Facing Fakes

How politics and politicians can respond to the deepfake age

By Matthew Feeney

CENTRE FOR POLICY STUDIES
Est. 1974

About the Centre for Policy Studies

The Centre for Policy Studies is one of the oldest and most influential think tanks in Westminster. With a focus on taxation, business and economic growth, as well as housing, energy and innovation, its mission is to develop policies that widen enterprise, ownership and opportunity. Founded in 1974 by Sir Keith Joseph and Margaret Thatcher, the CPS has a proud record of turning ideas into practical policy. As well as developing much of the Thatcher reform agenda, its research has inspired many more recent policy innovations, such as raising the personal allowance and National Insurance threshold, reintroducing free ports and adopting 'full expensing' for capital investment.

About the Author

Matthew Feeney is Head of Tech & Innovation at the Centre for Policy Studies. Before joining the CPS, Matthew was the director of the Cato Institute's Project on Emerging Technologies. His writing has appeared in The New York Times, The Washington Post, City A.M., and others. He received both his BA and MA in philosophy from the University of Reading.

Acknowledgments

The author would like to thank Robert Colvile and Karl Williams for their feedback and review. As ever, any errors are the author's own.

Contents

Executive Summary

We are in the midst of the UK's first deepfake general election. Although the election campaigns are only a few weeks old, deepfake or other AI-generated content is already spreading rapidly.

Some of this content is harmless, such as a satirical video portraying Rishi Sunak outlining troop deployments in the computer game Fortnite, in order to lampoon his proposals for National Service.[1] Or a satirical TikTok video showing Sunak saying that he could not care less 'about energy bills being over £3,000'.[2] But some of it is much less so. There was the video purporting to show Wes Streeting, Labour's Shadow Health Secretary, calling Diane Abbott a 'silly woman' during a Politics Live appearance.[3] Or a video that appeared to show Labour North Durham candidate Luke Akehurst – a particular hate figure for the Corbynites – using crude language to mock constituents.[4] By the time this report is published, there will doubtless be many, many more.

> **'Although the election campaigns are only a few weeks old, deepfake or other AI-generated content is already spreading rapidly'**

This is not a uniquely British problem. While the UK's first political 'deepfake moment' came last year, featuring fake audio purporting to reveal Sir Keir Starmer swearing at staffers, such content is springing up in election campaigns around the world. And it's not just politics. Invariably, new advances in image, audio and video generation have been used for all manner of sinister purposes. The vast majority of deepfake content, indeed, is pornographic. But it has also become a core part of fraud and blackmail scams.

Given the current general election campaign, we should be particularly wary of deepfakes aimed at spreading election misinformation and disinformation. The relatively short history of deepfakes is already full of examples of deepfakes featuring politicians around the world.[5]

1 Smith Galer, S. [@Sophiasgaler]. (2024, June 3). news media: people are making AI deepfakes of Rishi Sunak! This is awful!!! teenagers on fortnite: [Tweet]. Twitter. Link

2 Angus Colwell, 'The TikTok stars taking on the Tories', *The Spectator, June 1*, 2024. Link

3 George Hancorn, 'Wes Streeting calls out fake video which claims he called Diane Abbott a 'silly woman'', ITV, June 4, 2024. Link

4 James Moules, 'Second deepfake Labour video in two days as Streeting and Akehurst targeted', LabourList, June 4, 2024. Link

5 Matthew Feeney, 'Deepfake Laws Risk Creating More Problems Than They Solve', Federalist Society's Regulatory Transparency Project, March 1, 2021. Link

There have, inevitably, been calls to regulate this dangerous technology – ranging from outright bans to targeted regulations and laws that prohibit the use of deepfakes in specific contexts (e.g. deepfakes portraying political candidates within the 60 days before an election, revenge pornography, etc).

Unfortunately for lawmakers, the nature of social media, the state of deepfake detection tools, the low cost of deepfake creation, and the limited reach of British law mean that we should expect harmful deepfake content to proliferate regardless of how Parliament acts. The current general election might be the UK's first deepfake general election, but it will not be the last.

While the private sector is at work developing methods to detect deepfakes, it is likely that solutions such as watermarking content are unlikely to be useful at scale – or to deter foreign adversaries from using such techniques, which is a particular concern given the history of foreign interference in Western elections.

Yet while deepfake harms may cause headlines during the general election as well as over the months and years to come, we must bear in mind that deepfake technology can also have valuable applications, for example in the arts, journalism, documentaries and education. We should also take comfort from history. New technologies emerged to tackle deceptive editing and forgeries.

> **'The relatively short history of deepfakes is already full of examples of fraudulent content featuring politicians around the world'**

This paper will therefore examine the growth of deepfake technology, the benefits and drawbacks it can bring (the former often being underplayed, given the frenzy of concern around the technology).

In particular, drawing on the author's previous paper, 'Regulating for Growth', we will argue that the Government should avoid technology-specific regulations and limit the regulatory state's reach to the mitigation of the likely and significant harms the technology may cause. There are already all manner of laws governing elections, harassment, blackmail and fraud. It will be simpler, more transparent and more sensible to update such laws rather than creating entirely new legislation – which will, if past attempts at tech regulation are anything to go by, rapidly be outpaced by technological developments, and indeed be likely to have all manner of definitional problems and unintended consequences.

There is, however, a cultural problem here. We are now in a low-trust society. As this paper argues, hopes for rigorous watermarking of deepfake content are not only technologically challenging, but are likely to be ignored not only by foreign adversaries, but by a public whose mistrust in existing institutions seems to grow by the year, and who are unlikely to trust the gatekeepers and verifiers whose content they already suspect of being biased against their political tribe. There is also the risk that, as with conventional fraud, deepfakes become predominantly targeted at the elderly, who may be less sceptical of such content than the tech-savvy young.

