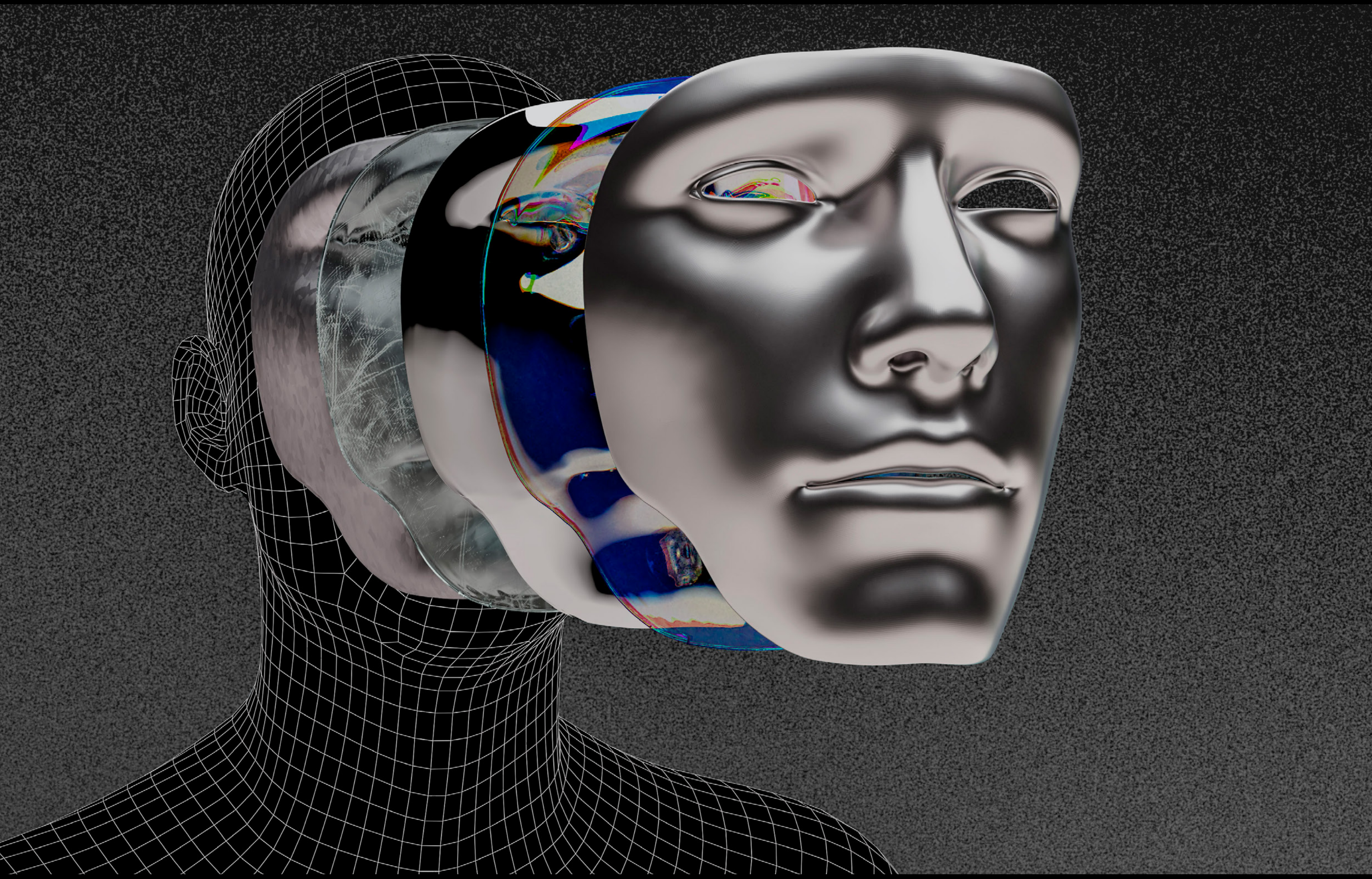
Weaponised deep fakes

National security and democracy

Hannah Smith and Katherine Mansted



ASPI AUSTRALIAN STRATEGIC POLICY INSTITUTE

INTERNATIONAL CYBER POLICY CENTRE

Policy Brief
Report No. 28/2020

About the authors

Hannah Smith is a Researcher working with the International Cyber Policy Centre.

Katherine Mansted is senior adviser for public policy at the Australian National University's National Security College, and a non-resident fellow at the Alliance for Securing Democracy at the German Marshall Fund of the United States, and the Harvard Kennedy School's Belfer Center for Science and International Affairs.

Acknowledgements

The authors would like to thank the support of the National Security College at the Australian National University.

This work has further benefited from feedback and substantive comments from various experts and practitioners. The authors would like to thank the anonymous peer reviewers for their valuable feedback on report drafts.

What is ASPI?

The Australian Strategic Policy Institute was formed in 2001 as an independent, non-partisan think tank. Its core aim is to provide the Australian Government with fresh ideas on Australia's defence, security and strategic policy choices. ASPI is responsible for informing the public on a range of strategic issues, generating new thinking for government and harnessing strategic thinking internationally.

ASPI International Cyber Policy Centre

ASPI's International Cyber Policy Centre (ICPC) is a leading voice in global debates on cyber and emerging technologies and their impact on broader strategic policy. The ICPC informs public debate and supports sound public policy by producing original empirical research, bringing together researchers with diverse expertise, often working together in teams. To develop capability in Australia and our region, the ICPC has a capacity building team that conducts workshops, training programs and large-scale exercises both in Australia and overseas for both the public and private sectors. The ICPC enriches the national debate on cyber and strategic policy by running an international visits program that brings leading experts to Australia.

Important disclaimer

This publication is designed to provide accurate and authoritative information in relation to the subject matter covered. It is provided with the understanding that the publisher is not engaged in rendering any form of professional or other advice or services. No person should rely on the contents of this publication without first obtaining advice from a qualified professional.

ASPI

Tel +61 2 6270 5100
Email enquiries@aspi.org.au
www.aspi.org.au
www.aspistrategist.org.au
facebook.com/ASPI.org
@ASPI_ICPC

First published April 2020.

ISSN 2209-9689 (online)
ISSN 2209-9670 (print)

Cover image: Produced by Leslie Sharp, www.lesliesharpe.net

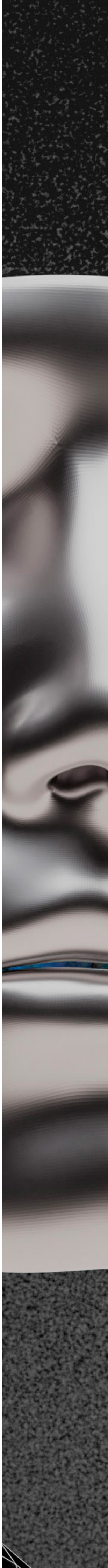
ASPI
AUSTRALIAN
STRATEGIC
POLICY
INSTITUTE

INTERNATIONAL
CYBER POLICY
CENTRE

Weaponised deep fakes

National security and democracy

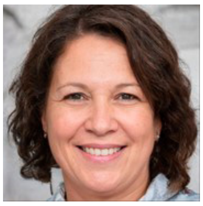
Hannah Smith and Katherine Mansted



Policy Brief
Report No. 28/2020

Contents

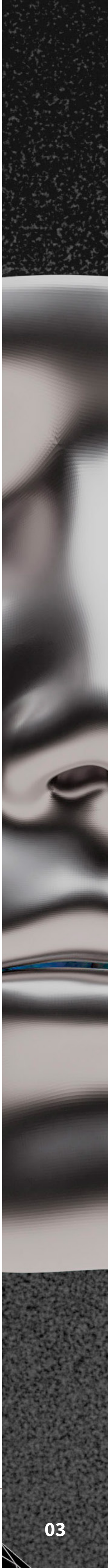
Foreword



Fakes are all around us. Academic analysis suggests that they're difficult to spot without new sensors, software or other specialised equipment, with 1 in 5 photos you see being fraudulent. The exposure of deep fakes and the services they facilitate can potentially lead to suppression of information and a general breakdown in confidence in public authorities and trust. We need to react not just to false or compromised claims but to those who would try to exploit them for nefarious purposes. We should not assume the existence of fake news unless we have compelling evidence to the contrary, but when we do, we should not allow the propaganda. I've never been more sure of this point than today.

