deception

History

Deception, disinformation, and even 'fake news' are perpetual societal problems, not new technological ones. Deepfakes should not be treated as a phenomenon entirely distinct from other forms of deceptive content. They are not indicative of any change in human behaviour. Misleading images of people, whether intended to flatter or to deceive, have been known for as long as human likenesses have been created. Propaganda and disinformation images long predate the invention of the camera, and the manipulation of photographs, for innocent or malicious purposes, is as old as photography itself. Whether in audio, video or still images, each new technology brings with it new means of deception. Digital editing of audio replaced the need to slice and splice snippets of reel-to-reel audio tape with razor blades and sticky tape, and computer-based retouching of images replaced the need to cut and paste literally, with scalpel and glue. The latest tools for creating deceptive video simply continue this trend of the techniques and technology required for deception becoming faster, easier, cheaper, more widely available, and much less reliant on specialist skills.

Two further key factors remain unchanged despite rapid technological advances. If a fake is well constructed, detecting it requires either special equipment, special training, or an original for comparison, in exactly the same way as since the earliest days of visual or auditory forgery. But an additional unchanging factor is human susceptibility. Now, as ever, the creators of fakes can rely on human willingness to suspend disbelief, to disengage critical faculties, or to follow base instincts. In an ideal case, the fakers will trigger all three self-destructive behaviours at once. In the words of one US Army officer explaining why he connected online with Katie Jones: 'I clicked on the link because she was hot.'

LinkedIn: A Permissive Environment

While all social media platforms provide a target-rich environment for those wishing to play on human weakness, LinkedIn presents a very specific type of permissive environment for malign influence by the nature of its design and purpose. Because the platform is driven by, and encourages, opportunistic networking, it is far easier for an impostor profile to build a network of connections with genuine users there, including known figures in selected communities of interest, than it is on Facebook or Twitter. This is all the more damaging because LinkedIn profiles are supposed to be of real individuals posting

under their own names, advertising genuine qualifications and experience to potential employers. By connecting with a fake profile, other users implicitly endorse it, bolstering the illusion that the profile represents a bona fide human being in the location and occupation claimed. In the case of Katie Jones it is not accurate to say that individuals lowered their guard on spotting a plausible profile in their area of interest at a leading think-tank; rather, because this was LinkedIn, their guard was never up in the first place.

The structure of LinkedIn makes it simple to map social and professional networks, making it easier for those with malicious intent to infiltrate those networks and contact their members. The range of possible objectives for such an operation is broad. While Russia's most evident hostile activity on LinkedIn consists of aggressive targeting of individuals perceived as critical of Moscow,[31] Western intelligence agencies have repeatedly warned users that China is taking advantage of LinkedIn for espionage—using the platform as a recruiting tool, precisely as it is intended.[32] Some of those targeted display classic characteristics or life circumstances that make them vulnerable to recruitment by a foreign power.[33] In other cases, individuals stumble into espionage simply by seeking legitimate career opportunities.[34] The German domestic intelligence agency, the BfV, has published information regarding Chinese attempts identifying false LinkedIn profiles that resemble precursors to Katie Jones: they 'are designed to look enticing to other users, and promote young Chinese professionals—who do not exist'.[35]

The example of Katie Jones demonstrates a lack of effective impediments to setting up deceptive profiles on LinkedIn, with no meaningful validation of credentials claimed. To take just one example, there is no immediately obvious reason why institutions and employers should not be notified when new accounts add them to their biographies or current workplaces; at present, profiles can claim any affiliation or qualification with no provision for validating that the claims are genuine.[36] The total number of connections allowed is another criterion that, far from constraining undesirable behaviour (even relatively innocuous behaviour such as spamming), almost encourages it. Facebook caps 'friends' at 5,000. LinkedIn caps connections at 30,000—a number vastly in excess of any realistic expectation of a genuine individual's network of contacts.