Hence our argument that there is no good solution to the rise of deepfakes, nor is there likely to be one. But bringing the legislative hammer down, in a knee-jerk response to electoral misinformation, is likely to be not just ineffective but actively counterproductive. We argue that the best solution to deepfake technology is to update existing laws, rather than creating new regulations – to follow the existing principle that it is the content itself that should be legal or illegal, not the form of its creation.

> **'We argue that the best solution to deepfake technology is to update existing laws, rather than creating new regulations'**

We also argue that whichever party is in power after the election should build on the creation of the AI Safety Institute in 2023, which established the UK as one of the world's hubs for AI safety public policy research, via the creation of a deepfake taskforce within that institution. This would mean that the Government is up to date on relevant threats and detection methods as well as the state of deepfake detection technology. It would also ensure the British state is as prepared as any to tackle deepfake harms – because as the election campaign shows, we will soon be in a situation where voters may struggle to have confidence in the veracity of anything they see or hear on screen.

Part One: What are Deepfakes?

Deepfakes are, to most people, fake images, videos or audio recordings that look and sound like the real thing – which advances in technology have made it astonishingly easy to create, at next to no cost.

There has obviously been fake content in the past. But deepfakes are qualitatively and technologically distinct.

Technically speaking, deepfakes consist of audio, visual or photographic content created by AI technologies such as Generative Adversarial Networks (GANs) and autoencoders.[6] GANs are made up of generators that create data and discriminators that evaluate that data.[7] These two systems train themselves in a feedback loop of offence and defence. Autoencoders are neural networks (computer systems that mimic how human brains work) tasked with compressing and reconstructing data.[8]

> **'Technically speaking, deepfakes consist of audio, visual or photographic content created by AI technologies such as Generative Adversarial Networks (GANs) and autoencoders'**

The results of these techniques, as mentioned above, are realistic fake pieces of content that can mimic people's voices and facial movements.

Deepfakes first emerged in 2017, when the Reddit user 'deepfakes' (whom they are named after) posted fake pornographic content featuring the faces of celebrities.[9] Since then, deepfakes have been used for a range of purposes.

In one sense, there is nothing new about this. Deepfakes belong in the family of content editing and manipulation technologies that include Photoshop and visual effects technologies. As the history of photography, journalism, film and television has revealed, these technologies can be used for a range of purposes. Deepfakes are no different. What is different is the ease and low cost with which they can be created. Deepfakes are low-cost and high value. Deepfake creation technology allows someone with relatively few computer skills or knowledge to generate very realistic-looking content that could be mistake for authentic material.

It is also important to say – because it is often missed – that deepfakes are a technology that can be used for both good and ill. They have been used by criminals, yes, but also by artists, documentarians and filmmakers. For example, South Park creators Trey Parker and Matt Stone used deepfake technology to produce an online comedy series called 'Sassy Justice', which follows a reporter of that name

6 U.S. Government Accountability Office. (2020). Science & Tech Spotlight: Deepfakes (GAO-20-379SP). Link

7 CVisionLab. (n.d.). Deepfake (Generative adversarial network). Link

8 Zucconi, A. (2018, March 14). Understanding the technology behind deepfakes. Retrieved June 5, 2024. Link

9 Payne, L. (2024, June 4). Deepfake. Encyclopaedia Britannica. Retrieved June 5, 2024. Link

with Donald Trump's face, based in Wyoming.[10] Bruce Willis gave permission for a deepfake firm to use his likeness in an advert for a Russian mobile phone network.[11] The upcoming film Here, directed by Robert Zemeckis and starring Tom Hanks, will reportedly feature extensive use of deepfake technology – just as the same duo previously used cutting-edge technology to insert Hanks' likeness into historical footage in Forrest Gump, or captured his image on computer for The Polar Express.[12]

In many of these instances, the use of deepfakes will be obvious given the context (i.e. comedy and creative filmmaking). Indeed, in the filmmaking context in particular, it is worth thinking of deepfakes as an improvement on CGI technology.

'The creative industries account for 5.7% of UK GVA (Gross Value Added), the equivalent of approximately £124.6 billion, and are one of the British industries that has been growing most strongly in recent years'

But of course, deepfakes also have applications beyond the creative arts. Documentarians recording the persecution affecting the gay community in Chechnya used deepfake techniques to hide the identities of gay Chechens.[13] Deepfakes could be used in the near future to revolutionise filmmaking, bring extinct music back to life, and help students learn languages. As should come as no surprise, the ability to replace someone's face at low cost has a variety of applications.

Unfortunately, the educational, artistic and documentary use of deepfakes is often overshadowed by more nefarious uses. This is especially concerning for the UK, where the creative industries account for 5.7% of UK GVA (Gross Value Added), the equivalent of approximately £124.6 billion, and are one of the British industries that has been growing most strongly in recent years.[14]

This is because the sad truth is that the vast majority of available deepfake content is non-consensual pornography, politically motivated disinformation, and scamming material – to which as the technology advances could easily be added knock-off versions of familiar films, cartoons or TV shows.[15] It is these kinds of deepfake content that have prompted calls for regulation and legislation.