—GPT-2 deep learning algorithm

The foreword to this report was written by a machine. The machine used a 'deep fake' algorithm—a form of artificial intelligence (AI)—to generate text and a headshot. Deep fakes are increasingly realistic and easy to create. The foreword took us approximately five minutes to generate, using free, open-source software.[1]



What’s the problem?

Deep fake technology isn’t inherently harmful. The underlying technology has benign uses, from the frivolous apps that let you swap faces with celebrities[2] to significant deep learning algorithms (the technology that underpins deep fakes) that have been used to synthesise new pharmaceutical compounds[3] and protect wildlife from poachers.[4]

However, ready access to deep fake technology also allows cybercriminals, political activists and nation-states to quickly create cheap, realistic forgeries. This technology lowers the costs of engaging in information warfare at scale and broadens the range of actors able to engage in it. Deep fakes will pose the most risk when combined with other technologies and social trends: they’ll enhance cyberattacks, accelerate the spread of propaganda and disinformation online and exacerbate declining trust in democratic institutions.

What’s the solution?

Any technology that can be used to generate false or misleading content, from photocopiers and Photoshop software to deep fakes, can be weaponised. This paper argues that policymakers face a narrowing window of opportunity to minimise the consequences of weaponised deep fakes. Any response must include measures across three lines of effort:

1. investment in and deployment of deep fake detection technologies
2. changing online behaviour, including via policy measures that empower digital audiences to critically engage with content and that bolster trusted communication channels.
3. creation and enforcement of digital authentication standards

What's a deep fake?

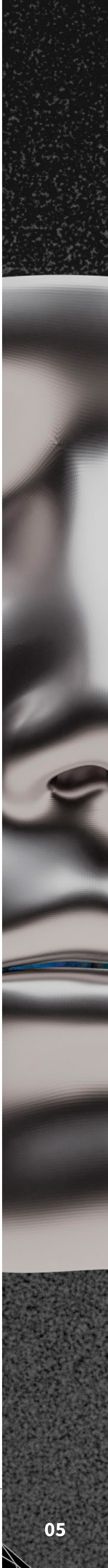
A deep fake is a digital forgery created through deep learning (a subset of AI).[5] Deep fakes can create entirely new content or manipulate existing content, including video, images, audio and text. They could be used to defame targets, impersonate or blackmail elected officials and be used in conjunction with cybercrime operations.

Some of the first public examples of deep fakes occurred in November 2017, when users of the popular online message-board Reddit used AI-based 'face swap' tools to superimpose celebrities' faces onto pornographic videos.[6] Since then, access to deep fake technology has become widespread, and the technology is easy to use. Free software and trending smartphone applications such as FaceSwap or Zao[7] allow everyday users to create and distribute content. Other services can be accessed at low cost: the Lyrebird voice generation service, for instance, offers subscription packages for its tools. In short: deep fake technology has been democratised.

Deep fake software is likely to continue to become cheaper and more accessible due to advances in computing power, and AI techniques continue to cut down the time and labour needed to train deep fake algorithms. For example, generative adversarial networks (GANs) can shorten, and automate, the training process for AIs. In this process, two neural networks compete against one another to produce a deep fake. A 'generator' network creates fake content. A 'discriminator' network then attempts to assess whether the content is authentic or fake. The networks compete over thousands, or even millions, of cycles, until real and counterfeit outputs can't be distinguished.[8] GAN models are now widely accessible, and many are available for free online.

The deep fake advantage

Not all digital forgeries are deep fakes. Forgeries created by humans using software editing tools are often called 'cheap fakes' (see box). Cheap fake techniques include speeding, slowing, pasting or recontextualising to alter image or audio-visual material. *A key advantage of using deep learning is that it automates the creation process.* This allows for realistic (or 'good enough') content to be quickly created by users with very little skill. Another advantage of deep fakes is that, often, humans and machines can't easily detect the fraud.[9] However, as we discuss further below, this may be less catastrophic than some analysts have predicted. Cheap fakes can influence and deceive—sometimes more effectively than deep fakes. Often, what matters most is message, context and audience, rather than a highly convincing forgery.



Deep or cheap?

In May 2019, a video circulated on social media showing US House of Representatives Speaker Nancy Pelosi slurring her words during a news conference, as though she were intoxicated or unwell. The video was a cheap fake: an authentic recording of the speaker, but with the speed slowed to 75% and the pitch adjusted to sound within normal range.[10]

Similarly, in November 2018, the far-right conspiracy website *InfoWars* disseminated a video edited to make it look like CNN journalist Jim Acosta was acting aggressively towards staff.

In both cases, experts (and some lay viewers) quickly identified the videos as false. Nonetheless, they had impact. The Pelosi video went 'viral' and was used by her political opponents to bolster a narrative that she was unfit to serve as the Speaker. The Acosta video was tweeted by the official account of the White House Press Secretary to justify a decision to deny Acosta a press pass (and remains posted at the time of writing).[11]

Audio-visual cheap fakes even pre-date the digital age. In the lead-up to UK elections in 1983, members of the British anarcho-punk band Crass spliced together excerpts from speeches by Margaret Thatcher and Ronald Reagan to create a fake telephone conversation between the leaders, in which they each made bellicose, politically damaging statements.

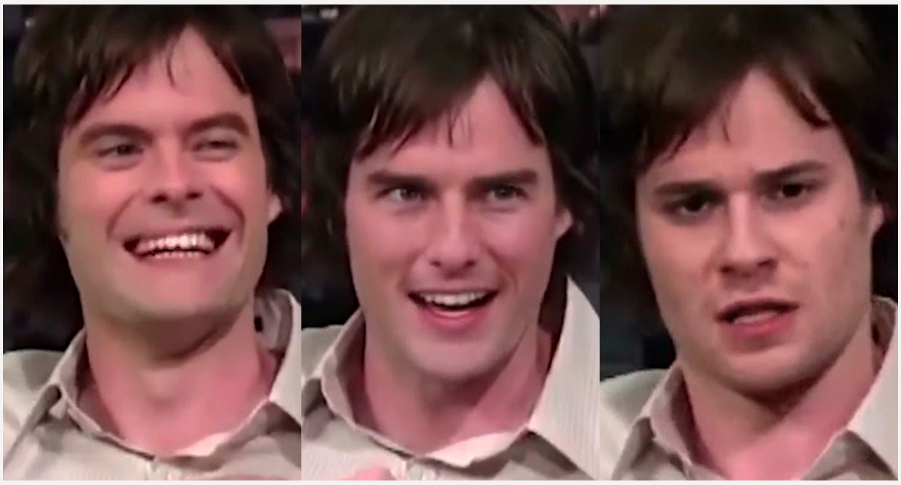
Common deep fake examples

Deep fake processes can be applied to the full spectrum of digital media. Below, we describe seven common deep fake tools. This isn't an exhaustive list; nor are the categories exclusive. Deep fakes are often amalgams of several tools.