A potential dilemma for LinkedIn is that any effort to constrain new signups, or even to bring attention to the problem, would be in direct conflict with its business model and pose a threat to profits. But for as long as LinkedIn is unwilling or unable to establish that its users are who they say they are, even at the most rudimentary level, genuine individuals must rely on outside sources for warnings that they may be targets for malign activity on the platform.[37]

Implications

General

Rather than treating deepfakes as an isolated phenomenon, they should be seen as an addition to the existing arsenal of deception, which may, under certain circumstances, offer more effective delivery of malign influence or disinformation. One of their most significant capabilities—generating a chimera capable of convincing viewers they are interacting with a real individual—is notably easy to scale and replicate. Campaigns costing close to zero could be aimed simply at inducing the most careless or gullible individual from a target set to click on the link. This too continues the well-established trend of disinformation tools becoming both more sophisticated and more accessible to a wider set of malign actors with a broader range of budgets.

Introducing effective countermeasures to deepfakes also requires clear allocation of responsibility for managing the challenge. Just as hybrid threats exploit the seams of responsibility between the armed forces and civilian agencies, blended technical and psychological attacks exploit the disconnect between technical defensive measures and those (if any) that are focused on societal resilience.

This can be demonstrated most clearly by asking a simple question: who should someone targeted by a 'Katie Jones' turn to for help? The platform Katie's profile was on? Her ostensible employer? The government? And if so, which one? As described elsewhere in this paper, the first two options are entirely bootless; the third opens up daunting jurisdictional questions. And in any case, few agencies provide contact details specifically for reporting apparent hostile activity on social media in the manner of the German BfV.[38] For the time being, as is often the case, the best available option was to present the story to a journalist from a major news agency with the time and investigative resources to ensure that Katie Jones was properly expiscated.

Meanwhile, Katie is a foretoken of the deepfake challenges that will affect corporations, governments, and individuals, whether or not they are preparing for them.

Governments and Corporations

For governments, as with disinformation overall, deepfakes imply a huge range of challenges across all levels from the tactical to the strategic. In tactical terms, there are clear military applications, even for simple deception: deepfake text or voice technology could generate false but convincing signals chatter on a massive scale, but with very little cost; or at a more sophisticated level,

key individuals in the chain of command could be impersonated giving voice orders. While current disinformation efforts, particularly by Russia, focus on denying that a particular event has taken place, deepfake technologies could facilitate convincing the adversary that something has happened when it in fact did not. At the level of strategic policy challenges, deepfakes exacerbate the problem of disinformation as a public and societal health issue, further amplified by the social engineering power of 'influencers', whether real or virtual.

Corporations are faced with increasing challenges surrounding trust, regarding both the information they receive from the outside world and that which they direct at the public. As a specific subset of this problem, deepfake audio impersonation is an immediate challenge, not a future problem.[39] The capability to convincingly reproduce the voice of a known individual could revolutionise principles of social engineering that have remained relatively static since first codified.[40]

Successful voice mimicry by AI increases the danger of 'CEO fraud' extending from e-mails to telephone calls (or, equally likely, a fraudulent e-mail being validated by a fraudulent telephone call). Symantec has reported three successful audio attacks on private companies where the attackers impersonated the CEO's voice requesting an urgent money transfer through senior financial officers. Each case resulted in a loss of 'millions of dollars' for the company.[41] Victim corporations may find that unlike in the case of a cyber attack illegally intruding into its network, voice-assisted CEO fraud is not covered by existing insurance policies. This is because instead of committing fraud through evading or disrupting security systems, a successful exploit of this kind involves inducing the company's own employees to carry out the fraudulent action in full accordance with standing security and assurance procedures.[42]

The exploitation of well-known brands to validate malign influence campaigns is another enduring problem that will only be exacerbated by deepfake proliferation. When notified of the existence of Katie Jones, think-tank CSIS responded 'first guess is a Russian troll', but took no further action.[43] Those companies and public

organisations that are concerned for their reputation should not follow CSIS's lead.

The involvement of corporations that manage online platforms in particular adds an additional layer of complexity to tackling the deepfake problem, primarily because the main objective of entities like Facebook and Twitter is generating profits rather than defending Western political systems. In the United States, the problem of deepfakes on social media is receiving official attention as part of a more general trend of platforms belatedly coming under political pressure to address their role in facilitating malign influence campaigns.[44] But there are clear limits to the amount of pressure that can be brought on major internet corporations—even through invocation of corporate social responsibility, or reminding them that they should not help subvert the societies within which they thrive and prosper. Instead, efforts to counter the weaponisation of social media platforms remain hamstrung by the lack of engagement of the platforms themselves.