This is not surprising given the threat such content poses to democracy, financial institutions, as well as personal dignity and safety. Many victims of deepfake pornography have suffered all kinds of horror and trauma due to the spread of content featuring what appears to be their faces and/or bodies, including the onset of severe depression, anxiety, and suicidal thoughts.[16]

10 Sassy Justice. (2020, October 26). Sassy Justice with Fred Sassy (Full Episode) | Deep Fake and Deep Fake: The Movie [Video]. YouTube. Link

11 Lees, D. (2023). Deepfakes in documentary film production: Images of deception in the representation of the real. Studies in Documentary Film. Link

12 Joseph, A. (2023, February 1). *Tom Hanks, Robert Zemeckis' new film will utilize deepfake AI technology.* CBR. Retrieved June 5, 2024. Link

13 Scott, A. O. (2020, July 1). *Deepfake technology adds realism to documentary 'Welcome to Chechnya'.* The New York Times. Link

14 Rachel Moyce, 'DCMS Sectors Economic Estimates Gross Value Added 2022 (provisional) ', Department for Digital, Culture, Media & Sport and Department for Science, Innovation and Technology, 15 February 2024. Link

15 Levy, S. (2023, April 24). *The internet is full of deepfakes, and most of them are porn.* PCMag UK. Retrieved June 5, 2024. Link

16 Iynn, A., & Henry, N. (2020). 'It's torture for the soul': The consequences of image-based sexual abuse. Social & Legal Studies, 29(5), 1-20. Link

In terms of politics, the American presidential election has already served as a venue for deepfake disinformation, with voters in New Hampshire receiving fake calls supposedly from President Joe Biden urging them not to vote in the Democratic primary and to 'save' their vote for the presidential election.[17] In October 2023 the UK experienced its 'first political deepfake moment' when an unidentified person released deepfake audio purporting to reveal Labour leader Sir Keir Starmer swearing at staffers.[18]

At the time of writing, the UK general election is only a fortnight old. Yet already, deepfake or less realistic 'cheapfake' content has already emerged. Satirical deepfake content has spread poking fun at the Prime Minister's proposal for National Service, showing Rishi Sunak supervising the deployment of teenage soldiers in the computer game Fortnite.[19] Another AI-generated video, this one spreading on TikTok, shows Sunak saying that he could not care less 'about energy bills being over £3,000'.[20]

Other deepfake content is less obviously satirical. In early June 2024 a video spread on social media purporting to show Wes Streeting, Labour's Shadow Health Secretary, calling fellow Labour politician Diane Abbott a 'silly woman' during a BBC Politics Live appearance.[21] In the early days of the election campaign, deepfake video appeared on X (formerly Twitter) showing Labour's North Durham candidate Luke Akehurst using crude language to mock constituents.[22]

'In early 2024, a financial professional wired $25 million to fraudsters who had used deepfake technology to make it appear as if he was on a video call with colleagues'

That said, not all political deepfakes are used to harm opponents. In India, one political candidate used deepfake technology to make it appear that he spoke a Hindi dialect, when in the original version he spoke English.[23] However, those deepfakes that have emerged in this year's British general election so far seem to have been overwhelmingly designed to ridicule candidates or portray them in a negative light. And while we should defend the right of citizens to ridicule and make fun of politicians, it is not hard to imagine how deepfake technology could be used to spread not just false claims about what politicians have said, but harmful election dis/misinformation, for example about the date of the election, or voting eligibility requirements. Just look at the concerted effort made by Russia and other malign actors to influence the results of recent American elections, at a time when the relevant technologies were much less advanced.

Deepfakes also, of course, make it much easier for criminals to fool the public. In early 2024, a financial professional wired $25 million to fraudsters who had used deepfake technology to make it appear as if he was on a video call with colleagues.[24] In 2023, the Federal Bureau of Investigation issued an advisory notice

17 Bohannon, M. (2024, February 6). *Biden deepfake robocall urging voters to skip New Hampshire primary traced to Texas company.* Forbes. Link

18 Bristow, T. (2023, October 9). *Keir Starmer suffers UK politics' first deepfake moment. It won't be the last.* Politico. Link

19 Smith Galer

20 Angus Colwell

21 George Hancorn

22 James Moules

23 Vincent, J. (2020, February 18). An Indian politician is using deepfakes to win votes. The Verge. Link

24 Westcott, B. (2024, February 4). Hong Kong CFO scammed out of millions in deepfake video call. CNN. Link

warning the public that 'malicious actors have used manipulated photos or videos with the purpose of extorting victims for ransom or to gain compliance for other demands (e.g. sending nude photos)'.[25] This brand of extortion is colloquially known as 'sextortion' and has had a particularly devastating effect on teenagers, some of whom have committed suicide after being targeted.[26]

No doubt artists in the creative industries will find ways to use deepfakes to creative effect that have not been considered yet, and there are educational applications of deepfakes that could make teaching more immersive and engaging. However, while we should be excited about the beneficial uses of deepfakes, the harmful effects of deepfakes are clear and obvious. Inevitably, this has resulted in calls for legislative and regulatory responses. So, what should politicians do? What, indeed, can they do?

> **'Inevitably, the rise of deepfakes has resulted in calls for legislative and regulatory responses'**

Attempts at Regulation and Legislation

So far, the regulations and laws on deepfakes generally fall into two categories: 1) disclosure and/or transparency requirements, 2) prohibitions of specific uses. The specific applications that have received the most attention include political campaign speech, content designed to facilitate fraud and sexually explicit material.

The EU's AI Act, for example, includes a deepfakes transparency requirement. It requires that 'deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated.'[27] This is not dissimilar from Chinese law, which requires individuals and groups (such as organisations and companies) to disclose if they have used deepfake technology and requires deepfake content to be accompanied by a disclaimer available to the viewer.[28]

The US state of Washington requires disclosure of which election-related content has been altered via deepfakes.[29] A similar bill is making its way through the New Mexico legislature, requiring that campaigns disclose if they have used deepfakes in their campaign adverts while also making it a crime to use deepfakes as a means to deliberately deceive voters.[30] Washington and New Mexico are hardly outliers. The majority of American state legislatures have either already passed deepfake election legislation or are considering such legislation.[31]

Some deepfake election legislation is more strict. California's Assembly Bill 730 bans the spread of deceptive deepfake content related to a political candidate in the 60 days before an election.[32] Minnesota's deepfake legislation is similar, prohibiting the

25 Federal Bureau of Investigation. (2023, June 5). *Malicious actors manipulating photos and videos to create explicit content and sextortion schemes.* FBI. Retrieved June 5, 2024. Link