1. Face swapping

Users insert the face of a target onto another body. This process can be applied to both still images and video. Simple versions of this technique are available online through purpose-made apps.

Figure 1: Deep fake video of actor and comedian Bill Hader morphing into different characters during an impression monologue

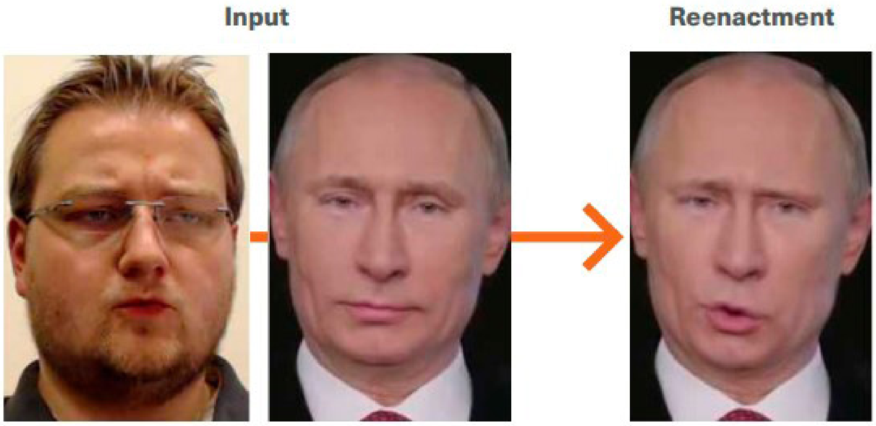


Source: 'Bill Hader channels Tom Cruise [DeepFake]', *YouTube*, 6 August 2019, online.

2. Re-enactment

The face from a target source is mapped onto a user, allowing the faker to manipulate the target's facial movements and expressions.

Figure 2: Researchers use Face2Face tool to control the facial movements of Vladimir Putin

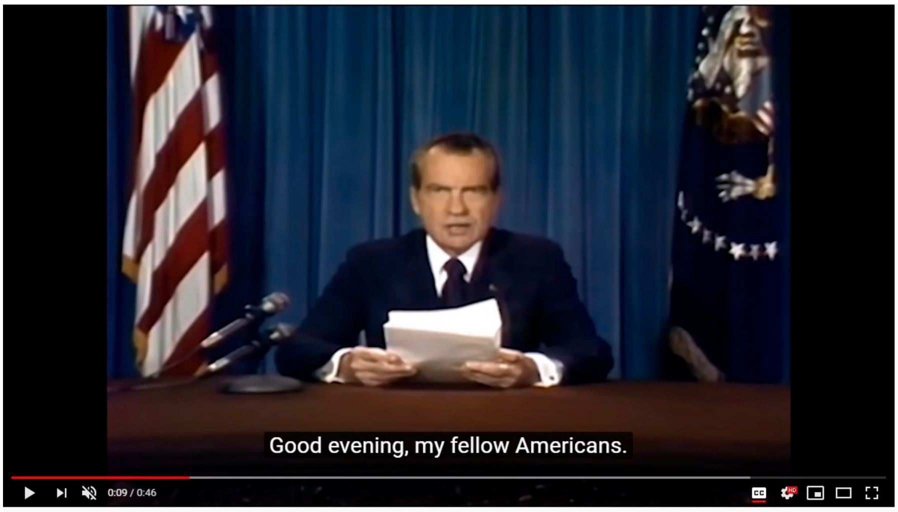


Source: TUM visual computing lab. Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, Matthias Nießner, 'Face2Face: Real-time face capture and reenactment of RGB Videos', *Graphics*, Stanford University, 2016, online.

3. Lip syncing

Users copy mouth movements over a target video. Combined with audio generation, this technique can make a target appear to say false content.

Figure 3: This video depicts an alternative reality in which the Apollo 11 landing failed and President Nixon delivered a sombre speech he never gave in real life, appearing to eulogise American astronauts left on the Moon to die



In Event of Moon Disaster - Nixon Deepfake Clips

Source: Suzanne Day, 'MIT art installation aims to empower a more discerning public', *MIT News*, 25 November 2019, online.

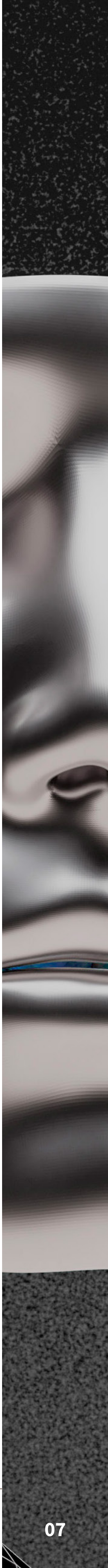
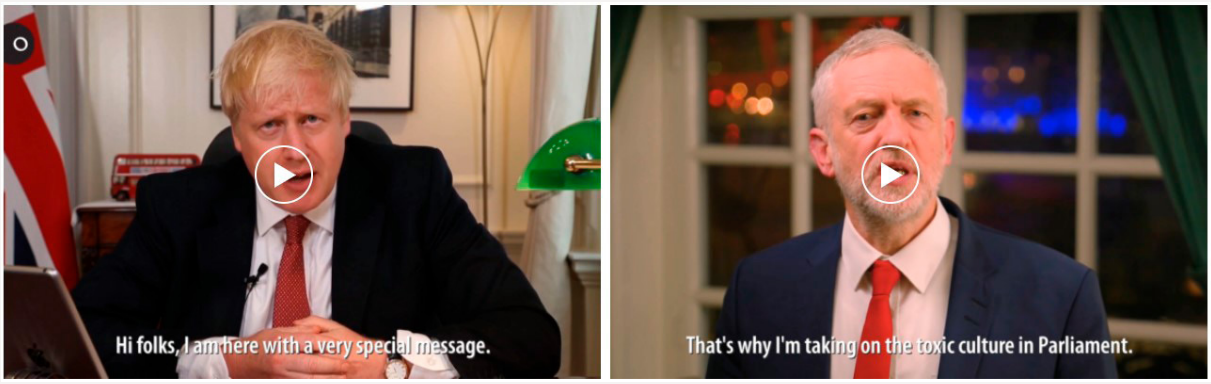


Figure 4: A video produced by AI think tank Future Advocacy depicts UK politicians Jeremy Corbin and Boris Johnson endorsing each other as the preferred candidate for the 2019 UK election