For example, when attempting to counter malicious actors and networks online, defenders are forced to rely on painstaking deductive analysis working only on the customer side of social media platforms, leading to assessments based on the balance of probability. By contrast the platforms themselves have full access to IP and MAC addresses, login logs, and supporting geographical indicators; consequently for them, once attention and resources are directed towards deceptive activity the confident identification and detection of fraudulent profiles is trivial. Social media platforms are constrained by privacy concerns and a duty of care to their innocent customers. But to understand the extent to which their lack of transparency and support for efforts to counter abuse hampers the defence of civil society, it is worth considering an analogy from the physical world. The current stance of most platforms is as though in the wake of the 9/11 attacks in the United States, airlines had not only refused to release passenger manifests or booking information to law enforcement agencies, forcing them to adopt infinitely slower and harder methods to identify the hijackers, but also refused to implement new security procedures

and systems in order to address known vulnerabilities, and thereby continued to allow themselves to be used to carry out easily preventable attacks.

Some steps that could be taken to limit the destructive potential of social media appear simple and obvious from the outside, including a more proactive awareness of what takes place on the companies' own sites and networks; again and again, third party observers are aware of serious problems with social media content well before the platforms themselves.[45] And there can be no reasonable objection to social media taking firmer steps to prevent the creation of profiles that are overtly deceptive, or indeed the hijacking of profiles of genuine organisations and individuals for disinformation aims.[46]

But, for the time being, the platforms remain unwilling or unable to address the fundamental problem of malign influence overall, even in its most simplistic characterisation as 'fake news'. They do not effectively address the ways in which their systems are abused to carry out organised deception targeted at their users—whether this is in breach of their policies, or, as has happened repeatedly, in full accordance with them.[47] In addition, with regard to the potential future deployment of deception campaigns, serious consideration should be given to social media's latent networks and how they could be leveraged. Although platforms have taken limited steps to remove accounts directly implicated in election manipulation from 2016 onwards, troll networks identified in earlier studies (including categories such as 'bikini trolls' and 'Wikipedia trolls') still appear to be active and gathering followers, perhaps held in reserve for deployment at some future date.[48]

Individuals

The individual risk of becoming a deepfake victim will be closely linked to the availability, quality, and accessibility of training data. This may seem to imply that individuals with an extensive or prominent online presence have a higher likelihood of becoming victims of deepfakes. However, in 2019 there are sufficiently large amounts of audio, video, picture, and text materials online, provided voluntarily or otherwise, that most humans with any online presence are at some risk. This easily accessible mass of data can be used as training material to produce deepfakes in text, speech, video, or a combination of these. In addition, the rapid pace at which the technology is advancing, aided by long-standing development techniques from 3-D video gaming, may mean sufficiently convincing fakes can be generated from much smaller data sets, potentially from a sample of one.[49]

The implication is that private citizens face two distinct categories of deepfake-enhanced risk. High-visibility individuals, whether in business, entertainment, politics, or government, may be imitated to deceive others or to attack their reputations. But less prominent people are also at risk,

although in a different form. In addition to high-profile national security implications, deepfakes offer potential for simple fraud against individuals.

At present, manufacturers of connected devices and financial service providers in some countries are increasingly relying on voice and facial recognition techniques to authenticate their customers.[50] In an environment where both voices and faces can be not just imitated but replicated, the new systems may soon be no more secure than the text-based authentication passwords they were supposed to replace; research in late 2018 showed that face recognition systems are highly susceptible to deepfake video material, with false acceptance rates ranging between 85% and 95%.[51] There may well be a lag between the potential for fraud being recognised by users and the introduction of countermeasures by providers. A similar lag occurred when scanning and editing technology for documents became universally available and affordable, but financial service providers continued to request photocopies of documents, not recognising the ease with which they could be manipulated by consumers at home.

Any large collection of selfies makes a convenient data set for a 'deepfake Cadmus' to sow a virtual army of Katie Joneses.[52] But these unique images can also be used for non-malign purposes. Creating a new identity and an online persona has always been an integral part of social media use. In the case of Katie Jones, technology has caught up with intention; it is no longer necessary to craft anonymous avatars when new and unique faces that appear to be of genuine human beings are available with one click.

Detection and Countermeasures

Deepfakes in any format are rapidly approaching the point where they are sufficiently convincing as to be indistinguishable from humans by humans. This highlights the need for a reliable and universally applicable assay method; in effect, a Voigt-Kampff test to tell deepfakes from content based on real people.