26 Hendery, S. (2023, June 7). *Deepfakes of victims used in sextortion attacks spike, FBI warns. SC Media.* Link

27 Whittaker, T., & Smith, L. (2024, February 16). *EU AI Act: What you need to know. Burges Salmon.* Link

28 Lawson, A. (2023, April 24). A look at global deepfake regulation approaches. Responsible AI. Link

29 Washington State Legislature. (2024). Senate Bill 5152-S: Relating to prohibiting the use of deepfake technology for political purposes. Link

30 New Mexico Legislature. (2024). House Bill 182. Link

31 Public Citizen. (2023, June 3). *Tracker: State legislation on deepfakes in elections.* Link

32 California Legislature. (2019). *Assembly Bill 730: Elections: Deceptive audio or visual media.* Link

spread of such content 90 days before an election.[33] At the federal level, Senators Amy Klobuchar (D-MN), Joshua Hawley (R-MO), Christopher Coons (D-CT) and Susan Collins (R-ME) have introduced the Protect Elections from Deceptive AI Act, which would limit the use of AI to create images of federal election candidates with the intent of influencing an election.[34]

However, while many American lawmakers may be keen to pass deepfake laws it is notable that the US Supreme Court has yet to rule on their constitutionality and it is possible that such laws run afoul of the First Amendment.[35]

Still, in the wake of the current general election, it is possible – perhaps even likely – that the next parliament will introduce a bill modelled on the American legislation discussed above. As we have seen with the Online Safety Act, politicians' willingness to regulate the tech sector often correlates strongly to the extent to which that technology makes their own lives difficult. So beyond specific proposals around elections, what might such a legislative response look like?

> **'The best-known content-specific bans are those taking aim at deepfake pornography'**

Well, there are already some measures in place – indeed, perhaps the best-known content-specific bans on deepfakes are those taking aim at deepfake pornography. The Online Safety Act bans the spread of such material, as do some American states. For example, the US state of Illinois' ban on deepfake pornography prohibits the sending of an 'intentionally digitally altered sexual image'.[36] New York's deepfake revenge pornography law prohibits the sending or publication of sexually explicit content of someone without their consent, including 'an image created or altered by digitization, where such person may reasonably be identified from the still or video image itself or from information displayed in connection with the still or video image'.[37]

Indeed, although lawmakers across the world have rushed to address deepfake harms, some believe that current law does not go far enough. At least two campaigns, ControlAI and Ban Deepfakes, have called for an outright ban on deepfakes.[38] Control AI's policy proposal wants lawmakers to 'make the creation and dissemination of deepfakes a crime', while the Ban Deepfakes campaign claims that 'the only effective way to stop deepfakes is for governments to ban them at every stage of production and distribution'.[39]

Yet aside from the sheer impracticality of such demands, these campaigns reveal how difficult it can be to make coherent policy on deepfakes. For example, there are many valuable uses of deepfake technology, all of which would be illegal if 'the creation and dissemination of deepfakes' were banned 'at every stage of production and distribution'.[40]

---

33 Minnesota Legislature. (2023). *House File 1370: Relating to public safety; establishing a cause of action for nonconsensual dissemination of deep fake sexual images; establishing the crime of using deep fake technology to influence an election.* Link

34 U.S. Congress. (2023). *S.2770 - Protect Elections from Deceptive AI Act.* Link

35 Baiocco, A. (2022, January 5). Political 'deepfake' laws threaten freedom of expression. Institute for Free Speech. Link

36 Illinois General Assembly. (2023). *House Bill 2123: Digital Forgeries Act.* Retrieved June 5, 2024. Link

37 New York State Senate. (2023). Senate Bill S1042A: Relating to unlawful dissemination or publication of intimate images created by digitization. Link

38 Control AI. (2024). Deepfakes. Link, Ban Deepfakes. (2024). Contact us. Link

39 Control AI. (2024). *Deepfakes policy.* Link, Ban Deepfakes. (2024). The solution. Link

40 Ibid.

Moreover, a closer look at these campaigns reveals that they exempt large swathes of deepfake content from their definition of 'deepfakes'. Control AI's proposal defines deepfakes in such a way so as to exclude satire and actors licensing their images, stating that 'these legitimate uses would remain legal, as they fall outside the deepfake definition'. Ban Deepfakes offers a similarly narrow definition of deepfakes: 'Deepfakes are non-consensually AI-generated voices, images or videos that are created to produce sexual imagery, commit fraud, or spread misinformation'.[41]

> **'At least two campaigns, ControlAI and Ban Deepfakes, have called for an outright ban on deepfakes'**

While it is reassuring that even campaigns calling for 'bans' are not willing to sacrifice the valuable applications of deepfakes in the pursuit of preventing harm, the campaigns should serve as a reminder to policymakers that there is a risk of the word 'deepfake' coming to be associated exclusively with harmful content rather than a broader category of content.

Instead, it would be best for policymakers to think of deepfake technology like other content-creation technologies such as the printing press, the Internet, the television, the camera, and the radio. All can be used for evil as well as valuable goals.

This is all the more important because there are a series of problems with the most common approaches to regulating deepfakes, which we shall explore in the next section.

Free Speech

First and most obviously, proposed remedies to the abuse of deepfakes raise concerns about free speech.

Many people are persuaded that restrictions on deepfakes are justified on the grounds that harmful deepfake content contributes little to important debate, humiliates innocent people, harms reputations, undermines liberal institutions, and facilitates blackmail and fraud.

Nonetheless, lawmakers should be hesitant to embrace a broad approach to deepfakes that tackles specific technologies rather than their use.

Targeting the use of technology rather than the technology itself is the approach taken by governments all over the world when it comes to speech regulation. That criminals can use books, television, phones, radio and online platforms to publish, broadcast or stream illegal speech does not warrant a ban on books, televisions, radios or social media platforms. Rather, governments tend to define categories of speech that are prohibited in each of these mediums.