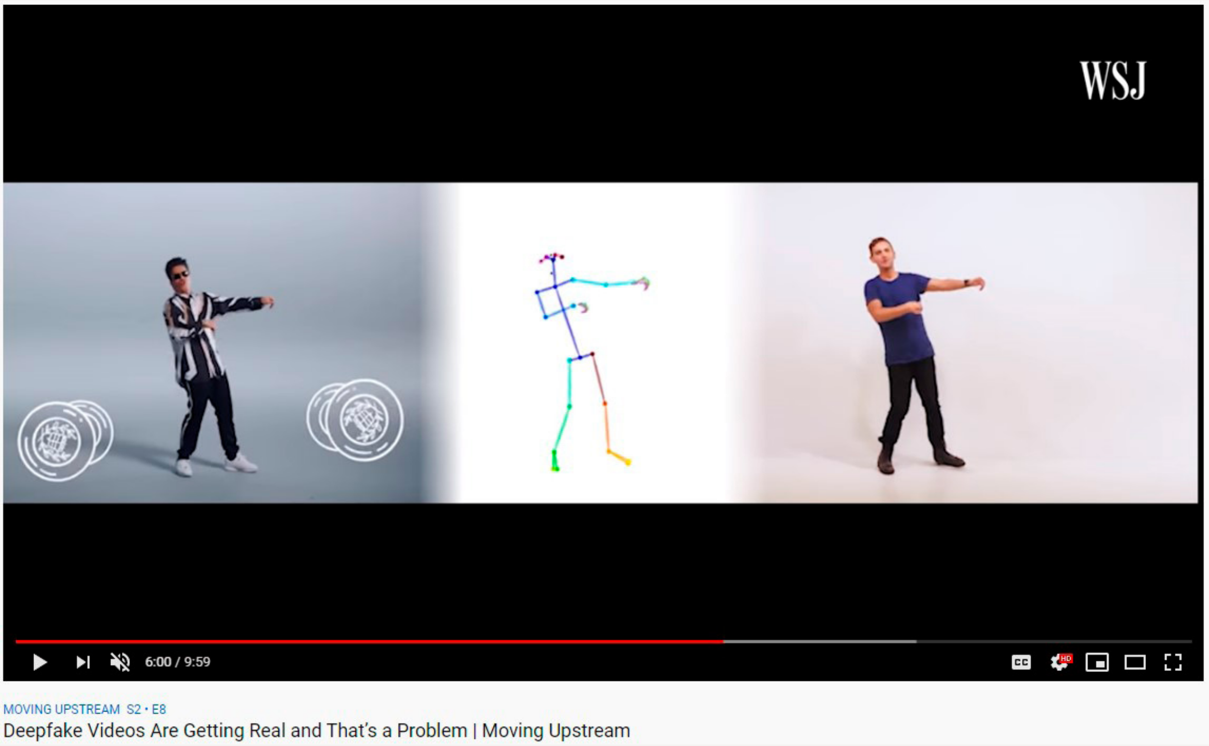

Hi folks, I am here with a very special message.
That's why I'm taking on the toxic culture in Parliament.

Source: ‘Deepfakes’, *Future Advocacy*, 2018, online.

4. Motion transfer

The body movements of a person in a source video can be transferred to a target in an authentic video recording.

Figure 5: Video depicts artist Bruno Mars dance routine mapped to a *Wall Street Journal* reporter through motion transfer technology



MOVING UPSTREAM S2 • E8

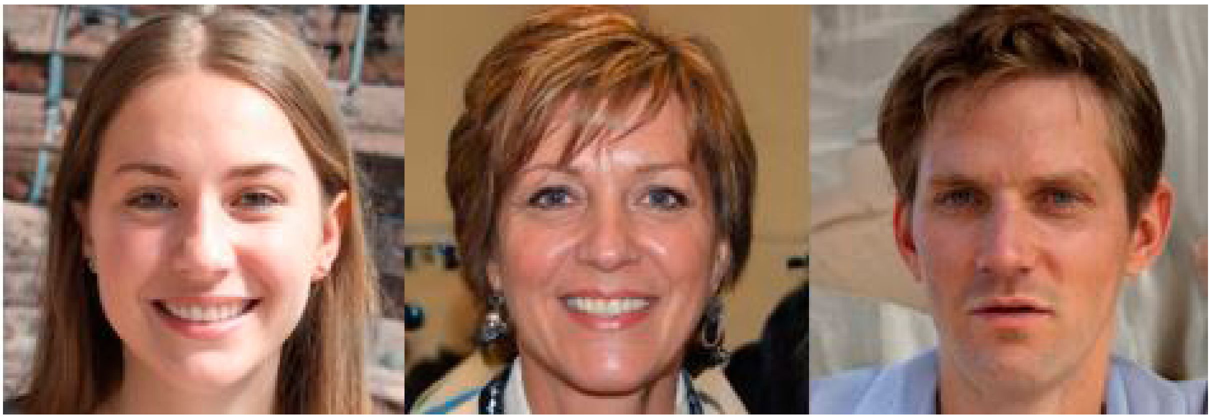
Deepfake Videos Are Getting Real and That’s a Problem | Moving Upstream

Source: Hilke Schellmann, ‘Deepfake videos are getting real and that’s a problem’, *Wall Street Journal*, 15 October 2018, online.

5. Image generation

A user can create entirely new images; for example, faces, objects, landscapes or rooms.

Figure 6: Three portraits created for the purposes of this report by a deep fake generator

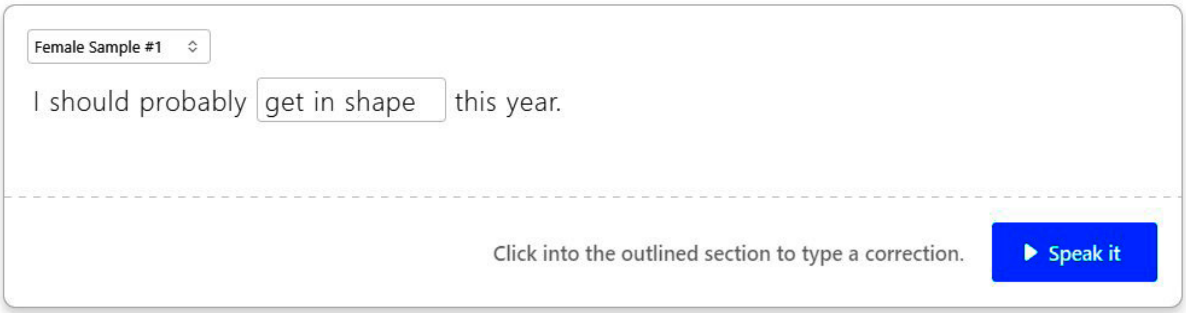


Source: 'This person does not exist', online.

6. Audio generation

Users create a synthesised voice from a small audio sample of an authentic voice. This technique can be combined with lip-sync tools, allowing users to 'overdub' audio into pre-existing clips.

Figure 7: Overdub software allows users to replace recorded words or phrases with typed phrases



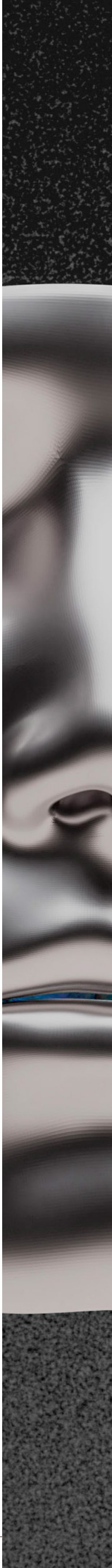
I should probably get in shape this year.

Click into the outlined section to type a correction.

Source: 'Lyrebird: Ultra-realistic voice cloning and text to speech', online.