Neural networks can be used not only to generate deepfakes, but also to detect them.[53] In detection as well as creation, technical capabilities are advancing rapidly; at the time of writing, under controlled circumstances, a neural-network-based tool can spot image manipulation at the level of individual pixels. This capability is likely to have developed further by the time of publication.[54] High recognition rates have also been claimed for deepfake videos,[55] although tests rely on assuming specific traits of deepfakes that may not be universal.[56]

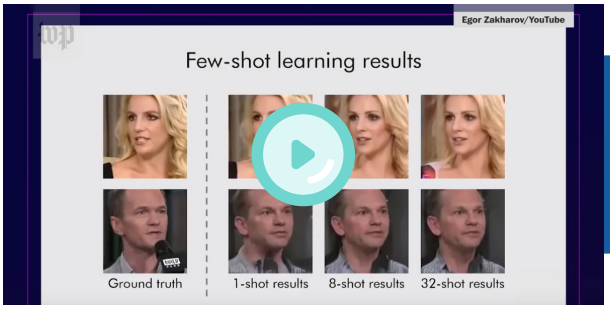
These two parallel tracks of rapid development have already led to an arms race between creation and detection. In August 2019 DARPA, the US Defense Advanced Research Projects Agency, announced a tender for 'technologies to automatically detect, attribute, and characterize falsified multi-modal media assets (text, audio, image, video) to defend against large-scale, automated disinformation attacks'. This was a response to the recognition that current detection techniques 'can often be fooled with limited additional resources', and a stated aim of the DARPA project is to reverse the current burden of effort so that successful falsification is more challenging than detection.[57]

An exclusive focus on technical means of detection may also obscure the importance of environmental factors and supporting evidence. In the case of Katie Jones, a key indicator that there was something unusual about her profile picture was the fact that it was apparently unique, as a reverse image search turned up no results.[58] But as with many other methods of self-defence against deception and disinformation, this type of detection relies on critical thinking about context, and on the intended victims of deepfake campaigns remaining both enquiring and sceptical. The Katie Jones case suggests that these qualities are far from universal among users of LinkedIn, and the same principles hold good for other online platforms, environments, and fora.

Even if detection becomes commonplace and rapid enough to be relevant in countering disinformation, critical decisions must still be made on how best to counter deepfakes and mitigate their impact.

At present, rather than pre-emption or live detection, proposed countermeasures appear to focus on forensics and after-event remedies, including potentially holding social media companies liable for harmful material disseminated over their platforms.[59] Making deepfakes illegal has also been proposed. While the principles behind use of deepfakes for deception may not be novel, the relatively new technology involved does raise new legal and juridical implications, including whether deepfakes can be legislated against on the grounds of falsehood,[60] and determining the evidential standards required.[61] In fact partial bans on deepfakes are already in place on specific platforms. Following democratisation of access to deepfake technology, for instance with the release of the deepfake creator app FakeApp in early 2018, 'numerous platforms including Twitter, Discord, Gfycat, and Pornhub explicitly banned deepfakes and associated communities. Gfycat in particular announced that it was using AI detection methods in an attempt to proactively police deepfakes'.[62] However, suggestions of a blanket ban on the technology overall seem impractical given the broad overlap with legitimate fields of study in audio and video synthesis and 3D-modelling, not to mention bona fide commercial applications. Attempts to limit access to training data are similarly unrealistic, given the constraints this would impose on news media and all forms of online presence.

The remaining realistic option is to continue the arms race: bolstering resilience to malign exploitation of deepfakes by educating the public, enhancing the resilience of video- and audio-based identification processes, and incorporating measures to identify and flag deepfakes online in real time.



Conclusion

Outlook

A number of key trends can be expected in the near and immediate future. The development and deployment of influence campaigns leveraging technology will accelerate still further, and machine learning algorithms enhancing profiles and interaction to build networks for commercial or malign purposes will go mainstream in a very short space of time. Meanwhile attitudes to deepfakes will remain confused and conflicted. Dramatic predictions of the consequences of their abuse for political purposes will continue, some justified and some overwrought. But in parallel, normalisation will also continue, driven by the increasing and widely accepted prevalence of virtual individuals, especially in marketing. One disturbing side-effect with unpredictable social consequences will be the continuing erosion of confidence in whether any online interaction is in fact with a real person.