Furthermore, ridiculing politicians and other powerful figures is one of the most protected categories of speech. Even in the UK, hardly a bastion of free speech these days, those seeking to criticise politicians in print, song or television enjoy broad protections.

Many people justify policies affecting speech on the basis of the content in question, not the technology that sends, creates, preserves, or edits it. The Online Safety Act is only one of the pieces of legislation around the world that includes examples of this,

41 Ban Deepfakes. (2024). *Frequently asked questions.* Link

with the law making special provisions for 'journalism' content. The Online Safety Act takes an approach to online content that seeks to categorise content as harmful by definition. Such an approach is similar to that outlined in the table below.

	A Authentic Footage	B Inauthentic Footage
1 Valuable	1) Documentaries 2) Journalism 3) Stand-up comedy	1) Fictional film and television programmes
2 Harmful	1) Terrorist propaganda 2) Revenge pornography 3) Self-harm/suicide instruction 4) Racist content	1) Deepfake revenge pornography 2) Election interference content 3) Scams 4) Content portraying fictional harm to humans and non-human animals

This kind of scheme is not watertight, however. For instance, some valuable documentaries, journalism programmes and films may show terrorist propaganda, racist content, deepfakes or animals being harmed. In addition, journalism outlets regularly have to edit footage, which can open them to allegations of deceptive or ideological bias. And many people would dispute the valuable contribution some films have made to society, or that every stand-up comedian is making a worthwhile contribution.

Nonetheless, the fact remains that when it comes to online speech, many lawmakers around the world are seeking to protect content considered valuable, whether it is authentic and inauthentic, while also trying to minimise the spread of harmful content. Given that creators can use deepfakes to produce valuable content as well as harmful content, lawmakers should be aware of how regulations on deepfakes may stifle valuable content. This should be of particular concern to British lawmakers given the size of the UK's creative industry sector.

Another issue that lawmakers must tackle, as alluded to above, is the definition of 'deepfake'. This is because most legislation tackling deepfakes does not actually mention 'deepfakes'. Rather, bills and laws targeting deepfakes define the media at issue in a variety of ways including 'advanced technological false personation record' or as an 'image, whether made or altered by computer graphics or in any other way, which appears to be a photograph or film'.[42]

The Protect Elections from Deceptive AI Act in the US includes an especially long and convoluted definition of 'deceptive AI-generated audio or visual media', defining it as 'an image, audio, or video that— [...] is the product of artificial intelligence or machine learning, including deep learning techniques, that [...] merges, combines, replaces, or superimposes content onto an image, audio, or video, creating an image, audio, or video that appears authentic; or [...] generates an inauthentic image, audio, or video that appears authentic; and [...] a reasonable person, having considered the qualities of the image, audio, or video and the nature of the distribution channel in which the image, audio, or video appears [...] would have a fundamentally different understanding or impression of the appearance, speech, or expressive conduct exhibited in the image, audio, or video than that person would have if that person were hearing or seeing the unaltered, original version of the image, audio, or video; [...] would believe that the image, audio, or video accurately exhibits any appearance,

42 UK Government. (2023). *Online Safety Act 2023*. Retrieved June 5, 2024. Link
U.S. Congress. (2019). *H.R.3230 - DEEP FAKES Accountability Act*. Link

speech, or expressive conduct of a person who did not actually exhibit such appearance, speech, or expressive conduct.'[43]

It is not a surprise that defining deepfake content is difficult. Too broad a definition will include a wide range of content, such as video or images that have undergone traditional digital editing. Too narrow a definition may exclude some content such as 'cheapfakes' or 'shallowfakes', which are created using AI tools but do not necessarily use GANs and autoencoders.

'It is not a surprise that defining deepfake content is difficult. Too broad a definition will include a wide range of content, such as video or images that have undergone traditional digital editing'

Problems with the definitions of deepfake content could have downstream effects on valuable content such as political commentary and satire. For example, California's deepfake election law, the first such law to pass in the US, prohibits the sending of 'materially deceptive audio or visual media' of a candidate within 60 days of an election 'with the intent to injure the candidate's reputation or to deceive a voter into voting for or against the candidate' if the content is not accompanied by a disclaimer.[44] The law makes an exemption for 'satire or parody', but these terms are not defined in the law. Some of the deepfake content that has already emerged in the UK election fits comfortably into the satire category. No serious person thinks that the video game Fortnite has anything to do with Sunak's national service proposal. But it is not hard to imagine difficult edge cases where one person's obvious satire looks to someone else like nefarious election interference. It is clear that the people making anti-Sunak videos, for example, want to injure his reputation – they just think people are smart enough to realise their spoofs aren't real.

There is another problem. As we pointed out repeatedly when campaigning against the Online Safety Act, such legislation invariably has a chilling effect. If you threaten technology companies with fines of up to 10% of their turnover if they fail to have a sufficiently robust content moderation system, then they will invariably err on the side of caution when deciding what material to ban – especially smaller firms which lack the big tech firms' resources in erecting systems of content monitoring and moderation.[45]

Whatever the line drawn by any new deepfake rules, creators, tech firms and technology providers will inevitably take a safety-first approach when it comes to creating or allowing users to create content that could be construed as violating the plethora of deepfake regulations and restrictions that are emerging. The result will be that some socially valuable content never makes it to audiences.

So given these qualifications and codicils, what should our approach to deepfakes actually look like? And are they even possible to regulate at all?

The history of communications technology is one of advances in editing and alteration being met with advances in detection techniques. Journalists, intelligence agencies, police forces and researchers all use technology to spot altered videos,

43 U.S. Congress. (2023). *S.2770 - Protect Elections from Deceptive AI Act.* Link

44 California Legislature. (2019). *Assembly Bill 730.*

45 Feeney, M. (2023). *A censor's charter? The case against the Online Safety Bill. Centre for Policy Studies.* Link

photos, and audio files. These same professionals are increasingly using technology designed to spot deepfakes.