Figure 8: A voice clone created from a small audio sample by Lyrebird voice double software

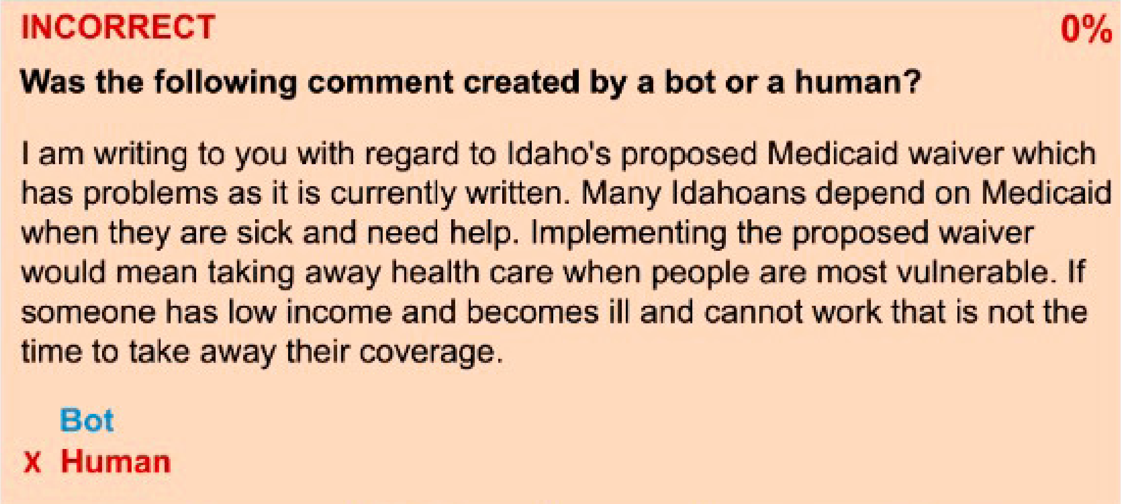
Source: 'Lyrebird: Ultra-realistic voice cloning and text to speech', online.



7. Text generation

A user can generate artificial text, including short-form 'comments' on social media or web forums, or long-form news or opinion articles. Artificially generated comments are particularly effective, as there's a wide margin for acceptable error for this type of online content

Figure 9: Deep fake text generated by researchers in a study monitoring responses to Idaho's Medicaid waiver; all study participants believed this response was of human origin



INCORRECT 0%

Was the following comment created by a bot or a human?

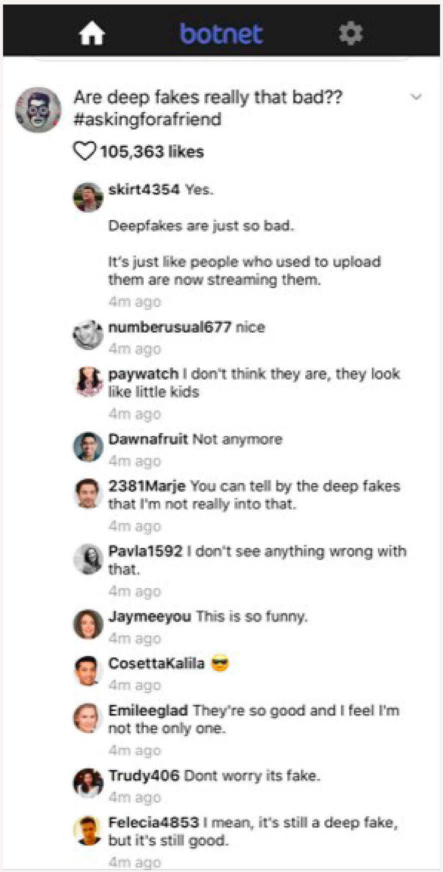
I am writing to you with regard to Idaho's proposed Medicaid waiver which has problems as it is currently written. Many Idahoans depend on Medicaid when they are sick and need help. Implementing the proposed waiver would mean taking away health care when people are most vulnerable. If someone has low income and becomes ill and cannot work that is not the time to take away their coverage.

Bot

X Human

Source: Max Weiss, 'Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions', *Technology Science*, 18 December 2019, online.

Figure 10: 'Botnet', a self-described social network simulator app, allows a single user to interact with fake comments generated by bots, who like and engage with the user's posts



Are deep fakes really that bad?? #askingforafriend

♡ 105,363 likes

skirt4354 Yes.

Deepfakes are just so bad.

It's just like people who used to upload them are now streaming them.

4m ago

numberusual677 nice

4m ago

paywatch I don't think they are, they look like little kids

4m ago

Dawnafruit Not anymore

4m ago

2381Marje You can tell by the deep fakes that I'm not really into that.

4m ago

Pavla1592 I don't see anything wrong with that.

4m ago

Jaymeeyou This is so funny.

4m ago

CosettaKalila 😎

4m ago

Emileeglad They're so good and I feel I'm not the only one.

4m ago

Trudy406 Dont worry its fake.

4m ago

Felecia4853 I mean, it's still a deep fake, but it's still good.

4m ago

Source: The Botnet social network simulator uses the open-source 'GPT-2' deep learning algorithm developed by California-based research lab OpenAI, online.

Weaponised deep fakes

Deep fake technology is not inherently dangerous. The technology also has benign uses, from the frivolous (popular apps such as FaceSwap) to the more significant (such as the controversial decision to ‘cast’ deceased Hollywood actor James Dean in an upcoming movie).[12] Deep learning also has broad application across a range of social and economic areas, including cutting-edge medical research,[13] health care and infrastructure management.[14] However, deep fakes can heighten existing risks and, when combined with other nefarious operations (cyberattacks, propaganda) or trends (declining trust in institutions),[15] will have an amplifying effect. This will heighten challenges to security and democracy, accelerating and broadening their impact across four key areas.

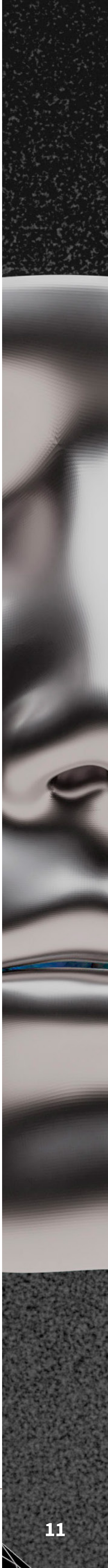
1. Cyber-enabled crime

Deep fakes will provide new tools to cyberattackers. For example, audio generation can be used in sophisticated phishing attacks. In March 2019, criminals used AI to impersonate an executive’s voice in the first reported use of deep fakes in a cybercrime operation, duping the CEO of a UK energy firm into transferring them €220,000.[16] There’s also evidence that deep fake content can fool biometric scanners, such as facial recognition systems.[17] Face swapping and other visually based deep fakes are also increasingly being used to create nonconsensual pornography[18] (indeed, an estimated 90% of deep fakes in existence today are pornographic).[19] As deep fake technology proliferates, we should also expect it to be used in acts of cyber-enabled economic sabotage. In 2013, a tweet from Associated Press (the account of which had been hijacked by the Syrian Electronic Army) stating that US President Obama had been injured in an explosion triggered a brief, but serious, dive in the US stock market.[20] While this example is political in nature, a more convincing fraud (imagine a deep fake video of the alleged explosion) could prove extremely damaging when paired with criminal operations.