The race between creation and detection of deepfakes will continue, with each side enjoying temporary advantage. Apparent success in developing detection techniques will on occasion provide false confidence that the problem has been solved, based on faith in the apotropaic powers of neural networks to detect and counter the phenomenon they themselves begat. But the realisation will develop that deepfakes are like terrorism; impossible to eradicate or resolve altogether, the answer is to find ways of living with them as a perennial problem and mitigating the most damaging likely outcomes. In another parallel with combatting terrorism, in creating and countering deepfakes the moral asymmetry will continue to favour the malign actor, with none of the constraints of rule of law to hamper their agility and ingenuity in devising new means to exploit the technology to do harm. As such, deepfakes will in time form a key element of how cyber-enabled information strategies are used in future war and hybrid conflict.[63]

Artificial systems will begin to assist or counter each other autonomously in real time. Automatic speaker verification systems have already been pitted against each other in simulation, as have sophisticated chatbots, in each case with disturbing results.[64] The continuing proliferation of machine-learning systems for generating content will require similar decisions regarding keeping a human in the loop as those already under discussion in consideration of AI-driven or autonomous combat or weapons systems. Meanwhile apps and platforms will continue to present themselves as neutral, but an increasing number of them will be developed and used

as tools of national competition, subversion, and espionage.[65]

Mainstream media awareness and popularisation of the term 'deepfake' will lead to definition creep, as precise and strictly bounded criteria for what can be termed a deepfake gives way to confusion in non-specialist discussion with simple edited audio, video, or still images.[66] But 'deepfake text', in the form of algorithmically-generated messages flooding recipients to give a false impression of political consensus, will present a further evolution of the manipulation of public discourse that will be conflated with other machine-learning enhancements for malign influence campaigns.[67] Of all forms of machine-enhanced deceptive content, text-based output is the first that will include interactions adjusted for the emotional state of the target, observing and analysing human responses and seeking the most effective method of influence through what is known in human-computer interface research as 'emotional modelling'.[68]

Even in the absence of dramatic involvement of deepfakes in causing political change or upheaval, long-term social implications may be profound. The more pervasive the present hype over deepfakes, the easier it becomes to claim that any legitimate information might in fact be doctored, with accusation and counter-accusation of fraud between disinformation spreaders and their debunkers.[69] This problem is of course not limited to deepfakes themselves, as disinformation researcher Renee DiResta notes: "whether it's AI, peculiar Amazon manipulation hacks, or fake political activism—these technological underpinnings [lead] to the increasing erosion of trust".[70] This points to a danger that user education in critical consumption of information may have an unintended consequence. If not managed carefully, emphasis on warnings that content online may be deceptive could contribute not only to this erosion of trust in genuine sources but also to the problem it seeks to address: the lack of belief in society at large in any form of objective truth.[71]

Policy recommendations

The challenge of information warfare is not a static situation, but a developing process. Adversary approaches evolve, develop, adapt, and identify successes to build on further. It follows that those nations and organisations that are preparing to counter currently visible threats and capabilities will find themselves out of date and taken by surprise by what happens next. Defences must instead be agile, alert to trends, and forward-thinking in how to parry potential future moves.

The deepfakes arms race will be a contest of agility and innovation. While it progresses, there are practical mitigation steps that can be taken. Pending the introduction of adequate defences against voice mimicry, individuals and corporations can review the extent of publicly available audio

recordings to assess whether a dataset is sufficient to generate fake voice interaction or authentication. Governments, NGOs, and think-tanks can adopt corporate attitudes on brand protection and compliance to increase awareness of who is purporting to represent them. Legal authorities can consider further whether and when deception carried out by means of deepfakes is, or should be, a criminal offence—and if so, which one. Social media platforms should continue to be challenged to address some of the most pernicious consequences of their laissez-faire attitude to hostile activity delivered across their networks.