Some journalism outlets have used a deepfake detection tool built by the Google incubator Jigsaw.[46] The Wall Street Journal, Reuters and The Washington Post are all directly addressing the rise of deepfakes.[47] In 2023, the BBC tested FakeCatcher, Intel's deepfake detection tool.[48]

> **'Even as deepfake quality improves, so will deepfake detection tools. Google DeepMind is working on SynthID, which focuses on watermarking individual pixels rather than whole images'**

The encouraging news on this front is that watermarking techniques are improving regularly. Members of the industry group the Coalition for Content Provenance and Authenticity (C2PA) – which includes Microsoft, Adobe, Sony, the BBC, AWS, Canon, Google and Intel - have committed to adopting a standard that embeds metadata into content, thereby allowing viewers to verify the provenance of content. OpenAI's image generation tool DALL·E 3 uses C2PA's credential standard.[49] This means that anyone who would like to verify if an image was created with DALL·E 3 can do so.

The bad news is that there are limitations to this kind of standard. It is easy to circumvent via cropping and screenshotting, so is hardly going to pose a significant deterrence to motivated bad actors, including hostile nation states.[50] Still, even as deepfake quality improves, so will deepfake detection tools. Google DeepMind is working on SynthID, which focuses on watermarking individual pixels rather than whole images.[51] Such novel techniques are necessary given that it is trivial to remove many traditional watermarks from images.

46 Jigsaw, 'Disinformation is More than Fake News,' February 2020. Link

47 Lucinda Southern, "A perfect storm': *The Wall Street Journal has 21 people detecting 'deepfakes'*," Dig Day, July 1, 2019. Link
'Seeing Isn't Believing: The Fact Checker's guide to manipulated video' The Washington Post. Link

48 Kelion, L. (2023, July 21). *Deepfake scams: How AI is being used to impersonate CEOs*. BBC News. Link

49 OpenAI. (2024). C2PA in DALL·E 3. Link

50 Ibid.

51 Gowal, S., & Kohli, P. (2023, August 29). Identifying AI-generated images with SynthID. Google DeepMind. Link

Part Two: How to Regulate Deepfakes

The history of content editing techniques is one full of such techniques being met with technologies designed to detect alterations in content. We should expect a similar arms race to continue with deepfakes.

However, we cannot rely on technology to do the job for us. In particular, a November 2023 paper from an international coalition of academics argued that, under a series of reasonable assumptions, 'strong and robust watermarking is impossible to achieve', and that as AI systems improve their capabilities, it will become harder rather than easier to create watermarking systems that are immune to attack.[52]

> **'The low cost of deepfake creation and the global nature of the Internet make creating effective deepfake legislation difficult'**

In other words, having proper watermarking and chain-of-custody will be of value to journalists, researchers and other professionals, who are creating deepfakes for legitimate purposes and have a clear interest in their content being labelled as such. But most people worried about deepfakes, who are seeking a reliable way to determine whether what they are seeing is authentic footage or not, may not be able to rely on such tools.

So what can be done?

The low cost of deepfake creation and the global nature of the Internet make creating effective deepfake legislation difficult. Yet as the general election campaign continues, we will almost certainly see more deepfake content, which will no doubt be accompanied by increased and urgent calls for the government to act. This will be true despite, as noted above, flaws in deepfake detection tools and the fact that restricting the use of a content-creation technology would be a novel but ultimately unhelpful policy.

Nonetheless, there are steps that the next Government can take to ensure that Parliament, intelligence agencies, police, industry and consumers are better prepared for the spread of malicious and harmful deepfake content such as election disinformation and misinformation, as well as deepfake content used to facilitate serious crimes.

52 Zhang, H., Edelman, B., & Barak, B. (2023). *Watermarking in the sand: Impossibility of strong watermarking for generative models.* Kempner Institute. Link

Transparent and Cooperative Defence

Many of the legislations and regulations aimed at tackling deepfakes are associated with a host of problems lawmakers should seek to avoid. In particular, while there are widespread calls for technology companies to track and watermark content, the research cited above should motivate us to dampen our expectations.

Not only is watermarking unlikely to result in robust deepfake detection, but watermarking requirements may well only be embedded within content created in friendly countries and by companies within our regulatory ambit. There appears to be little realistic prospect of deterring foreign adversaries from creating deepfakes designed to destabilise elections or sow distrust of important institutions.

But lawmakers should not lose hope. There are ways for Parliament, intelligence agencies, the Home Office and industry to work together in identifying the most relevant deepfake trends, detection methods, and developing risks.

‘In the UK, government departments such as the Department for Science, Innovation and Technology have organised collaborations with social media companies in order to tackle dis/misinformation’

One solution would be the creation of a designated deepfake taskforce, bringing together relevant officials and industry representatives. This taskforce could not only become a world-leading venue for deepfake detection research but also serve as a hub for legislative proposals aimed at ensuring that modern law adequately accounts for deepfakes. Indeed, as explained below, while deepfakes might be new, they pose familiar problems (e.g. election interference, blackmail, fraud) – almost all of which are already prohibited under current law. One of the taskforce’s priorities should therefore be to examine relevant current law to see if it is in need of updating thanks to the emergence of deepfakes.