2. Propaganda and disinformation

Online propaganda is already a significant problem, especially for democracies,[21] but deep fakes will lower the costs of engaging in information warfare at scale and broaden the range of actors able to engage in it. Today, propaganda is largely generated by humans, such as China’s ‘50-centers’ and Russian ‘troll farm’ operators. However, improvements in deep fake technology, especially text-generation tools, could help take humans ‘out of the loop’.[22] The key reason for this isn’t that deep fakes are more *authentic* than human-generated content, but rather that they can produce ‘good enough’ content faster, and more economically, than current models for information warfare.

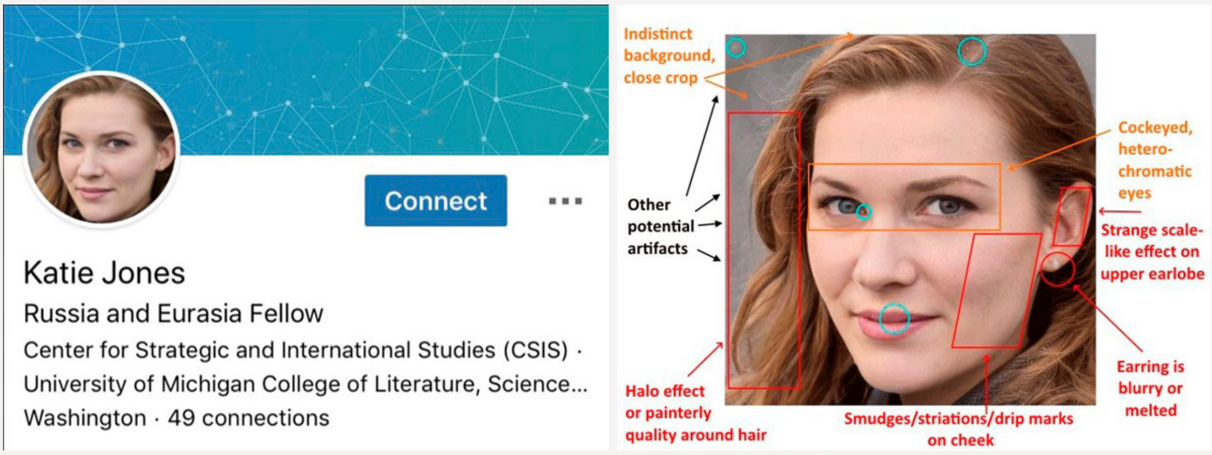
Deep fake technology will be a particular value-add to the so-called Russian model of propaganda, which emphasises volume and rapidity of disinformation over plausibility and consistency in order to overwhelm, disorient and divide a target.[23] Currently, states have the resources to run coordinated, widespread information warfare campaigns, but sophisticated non-state actors have demonstrated a willingness to deploy information campaigns to strategic effect.[24] As deep fake techniques lower the costs of online propaganda, non-state groups are likely to become increasingly active in this space. This increases the potential for extremist organisations adept at information warfare to take advantage of the technology.



Of particular concern is the use of automatic text generation to produce false online engagement, such as 'comments' on news articles, forums and social media. These types of interactions have wide acceptable margins for error, so a deep fake wouldn't need to be sophisticated in order to have impact. Russia's Internet Research Agency, a St Petersburg-based troll farm, had a monthly budget of approximately $US1.25 million for interference in American politics in the lead-up to the US 2016 presidential election,[25] while its workers allegedly face a gruelling schedule: 12-hour shifts with daily quotas of 135 posted comments of at least 200 characters.[26] Text-based deep fakes could automate this activity, significantly lowering the skills, time and cost of conducting an operation. AI-generated text would also be able to 'game' social media and search engine trending algorithms, which preference content based on popularity and engagement. This method is already leveraged in Russian influence campaigns.[27]

Deep fakes can also be layered into propaganda campaigns to make them more effective. For example, online propaganda often uses fake accounts and 'bots' to amplify content. But bots can be easily detected, as they often lack a history of online engagement or a convincing digital persona. Deep fake generated images and text can help bridge that gap. In 2019, journalists discovered that intelligence operatives had allegedly created a false LinkedIn profile for a 'Katie Jones', probably to collect information on security professional networks online. Researchers exposed the Katie Jones fake through technical photo analysis and a rather old-fashioned mechanism: asking the employer listed on LinkedIn (the Center for Strategic and International Studies) if such a person worked for it.[28] Importantly, deep fakes don't need to be undetectable to provide a benefit to agents of propaganda. They merely need to be 'good enough' to add extra layers of plausibility to a deceptive message.

Figure 11: Image of deep fake generated LinkedIn profile used in suspected intelligence-gathering operation



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

Connect
Indistinct background, close crop
Other potential artifacts
Halo effect or painterly quality around hair
Smudges/striations/drip marks on cheek
Cockeyed, hetero-chromatic eyes
Strange scale-like effect on upper earlobe
Earring is blurry or melted

Source: Raphael Satter, 'Experts: Spy used AI-generated face to connect with targets', *AP News*, 14 June 2019, online.

Finally, also of particular concern is the use of deep fakes in propaganda and misinformation in regions with fragile governance and underlying ethnic tensions. Misleading content spread via social media, such as decontextualised photos and false claims, has fuelled ethnic violence and killings in countries including India, Myanmar and Sri Lanka.[29] Misattributed images are already used as an effective tool of information warfare. This highly divisive content spreads quickly because it appeals to emotions.

3. Military deception and international crises

Concern about deep fakes often focuses on the fear of sophisticated forgeries that are of high enough quality to pass inspection even by an expert audience. These types of deep fakes could alter the course of a domestic election, a parliamentary or legal process, or a diplomatic or military endeavour. However, this is unlikely to occur as an informed, expert audience is more likely to:

- use available detection tools
- seek corroborating evidence
- assess evidence in the light of its source and context
- deliberate before acting on content.

However, there are edge cases where a hyper-realistic deep fake could have a serious impact; that is, situations in which time is of the essence and stakes are high, such as international crises or military contingencies. Forged audio-visual content could be used to degrade military commanders' situational awareness (either by *constructing* 'facts' on the ground or by manipulating legitimate data streams to obscure real facts). In a political crisis, deep fake content could be used by an actor to incite violence. Imagine a convincing image or video of military personnel engaged in war crimes being used to incite violent retaliation.[30]

4. Erosion of trust in institutions

In May 2018, Belgium's Socialistische Partij Anders became the first political party to use deep fake technology to influence public debate. The party posted a video to Facebook allegedly showing US President Trump encouraging Belgium to withdraw from the Paris Agreement on climate change.[31] According to the party, the video was designed to spark debate, not dupe: the lip-syncing was imperfect, it included a disclaimer stating that it was fake,[32] and it was quickly debunked by online communities and news sites. There's no evidence that the deep fake affected the Belgian election.

