# COUNTERING AI-GENERATED DEEPFAKES INTENDED TO UNDERMINE OUR ELECTIONS: AN UPDATE

- The US experienced its first AI-generated deepfake robocall effort targeting voters, which served as a wake-up call for lawmakers and regulators, and to a lesser extent election officials, who have for some months been bracing for election-related disinformation at scale. The fear is that deepfakes deployed just before the election will seek to discourage voters from voting or undermine trust in voting systems, with little or no time for media, the campaigns or election officials to debunk them.
- The Supreme Court will hear oral arguments tomorrow in a landmark case that will determine the extent to which agencies of the US government can liaise with the social media platforms as part of content moderation efforts.
- States are scrambling to regulate election-related deepfakes; federal legislation is less likely.
- There are a range of civil society and academic efforts underway to identify and rollout solutions to counter election-related disinformation that do not touch on content moderation and, therefore, are more likely insulated from lawsuits and Republican-led House investigations of alleged suppression of "conservative free speech."

Since I posted my last briefing note on the threats posed to our upcoming elections by generative AI technology and resultant deepfake tools, the United States experienced its first incident involving AI-driven robocalls, this one targeting voters during the January New Hampshire primary.[1] The United States thus joins a growing list of countries experiencing deployment of deepfakes, which also includes the United Kingdom, Nigeria, India, Sudan, Ethiopia and Slovakia. The good news is that the New Hampshire incident served not only as reminder, in the words of Lisa Monaco, Deputy Attorney General, that "AI is the ultimate double-edged sword," but also as a much-needed wake up call for lawmakers and regulators to understand the threat and act to counter it.

Technology is moving at lightning speed and, as often is the case, innovation is outstripping the ability of society to fully grasp the consequences of the technology, let alone determine whether, and if so how, it needs to protect itself from its more malign uses. Generative AI poses a variety of challenges – almost any malign actor can deploy it – cheaply, easily, more quickly and potentially at scale. That malign actor could be foreign or domestic, or foreign but appearing to be domestic, and need not be working for a campaign or candidate.

[1] *See also* two other previous briefing notes, "The Impact of Generative AI on Upcoming Elections: Disinformation and other malign influence operations coming our way at scale" and "Deepfakes: Driving the proliferation of fraud, deception and disinformation at warp speed."

And generative AI is far more sophisticated – deepfakes, in a significant departure from the use deployed to date, which largely has involved nonconsensual pornography,[2] need not purport to make it appear a candidate said something that he/she did not say or to have done something that he/she did not do, in each case to embarrass a candidate or undermine or otherwise try to damage the candidate's reputation. It could be far more subtle or nuanced – and it likely is not intended to sway votes (that would be too obvious), but to prompt potential voters not to vote or to deepen distrust across the electorate of election outcomes. In an election cycle where Donald Trump, at campaign rally, can praise the insurrectionists sentenced for their roles in the January 6th attack on the Capitol as "hostages" and "unbelievable patriots," can say that some migrants are "not people" and can predict a "bloodbath" for the country "if [he does] not get elected" (and this is only in the past 24 hours), or can invite Vladimir Putin "to do whatever the hell they want" to NATO countries that "don't pay," there is little left of a reputation from him to tarnish.[3]

This briefing note is timely for various reasons, including that tomorrow oral arguments will take place at the Supreme Court in respect of two so-called "jawboning" cases, the outcome of one of which will determine the extent to which US government agencies can liaise with the social media platforms regarding disinformation targeting Americans (whether foreign or domestic). More about this below.

## Harbingers of Election Interference to Come

In October 2023, two days before the national election in Slovakia, a deepfake audio recording appeared on Facebook. On the recording were two voices: allegedly, Michal Šimečka, who leads the liberal Progressive Slovakia party, and Monika Tódová from the daily newspaper Denník N. The recording appeared to be of a discussion of how to rig the election, partly by buying votes from the country's marginalized Roma minority. The post was deployed during the 48-hour blackout of media and campaign statements ahead of the

[2] There is a related threat tied to nonconsensual pornography, namely the risks posed by deepfakes to the participation of women in politics. A 2021 report issued by the Wilson Center (the lead author of which was Nina Jankowicz) highlighted the online abuse directed at women in public life. It found over 336,000 individual pieces of gendered and sexualized abuse posted by over 190,000 users directed at 13 research subjects across six social media platforms during a two-month period. (*See* Deepfakes and Elections: The Risks to Women's Political Participation). The report noted that malign creativity – the use of coded language, iterative, context-based visual and textual memes, and other tactics designed to avoid detection by social media platforms – is the greatest obstacle to detecting and enforcing against online gendered abuse and disinformation. This is a recurrent theme relevant to the broader election space.

[3] Never mind that in his speech in Ohio last night, during which the teleprompter seemingly failed, Trump struggled to pronounce some words and insisted that Joe Biden had beaten "Barack Hussein Obama" in every swing state. While experts also cite the threat posed to investigative journalists and other in the media in allowing politicians or their acolytes to disavow statements they actually made, claiming they are fake (the so-called "liar's dividend"), Trump only doubles down (even when his campaign tries to walk back some of the more egregious statements).

election so there was no way to debunk the recording. As it was audio, it apparently was not covered by Meta's policy, which only picks up AI-generated video that appears to show people saying things they did not say (both conditions must be satisfied for the content to be removed[4]). The recording is believed to have swung the election in favor of a pro-Russian candidate.