Where should this taskforce sit? Well, in 2023, the Government established the AI Safety Institute – a world-leading organisation. This would be a natural home for a coordinated effort on deepfakes. Indeed, similar organisations at home and abroad could serve as templates or inspiration for the taskforce. In the US, Stanford University’s Internet Observatory and Program on Democracy and the Internet, the social media analytics firm Graphika, the Atlantic Council’s Digital Forensic Research Lab and the University of Washington’s Center for an Informed Public created the Election Integrity Partnership (EIP), which partnered with government officials in order to identify election dis/misinformation.[53] Other disinformation collaborations between government and civil society include the Global Disinformation Index.[54]

In the UK, government departments such as the Department for Science, Innovation and Technology have organised collaborations with social media companies in order to tackle dis/misinformation. In addition, the government has established the Defending Democracy taskforce, which reports to the National Security Council.[55]

53 Center for an Informed Public, Digital Forensic Research Lab, Graphika, & Stanford Internet Observatory (2021). The Long Fuse: Misinformation and the 2020 Election. Stanford Digital Repository: Election Integrity Partnership. v1.3.0 Available at: Link

54 Global Disinformation Index. (2024). About us. Link

55 UK Government. (2022, November 28). Ministerial Taskforce meets to tackle state threats to UK democracy. Link

Such efforts could also help guide lawmakers interested in tackling deepfakes. A deepfake taskforce could, like the Defending Democracy taskforce, report to the National Security Council and inform decision-making related to foreign adversary deepfake threats. It could also establish a formal relationship with Ofcom, which since the passage of the Online Safety Act is the lead regulator of social media.

It would also be critical that such a taskforce included representatives from those social media platforms and messaging services that are most often the venues for foreign adversary attacks and the spread of revenge pornography, as well as content used to facilitate scams and blackmail.

Another idea would be for the Government to build incentives for deepfake detection research by hosting regular competitions aimed at rewarding labs, researchers, universities, and companies working on deepfake detection methods and tools. For example, the Government could host regular public deepfake detection challenges with cash prizes as rewards.

'In early 2024 the Home Office, DSIT, the Alan Turing Institute, and The Accelerated Capability Environment (ACE) launched a Deepfakes Detection Challenge'

These challenges would allow researchers building deepfake detection tools to test and show off their products while allowing observers to see which deepfake detection tools and methods are the most effective.

Such challenges are not unprecedented. In early 2024 the Home Office, DSIT, the Alan Turing Institute, and The Accelerated Capability Environment (ACE) launched a Deepfakes Detection Challenge.[56] The government should pursue more of these kinds of challenges in order to help establish the UK as a home for AI safety, building off the successful launch of the AI Safety Institute in 2023.

Update Existing Law

While deepfakes might be new, they do not pose a distinct and original threat. As mentioned above, similar attempts at fraud and manipulation are rife throughout the history of technology. Deepfakes make deception and misinformation easier and cheaper to carry out, but they do not change its fundamental nature.

We have also seen in recent attempts to introduce sweeping new regulation of the technology sector – such as the Online Harms Act or Digital Markets, Competition and Consumers Act (which just squeaked through the last parliament in the final days) – that creating new tech regulation is rife with challenges, complication and complexity, and can indeed have many adverse effects that were not anticipated by those drafting the proposals. Such legislation also tends to be enshrined in law well after the problem at hand has become significant.

In the case of deepfakes, then, their rise ought to motivate lawmakers to amend existing law, rather than crafting new regulations and legislation aimed at deepfakes – building on the strong existing principle that what is legal or illegal is the content or speech itself, rather than the means of creating or distributing it.

56 Shanks, K. (2024, April 16). *Unmasking deception: Join the Deepfake Detection Challenge!.* Accelerated Capability Environment. Retrieved June 5, 2024. Link

In this instance, the UK already has comprehensive and robust laws governing election interference, harassment, revenge pornography, blackmail and fraud. Those lawmakers worried about criminals using deepfakes to facilitate crimes should therefore focus on amending these laws in light of deepfake threats rather than draft deepfake-specific laws. The deepfake taskforce proposed above would serve as an ideal venue for lawmakers to examine the state of current law.

Amid the ongoing election we should expect for policy experts and lawmakers to cite election law in particular. But here again, there are robust theoretical protections in place. Elections in the UK are governed by a range of legislation, including laws that prohibit making false claims about political candidates.[57] In addition, the Representation of the People Act 1983 prohibits 'placing undue spiritual pressure on a person' and 'doing any act designed to deceive a person in relation to the administration of an election'.[58]

If the next Government believes that deepfakes pose a significant risk to the integrity of elections, it should take steps to update existing election law to make it clear that deepfake content designed to deceive voters about the date of an election, the requirements for voting, or other important details about voting is prohibited. Although of course in doing so, the government must avoid drafting legislation that includes traditional image editing software in its definition of 'deepfake'.

> **'Elections in the UK are governed by a range of legislation, including laws that prohibit making false claims about political candidates'**

Likewise, lawmakers can update laws governing harassment, blackmail, and fraud to explicitly prohibit the use of deepfakes to facilitate those crimes. The sending of deepfake revenge pornography is already illegal under the Online Safety Act and the previous Government considered additional legislation which would have made the creation of sexually explicit deepfake content illegal, even if such content was never sent.[59]

Changes to existing laws are preferable to an overarching deepfake law or regulation. As with past speech technologies, deepfakes are best-regulated on a use-by-use basis rather than as a whole.

A Cultural Solution to a Cultural Problem

The history of speech technologies is full of institutions and norms catching up to new inventions. First came the printing press, then came professional journalism, newspapers, academic journals, censorship boards and publishers that allowed citizens and lawmakers to gauge the legitimacy of printed speech. Similarly, the British Board of Film Classification, the BBC charter, and Ofcom developed after the invention of television and radio.

Not each of these institutions and norms necessarily promoted a libertarian ideal of free speech, but they allowed citizens and lawmakers to navigate new technologies and provided a way for consumers to gauge the legitimacy of speech.