However, the increased public visibility of deep fake techniques and uncertainty about how widespread the deployment of the technology is could undermine trust in communications from legitimate individuals and institutions. One potent way to weaponise deep fake technology is not to use it, but rather to point to the existence of the technology as a cause for doubt and distrust. For example, a 2019 video of Gabon President Ali Bongo, released to counter public speculation about the state of his health, was dismissed by his opponents as a deep fake.[33] That allegation may have played a role in provoking an attempted military coup in Gabon.[34]

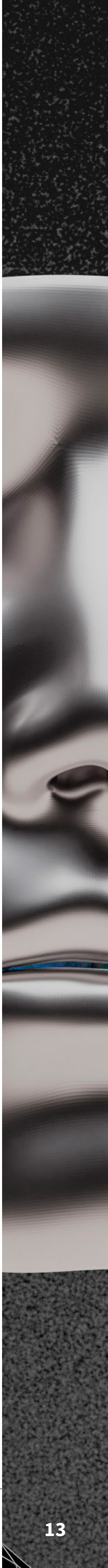
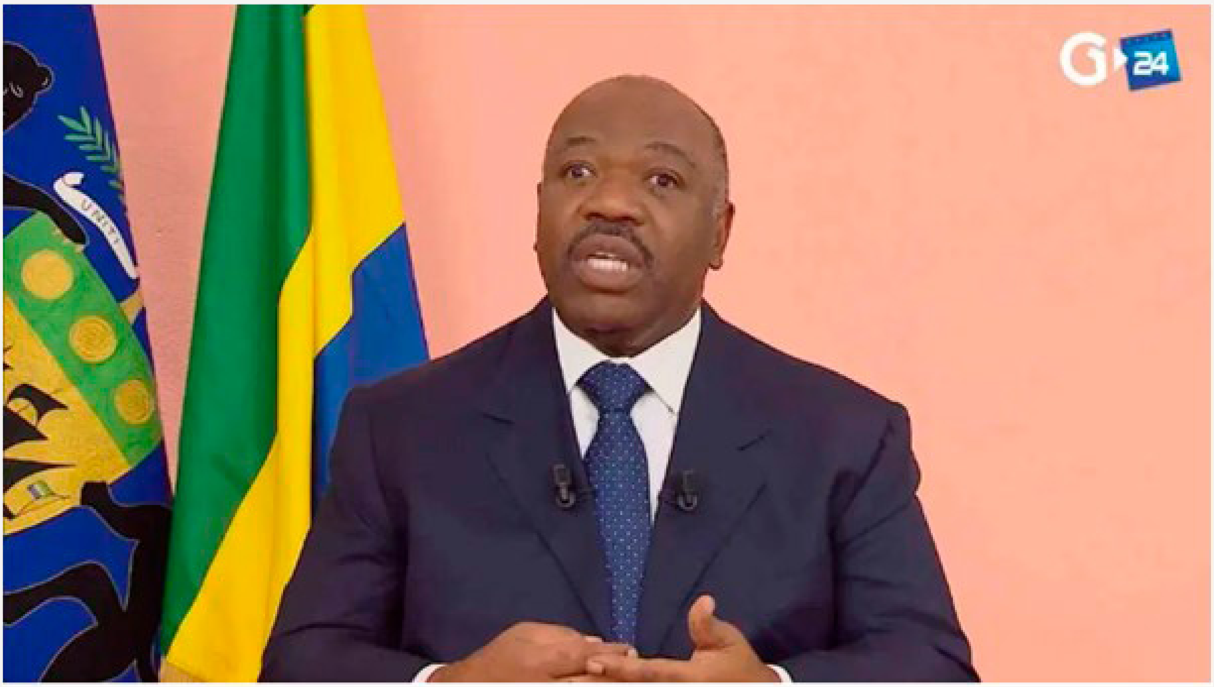


Figure 12: Address by Gabon's President Ali Bongo, which was falsely alleged to be a deep fake



Source: 'Gabon 24', *Facebook*, 31 December 2018, online.

This dynamic is exacerbated by what researchers term the 'liar's dividend': that is, efforts to debunk misinformation or propaganda can make it more difficult for audiences to trust all sources of information. This underscores the need for effective policy responses to weaponised deep fakes. Governments must act early to reassure the public that they're responding to the challenges of weaponised deep fakes, lest panic or credulity outstrip the impact of the fakes.

Recommendations

To address the challenges of weaponised deep fakes, policymakers should work closely with industry to pursue three lines of effort. Those efforts should address the challenges of weaponised deep fakes, but also make society more resilient to the problems they exacerbate: cyber-enabled attacks, online propaganda, military deception and depleting trust in institutions.

1. Detection technologies

Tools are available to detect some deep fake processes.[35] However, on balance, detectors are losing the 'arms race' with creators of sophisticated deep fakes.[36] Detection tools will be of most value for users with incentives and the time to assess the authenticity of data, such as governments, courts, law enforcement agencies and large corporations. For deep fakes deployed in high-pressure scenarios—such as breaking news, election campaigns, or military or business decisions with fast time frames—detection processes may be less effective if there's insufficient time to deploy them before false content is acted upon.

Detection won't fully mitigate the use of deep fakes in online disinformation (where 'good enough' is often sufficient to persuade) and misinformation, which tend to be fuelled by emotion and the speed of propagation rather than reason. Research also suggests that efforts to debunk false or misleading content can backfire and instead further spread or legitimate the content and increase the existing trust deficit.[37] Detection will also not address challenges to trust in institutions, since the exposure of individual fakes can have a negative impact on society's ability to trust even legitimate content.[38] That said, automatic detection tools that result in more consistent, principled labelling and flagging of content for review online (especially in the context of electoral advertising and political claims) may help reduce the effectiveness of deep fakes in propaganda and misinformation and increase public trust in the veracity of online material.

Governments, in collaboration with industry, should:

- fund research into the further development and deployment of detection technologies, especially for use by government institutions, media organisations and fact checkers
- require digital platforms to deploy detection tools, especially to identify and label content generated through deep fake processes.

2. Behavioural change

Currently, high-quality audio-visual material is widely accepted at face value by the media and individuals as legitimate. In other words, seeing is still believing. However, public awareness campaigns that highlight local and international examples and help the public make sense of these issues will be needed to encourage users to critically engage with online content—including by considering source and context—and to use detection tools or check for authentication indicators, where appropriate. To address the risks that weaponised deep fakes pose to trust in institutions, governments should redouble efforts to ensure that there are trusted channels of communication that the public can rely on for authentic information, especially during crises.

Governments, in collaboration with industry, should:

- support trusted purveyors of information, such as local and national news media providers
- increase support for dedicated transparency bodies and initiatives
- encourage social media platforms to expand verified account programs, with stringent checks for achieving verification, to help users identify the source of information in order to better assess whether it's likely to be trustworthy and credible
- create established communications protocols for governments to provide public messages during crises (for example, via trusted messaging platforms, social media accounts or national radio channels)
- create legislative and policy 'firebreaks' for time-sensitive or politically sensitive situations in which detection or authentication related solutions are likely to be insufficient (for example, by implementing 'media blackouts' in the hours before an election).

3. Authentication standards

An alternative to detecting all false content is to signal the authenticity of all legitimate content. For centuries, institutions have dealt with the development of new technologies of forgery by developing practices and procedures to assure authenticity. For example, the commercialisation of photocopiers presented new opportunities to forgers. That challenge was met by technical responses (such as simulated watermarks and polymer banknotes) and new laws and policies (for example, processes by which a trusted third party, such as a justice of the peace, can 'certify' copies of original documents). Over time, it's likely that certification systems for digital content will become more sophisticated, in part mitigating the risk of weaponised deep fakes. In particular, encryption and open ledger 'blockchain' technologies may be used to authenticate digital content. Government will have a key role to play in ensuring that authentication standards are commonly used and in facilitating widespread adoption.