But the most powerful defence against the possible pernicious influence of deepfakes remains the same as against malign influence campaigns overall: awareness, and an appropriately developed and well-informed threat perception. Individuals up and down the chain of command of any and all organisations should be briefed on the potential impact of a deepfake-enhanced attack and, as an adjunct to cyber security awareness campaigns, education for the general public should include accessible explanations of the nature and implications of deepfake technology.[72] Media organisations, especially national ones, should follow the example of Yle in Finland and produce their own demonstration deepfake videos, released under controlled circumstances, illustrating their potential to deceive in order to educate their audiences.[73] In particular, individuals should be reminded of the basic principle that any personal image or information posted publicly online can become hostage to abuse for nefarious purposes.[74] Katie Jones is a harbinger: it is in everyone's interest to ensure that no-one is taken by surprise by her inevitable multitude of successors.

Endnotes

1 Keir Giles and Kim Hartmann are Research Director and Cyber and Technology Director respectively at the Conflict Studies Research Centre in the UK. Munira Mustaffa is a former analyst with the Southeast Asia Regional Centre for Counter Terrorism (SEARCCT) in Kuala Lumpur, Malaysia, currently pursuing a PhD in Justice, Law & Criminology at the School of Public Affairs, American University, Washington DC.

2 E-mail to author from Raphael Satter, 11 June 2019.

3 Azreen Hani, 'The Lady Who Decoded the Deepfake Katie Jones', The Malaysian Reserve, 23 July 2019; Raphael Satter, 'Experts: Spy Used AI-generated Face to Connect With Targets', AP, 13 June 2019.

4 The example of 'Robin Sage', a persona created by security researchers in 2010, showed event organisers not carrying out even rudimentary due diligence before issuing fake profiles with invitations to conferences. See Thomas Ryan, Getting In Bed with Robin Sage, (Provide Security, 2010).

5 Adrian Yijie Xu, 'AI, Truth, and Society: Deepfakes at the Front of the Technological Cold War', Medium, 2 July 2019.

6 Although content of a pornographic nature did not form part of the research corpus for this investigation, the authors have been informed that a substantial volume of such material is available online for examination should readers wish to undertake further independent study.

7 Chris Baraniuk, 'How To Fake It In The Real World', New Scientist, Volume 239/3193 (1 September 2018): 5.

8 Victor Tangermann, 'Congress Is Officially Freaking Out About Deepfakes', Futurism.com, 13 June 2019.

9 Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman, 'Synthesizing Obama: Learning Lip Sync from Audio', ACM Transactions on Graphics, Volume 36/ 4, Article 95, July 2017.

10 Bill Posters, '"Imagine this..." (2019) Mark Zuckerberg reveals the truth about Facebook and who really owns the future', Instagram, 7 June 2019.

11 Tim Hwang, 'The Future of the Deepfake—And What It Means for Factcheckers', Poynter, 17 December 2018.

12 Robert Chesney and Danielle Keats Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21.

13 BBC News, 'Deepfake Videos Could "Spark" Violent Social Unrest', 13 June 2019.

14 Franklin Foer, 'The Era of Fake Video Begins', The Atlantic, May 2018.

15 Keir Giles, 'Hack and Fake (Facts): Warfare in the Information Space', Lennart Meri Conference, YouTube video, 13 May 2017.

16 Nathalie van Raemdonck, cyber expert at European Union Institute for Security Studies (EUISS), interview by the authors, 2 August 2019.

17 Beatrice Dupuy and Barbara Ortutay, 'Deepfake Videos Pose a Threat, But "Dumbfakes" May Be Worse', AP, 19 July 2019.

18 Kevin Poulsen, We Found the Guy Behind the Viral 'Drunk Pelosi' Video, The Daily Beast, 1 June 2019.

19 Jeffrey Kluger, "How That Viral Video of a White House Reporter Messes With Your Mind", Time, 8 November 2018, https://time.com/5449401/jim-acosta-cnn-trump-video/

20 Adi Robertson, 'A Million Facebook Users Watched a Video That Blurs the Line Between Bad Satire and "Fake News"', The Verge, 24 July 2018.

21 Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen, 'Can We Steal Your Vocal Identity from the Internet? Initial Investigation of Cloning Obama's Voice Using GAN, Wave Net and Low-quality Found Data', Odyssey 2018 The Speaker and Language Recognition Workshop, 26–29 June 2018.

22 Paige Leskin, 'Here's How to Use Duplex, Google's Crazy New Service that Impersonates a Human Voice to Make Appointments on Your Behalf', Business Insider, 2 December 2018.