57 Representation of the People Act 1983 Section 106
See also: Electoral Commission. (2024). *Claims made in online political ads.* Link

58 Goodman, A., & Harrod, J. (2023, November 27). *Tackling deepfakes and disinformation in elections.* Local Government Lawyer. Link

59 UK Government. (2024, April 16). Government cracks down on deepfakes creation. Link

Such institutions have yet to emerge for much of the content created in the social media age. Some private companies have led efforts to create legitimising institutions for social media (e.g. the Meta Oversight Board), but none have succeeded in appeasing lawmakers intent on regulation or citizens who remain concerned about the power of 'Big Tech'. Indeed, although the Online Safety Act became law after years of heated debate, it remains to be seen whether it is effective at minimising child access to harmful content and indeed whether its structure lends more legitimacy to online speech platforms in the eyes of the public.

Deepfakes are not a new vehicle for content like the printing press, the radio or the television, but they are nonetheless examples of a new form of content that risks undermining the legitimacy of online platforms.

Without robust and trusted means to tackle deepfake content, the public will likely become increasingly sceptical of online platforms as well as institutions such as political parties and journalism outlets that rely on such platforms to reach their audiences.

Fortunately, the history of technology is full of examples of societies adapting to dramatic changes. The rise of Photoshop and modern visual effects technology did not destroy confidence in journalism. But the transition from no use of these technologies to widespread use was not without some incidents of criminals and foreign adversaries using them to their advantage.

'There is a risk that the spread of deepfakes prompts an unhelpful and widespread distrust of authentic content and a belief in fake content'

The worry with deepfakes is that the cost of production is so low and their quality so high that our online platforms will be saturated with deepfakes before we have developed robust technological and societal defences.

The cultural defences are harder to predict, especially when some of the most motivated foreign adversaries will deploy deepfake content regardless of criminal penalties.

One possible outcome of the emergence of deepfakes is that the generations born into the world of mobile phones and social media will quickly adapt to deepfakes and have their digital guard up when they consume online content, while the older generations take longer to adapt.

Such a situation would not be dissimilar to the way that fraudsters disproportionately target the elderly, who are less likely than the young to be conditioned against phone-based and online scams.[60]

More scepticism of online content in the right dose would be welcome, but there is a risk that the spread of deepfakes prompts an unhelpful and widespread distrust of authentic content and a belief in fake content.

60 Metropolitan Police. (2024, March 25). *Met officers target phone scammers who prey on elderly.* Link

Indeed, it is not hard to imagine politicians taking advantage of the 'Liar's Dividend', whereby they claim that authentic compromising or embarrassing content is a deepfake.[61] Think of a controversy similar to something like the following happening in the near future, perhaps even during the next few weeks:

Candidate A (Alex Adams) and Candidate B (Bryan Biggs) are competing in an election. A pseudonymous account posts authentic 20-year-old footage of Adams using a racial slur while at a university party. Within 24 hours of the post appearing Biggs and his party colleagues rush to condemn Adams, insisting that he apologise and drop out of the race. Adams refuses and his team launches a media campaign, accusing Biggs supporters of spreading a deepfake designed to destabilise the election. Media organisations and intelligence agencies rush to examine the video. They conclude that the footage is authentic and publish their findings within a week. But their findings are of little political consequence. Adams supporters believe Adams, and Biggs supporters believe Biggs.

This kind of controversy is especially likely in a society where political polarisation is acute and trust in government and the media is low. Such a society is at a disadvantage when it comes to tackling deepfakes.

Indeed, even if deepfake detection methods are accurate and widely used, they will be of limited aid in a society where a significant portion of the public dismiss verifications of authentic media and detection of fake media if they are contrary to their political and cultural priors.

'According to a survey conducted by King's College London, only 13% of people in the UK had a great deal/quite a lot of confidence in the press'

The sorry state of trust in institutions across the world should not detract researchers and investigators from working on tools and methods to detect deepfakes. However, such relatively low trust should temper the optimism of policymakers who would like technology companies to track and label deepfake content.

Trust in many British institutions is unfortunately low. According to a survey conducted by King's College London, only 13% of people in the UK had a great deal/quite a lot of confidence in the press, the second lowest rating among the 24 other countries included in the report.[62] A majority of British people do not trust political parties and almost half do not trust Parliament.[63]

Addressing the lack of trust in crucial institutions is beyond the scope of this paper. However, it should not go unnoticed by those who hold out hope that deepfake transparency and tracking requirements will protect against election disinformation and misinformation.

61 Chesney, B., & Citron, D. (2019). *Deep fakes: A looming challenge for privacy, democracy, and national security.* California Law Review. Link

62 Majid, A. (2023, March 30). *UK has second-lowest level of trust in press in survey of 24 countries.* Press Gazette. Link

63 Office for National Statistics. (2022). *Trust in government, UK: 2022.* Link

Conclusion

While we can look forward to valuable and worthwhile applications of deepfake technology, we should also be concerned about how it can be used for nefarious ends. The ongoing general election is only one example of an important event that risks being undermined by realistic fake footage that is cheap to create and spread across the internet.

Beyond the risks posed to critical institutions, abusers and criminals can use deepfake content to humiliate and extort innocent people and engage in fraud. It is not a surprise that in the face of these harms that lawmakers across the world have reached for regulation and legislation.

‘Lawmakers should resist the urge to implement technology-specific legislation and instead focus on updating existing legislation, policing the content rather than the technology used to create it’

However, there are risks associated with lawmakers taking aim at deepfakes. Deepfakes have many valuable uses, which risk being undermined by legislation or regulation. Many content-creation and alteration technologies such as the printing press, radio, photography, film editing, CGI, etc. pose risks, but lawmakers have resisted bans on these technologies because of these risks.

Lawmakers should also resist the urge to implement technology-specific legislation and instead focus on updating existing legislation, policing the content rather than the technology used to create it, and making the Government and industry as best-prepared as possible for harmful deepfakes – because there is no realistic way to turn back the tide.

CENTRE FOR POLICY STUDIES

Est. 1974

57 Tufton Street, London, SW1P 3QL
June 2024
ISBN 978-1-914008-52-8