Governments, in collaboration with industry, should:

- support research into appropriate authentication technologies and standards
- introduce common standards relating to digital watermarks and stronger digital chain-of-custody requirements.

Notes

[1] The foreword was made by copying a primer sentence about deep fakes into a web-hosted text generator called ‘Talk to Transformer’. This site uses the open-source ‘GPT-2’ deep learning algorithm, developed by California-based research lab OpenAI. The headshot was created by a deep fake generator, online.

[2] Allan Xia, *Twitter,* 1 September 2019, online.

[3] BA Zagribeinyy, A Zhavoronkov, A Aliper, D Polykovskiy, VA Terentiev, V Aladinskiy, MS Veselov, A Aladinskaia, A Asadykaev, A Zhebrak, LH Lee, R Soll, D Madge, Li Xing, Tso Guo, A Aspuru-Guzik, YA Ivanenkov, R Shayakhmetov, ‘Deep learning enables rapid identification of potent DDR1 kinase inhibitors’, *Nature Biotechnology*, 2019, 37(9):1038–1040.

[4] ‘AI catching wildlife poachers’, *Silverpond*, 2018, online.

[5] Deep learning is a subfield of machine learning in which artificial neural networks—algorithms inspired by the human brain—learn from large amounts of data. Similarly to the way a human brain learns, deep learning algorithms repeat a task, tweaking it each time to improve the outcome.

[6] Samantha Cole, ‘AI-assisted fake porn is here and we’re all fucked’, *Vice*, 12 December 2017, online.

[7] ZAO app, online.

[8] Kelly M Sayler, Laurie A Harris, *Deep fakes and national security*, Congressional Research Service, Washington DC, 14 October 2019, online.

[9] James Vincent, ‘Deepfake detection algorithms will never be enough’, *The Verge*, 27 June 2019, online.

[10] ‘Pelosi videos manipulated to make her appear drunk are being shared on social media’, *Washington Post*, *YouTube*, 23 May 2019, online.

[11] Stephanie Grisham, *Twitter,* 8 November 2018, online.

[12] Jason Guerrasio, ‘64 years after James Dean’s death, the actor will star in a new movie. Some in Hollywood are horrified but the advances in visual effects could make it commonplace’, *Business Insider Australia*, 12 November 2019, online; Dani Di Placido, ‘James Dean and the rise of “deep fake” Hollywood’, *Forbes*, 8 November 2019, online.

[13] Zagribeinyy et al., ‘Deep learning enables rapid identification of potent DDR1 kinase inhibitors’.

[14] M Chui, M Harryson, J Manyika, R Roberts, R Chung, A van Heteren, P Nel, *Notes from the AI frontier: Applying AI for social good*, McKinsey Global Institute, 2018.

[15] Sarah Cameron, Ian McAllister, *The 2019 Australian Federal Election: Results from the Australian Election Study*, Australian National University, 15, online.

[16] Catherine Stupp, ‘Fraudsters used AI to mimic CEO’s voice in unusual cybercrime case’, *Wall Street Journal*, 30 August 2019, online.

[17] ‘Deepfake videos easily fool face systems, researchers warn’, *Biometric Technology Today*, 2019(10):3.

[18] Douglas Harris, ‘Deepfakes: False pornography is here and the law cannot protect you’, *Duke Law & Technology Review*, 2018, 17(1):99.

[19] Dave Lee, ‘Deepfakes porn has serious consequences’, *BBC News*, 3 February 2018, online.

[20] Heidi Moore, Dan Roberts, ‘AP Twitter hack causes panic on Wall Street and sends Dow plunging’, *The Guardian*, 24 April 2013, online.

[21] Fergus Hanson, Sarah O’Connor, Mali Walker, Luke Courtois, *Hacking democracies*, ASPI, Canberra,.15 May 2019, online.

[22] Ze Yang, Can Xu, Wei Wu, Zhoujun Li, ‘Read, attend and comment: a deep architecture for automatic news comment generation’, conference paper, 26 September 2019, online. In research supported by China’s National Natural Science Foundation and National Key R&D Program, researchers propose a method for using deep learning to distil key points in a news article and automatically generate attention-maximising comments.

[23] Christopher Paul, Miriam Matthews, *The Russian ‘firehose of falsehood’ propaganda model*, RAND Corporation, Santa Monica, 2016, online.

[24] Anne-Marie Slaughter, Asha Castleberry, ‘ISIS 2.0 and the information war’, *Project Syndicate*, 27 September 2019, online.

[25] Brennan Weiss, ‘A Russian troll factory had a $1.25 million monthly budget to interfere in the 2016 US election’, *Business Insider Australia*, 17 February 2018, online.

[26] Dmitry Volchek, Daisy Sindelar, ‘One professional Russian troll tells all’, *Radio Free Europe / Radio Liberty*, 25 March 2015.

[27] *Report of the Select Committee on Intelligence, United States Senate: Russian active measures campaigns and interference in the 2016 US election, volume 2: Russia’s use of social media with additional views*, 57, online.

[28] Raphael Satter, ‘Experts: Spy used AI-generated face to connect with targets’, *AP News*, 14 June 2019, online.

[29] ‘Deep fake videos could “spark” violence’, *BBC News*, 13 June 2019, online.

[30] Ian Brown, ‘Imagining a cyber surprise: How might China use stolen OPM records to target trust?’, *War on the Rocks*, 22 May 2018, online.

[31] ‘Teken de Klimaatpetitie’, *Facebook*, 19 May 2018, online.

[32] The disclaimer was, however, spoken only in English and not translated into Dutch.

[33] ‘Gabon 24’, *Facebook*, 31 December 2018, online.

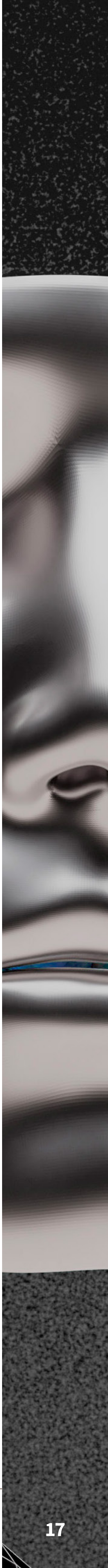
[34] Ali Breland, ‘The bizarre and terrifying case of the “deepfake” video that helped bring an African nation to the brink’, *Mother Jones*, 2019, online.

[35] For example, Google-supported non-profit Jigsaw is in the early stages of developing an ‘Assembler’ tool to help journalists detect fake images, video and audio. Jared Cohen, ‘Disinformation is more than fake news’, *Medium*, 2020, online.

[36] M Westerlund, ‘The emergence of deepfake technology: a review’, *Technology Innovation Management Review*, 2019, 9(11).

[37] Robert Chesney, Danielle Keats Citron, *Deep fakes: a looming challenge for privacy, democracy, and national security*, Social Science Research Network, New York, 2018, online.

[38] Kelly McBride, ‘The “liar’s dividend” is dangerous for journalists. Here’s how to fight it’, *Poynter*, 17 May 2019, online.



Acronyms and abbreviations

AI	artificial intelligence
GAN	generative adversarial network

ASPI
AUSTRALIAN
STRATEGIC
POLICY
INSTITUTE

INTERNATIONAL
CYBER POLICY
CENTRE