23 Keir Giles, Russia's 'New' Tools for Confronting the West, (Chatham House, March 2016), p. 30.

24 Oscar Schwartz, 'Could "Fake Text" Be the Next Global Political Threat?', The Guardian, 4 July 2019.

25 Marco Rubio, Twitter post: Want to Know What a #Putin Disinformation Campaign Looks Like?, 19 June 2019.

26 Tim Hwang, 'A Vote Against Deepfakes', New Scientist, Volume 239/3190 (11 August 2018): 22–23; Zarine Kharazian, '360/OS 2019: Deepfakes—It's Not What It Looks Like!', Medium, 1 July 2019; Russell Brandom, 'Deepfake Propaganda is Not a Real Problem', The Verge, 5 March 2019.

27 James Vincent, 'Why We Need a Better Definition of "Deepfake"', The Verge, 22 May 2018.

28 Katie Baron, 'Digital Doubles: The Deepfake Tech Nourishing New Wave Retail', Forbes, 28 July 2019.

29 Tiffany Hsu, 'These Influencers Aren't Flesh and Blood, Yet Millions Follow Them', The New York Times, 17 June 2019.

30 Lily Kuo, 'World's First AI News Anchor Unveiled in China', The Guardian, 8 November 2018.

31 Jeff Stein, 'How Russia Is Using LinkedIn as a Tool of War Against Its U.S. Enemies', Newsweek, 3 August 2017.

32 Warren Strobel, Jonathan Landay, 'Exclusive: U.S. Accuses China of "Super Aggressive" Spy Campaign on LinkedIn', Reuters, 31 August 2018; Edward Wong, 'How China Uses LinkedIn to Recruit Spies Abroad', The New York Times, 27 August 2019.

33 Ken Dilanian, 'How a $230,000 Debt and a LinkedIn Message Led an Ex-CIA Officer to Spy for China', NBC News, 4 April 2019.

34 Mika Aaltola,'Geostrategically Motivated Co-option of Social Media: The Case of Chinese LinkedIn Spy Recruitment', Finnish Institute of International Affairs, FIIA Briefing Paper 267, 19 June 2019.

35 BBC News, 'German Spy Agency Warns of Chinese LinkedIn Espionage', 10 December 2017.

36 Further evidence of the problem this presents came in September 2019 with another well-publicised instance of a profile being set up for a fake employee at a prestigious US think-tank. See Robert K. Knake, "Hey LinkedIn, Sean Brown Does Not Work at CFR: Identity, Fake Accounts, and Foreign Intelligence", Council on Foreign Relations, 10 September 2019.

37 Christopher Burgess, 'Who's in Your Social Network? Why You Can't Always Trust Your Online Friends', ClearanceJobs, 11 April 2019.

38 'Vorsicht bei Kontaktaufnahme über Soziale Netzwerke!' [Be careful with contacts on social networks!], Bundesamt für Verfassungsschutz (BfV) [Federal Office for the Protection of the German Constitution], 3 July 2017.

39 Scott Ikeda, 'The Cutting Edge of AI Cyber Attacks: Deepfake Audio Used to Impersonate Senior Executives', CPO Magazine, 18 July 2019.

40 Kevin D. Mitnick and William L. Simon, The Art of Deception: Controlling the Human Element of Security, (Indianapolis: Wiley Publishing Inc., 2002).

41 Grothaus, M., 'Criminals Are Using Deepfakes to Impersonate CEOs', Fast Company, 19 July 2019; BBC News, 'Fake Voices "Help Cyber-crooks Steal Cash"', 8 July 2019.

42 Sjouwerman, S., 'Deepfake Videos—An Increasing Cyber Threat for Corporate Clients', KnowBe4, 6 May 2019.

43 E-mail to author, 2 April 2019.

44 Elizabeth Culliford, 'House Intelligence Chief Presses Social Media Companies on Deepfake Policies', Reuters, 15 July 2019.

45 Charlie Warzel, 'Why Can Everyone Spot Fake News But YouTube, Facebook, And Google?', 22 February 2018.

46 Dean Obeidallah, 'How Russian Hackers Used My Face to Sabotage Our Politics and Elect Trump', Daily Beast, 28 September 2017.

47 Mike Allen, 'How Big Tech is Prepping for Russian Propaganda Backlash', Axios, 26 September 2017.

48 'Internet Trolling as A Tool Of Hybrid Warfare: The Case Of Latvia', NATO StratCom Centre of Excellence, 2015.

49 Katyanna Quach, "Deepfake 3.0 (beta), the bad news: This AI can turn ONE photo of you into a talking head. Good news: There is none", The Register, 19 June 2019.

50 Hannah Durevall, 'Growing Demand for Biometric Security in Banking', Mapa Research, 23 January 2017.

51 Pavel Korshunov and Sebastien Marcel, 'DeepFakes: A New Threat to Face Recognition? Assessment and Detection', unpublished paper, 20 December 2018.

52 Karen Hao, 'Yes, FaceApp Could Use Your Face—But Not For Face Recognition', MIT Technology Review, July 2019.

53 Yuezun Li, Siwei Lyu, 'Exposing DeepFake Videos By Detecting Face Warping Artifacts', Computer Vision Foundation, State University of New York at Albany, 1 November 2018.

54 Mark Anderson, 'A Two-Track Algorithm to Detect Deepfake Images', Spectrum.leee.org, 29 July 2019.

55 David Guera and Edward Delp, 'Deepfake Video Detection Using Recurrent Neural Networks', 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 27–30 November 2018.

56 Yuezun Li, Ming-Ching Chang and Siwei Lyu, 'In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking', State University of New York at Albany, 11 June 2018.

57 'Special Notice: Semantic Forensics (SemaFor) Proposers Day', DARPA, 5 August 2019.

58 Katie Jones was suspicious for the opposite reason to 'Robin Sage', whose profile pictures were easily identifiable as lifted from pornographic websites.

59 Abigail Summerville, ' "Deepfakes" Trigger a Race to Fight Manipulated Photos and Videos', Wall Street Journal, 27 July 201.

60 Marc Jonathan Blitz, 'Lies, Line Drawing, and (Deep) Fake News', Oklahoma Law Review, Vol. 71/59 (2018): 59–116.

61 Marie-Helen Maras, Alex Alexandrou, 'Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and In the Wake of Deepfake Videos', The International Journal of Evidence & Proof, Vol. 23/3 (2019): 255–62.

62 Benjamin Goggin, 'From Porn to "Game of Thrones": How Deepfakes and Realistic-looking Fake Videos Hit It Big', Business Insider, 23 June 2019.

63 Martti Lehto, 'The Modern Strategies in the Cyber Warfare' [sic] in Cyber Security: Power and Technology, Martti Lehto and Pekka Niettaanmäki (eds), (Springer, 2018), pg. 3–20.

64 Ville Vestman, Tomi Kinnunena, Rosa González Hautamäki, and Md Sahidullah, 'Voice Mimicry Attacks Assisted by Automatic Speaker Verification', Computer Speech and Language, Volume 59 (January 2020): 36–54. Samuel Bendett, "What Russian Chatbots Think About Us", Defense

One, 2 September 2019.

65 Samuel Scott, 'A Look at FaceApp, TikTok and the Rise of "Data Nationalism"', The Drum, 23 July 2019.

66 Vincent, 'Why We Need a Better Definition of "Deepfake"'.

67 Charlie Warzel, 'He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse', Buzzfeed, 11 February 2018.

68 Ingo Siegert, Kim Hartmann et al., 'Modelling of Emotional Development within Human-Computer-Interaction', Kognitive Systeme, 2013-1.

69 'The Unreal Deal', New Scientist, Vol. 239/3193 (1 September 2018): 3–57.

70 Warzel, 'He Predicted The 2016 Fake News Crisis'.

71 James Ball, 'What do we do when everything online is fake?', The World Today, June & July 2019, Chatham House

72 Summer Hirst, 'Deepfakes—Seeing is Believing. Or Not?', Surfshark blog, 30 July 2019; Goggin, 'From Porn to "Game of Thrones"'.

73 "Tämän jutun jälkeen katsot liikkuvaa kuvaa uusin silmin: Yle teki deepfake-videon, jolla Sauli Niinistö haaveilee kolmannesta kaudesta" (After this you will see moving pictures with new eyes: Yle made a deepfake video where [President] Sauli Niinistö dreams of a third term), Yle Uutiset, 6 September 2019.

74 Kara Swisher, 'Does Russia Want More Than Your Old Face?', The New York Times, 19 July 2019.

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel.
Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org