What caught the attention of experts were the timing of the release, highlighting the period when those voting on election are most vulnerable, that the deepfake involved the cloning of the voice of a journalist and that the operation had the hallmarks of a foreign influence operation.

Three months later, in New Hampshire, days before the January 23 primary, a robocall purporting to be the voice of President Biden urged voters to stay at home ("it's important that you save your vote for the November election"). The message concluded with the phone number belonging to the former chair of the New Hampshire State Democratic Party; that number had also been spoofed to make it appear as if the call was coming from her. The incident triggered an investigation by the office of the New Hampshire Attorney General.

NBC News reported that the person who created the deepfake did so in less than 20 minutes for a cost of only $1.00, using a tool available on the ElevenLabs platform (which is a text-to-speech generator). He reportedly used a reporter's voice instead of his own to circumvent the platform's terms of service. He was paid $150 by a Democratic consultant who did GoTV work for the Dean Phillips campaign, but reportedly was acting on this own initiative. The New Hampshire AG has identified two Dallas-based companies believed to be responsible for distributing the fake message to between 5,000 and 25,000 voters.

In early March, a BBC investigation identified deepfake videos circulating online that purport to show Donald Trump posing with Black voters. None of the videos were tracked back to the Trump campaign. The Center for Countering Digital Hate (CCDH) determined that at least one of the videos initially was created as satire, but since had been shared among Trump supporters.

## Impact of Deepfakes

Elections provide a target-rich environment for malign activity because election season and the accompanying emotions present prime opportunities for actors to exacerbate existing differences and exploit tribal dynamics. Into this cauldron, we find ourselves in the age of

[4] In February, the Meta Oversight Board upheld a decision by Meta to not remove a video edited to make it appear as if President Biden inappropriately touched his adult granddaughter's chest, accompanied by a caption describing him as a "paedophile." The post did not meet the two conditions. However, the Oversight Board went on to find that "Meta's Manipulated Media policy is lacking in persuasive justification, is incoherent and confusing to users, and fails to clearly specify the harms it is seeking to prevent. In short, the policy should be reconsidered." See also the op-ed by Helle Thorning-Schmidt, co-chair of the Oversight Board and former prime minister of Denmark, "We need to act on online disinformation now."

the democratization of AI, where fake videos, text messages and audio (which is the easiest to fake and offers fewer contextual clues to identify) are cheap, effective, deployable at scale and becoming more sophisticated by the day.[5] Add to this the ease with which malign actors can create fake social media user accounts and bots that appear to be human.

These and other disinformation tactics have three straight-forward objectives: to sow discord, suppress the vote and incite political violence. In contrast to the hundreds of millions of dollars spent on campaign ads, the goal of malign actors deploying deepfakes need not be to change anyone's mind – the objective could be to keep targeted audiences at home on Election Day[6] or to heighten distrust more broadly in the election results.

In early March, CCDH issued a report ("Fake Image Factories: How AI image generators threaten election integrity and democracy") highlighting how easy it is to generate election disinformation using Midjourney, OpenAI's DALL-E 3 through ChatGPT Plus, Microsoft's Image Creator and Starbility AI's DreamStudio, despite restrictions in their terms of service on content manipulation. The CCDH conclusions highlight two vulnerabilities – that generative AI platforms (the generators) will be unable to prevent the creation of misleading or false images (the deepfakes), and social media platforms (the delivery channels) will be unable to detect and remove the content before it reached online audiences. CCDH cited, as examples, fake images of ballots thrown in a dumpster, militia members "guarding" polling stations and voting machines being tampered with. These more generic images or videos could be harder for platforms to spot and take down.

And as I cited in an earlier briefing note, robocalls and robotexts present a separate threat that is beyond the scope of social medial – microtargeted calls can make a range of false assertions designed, in increasingly sophisticated nuanced fashion, to keep voters at home on Election Day. Introducing music or adding muffled background noise reportedly undermines detection tools.

[5] Initially, the concerns around deepfakes were for those with broad public profiles – in other words, someone with an extensive audio and visual profile on the internet. Today, with the ability of malign actors to create a full narration from one to two minutes of captured audio, the range of potential cloned voices increases exponentially. Consider, for example, the local election official or trusted community voice.

There are two forms of cloned speech – text-to-speech (actor uploads cloned speech and types the script for the fake output) or speech-to-speech (actor dictates speech and it is converted to create the fake output). A third model that is evolving is the hybrid avatar that can be made to look like it is speaking the cloned speech, which can be far longer than the fake messages superimposed on actual people in short video clips. (*See* "AI Audio Deepfakes Are Quickly Outpacing Detection" and "I Cloned Myself With AI. She Fooled My Bank and My Family.")

[6] We have actually seen this tactic before. According to a report produced by New Knowledge, researchers at Columbia University and Canfield Research LLC for the Senate Intelligence Committee, the Russian influence campaign ahead of the 2016 election used a range of tactics targeting Black communities to suppress the vote among Black (Democratic) voters. (*See* "Russian 2016 Influence Operation Targeted African-Americans on Social Media.")

The threats of deepfakes and other disinformation tactics are accelerating at an unimaginable pace as we approach the most consequential election of our time, when polarization is at unprecedented levels, the level of distrust in institutions, politicians, election systems and voting processes are at unimaginably high levels, turbo-charged by the ongoing effects of the Big Lie and continued election denial, as well as countless efforts to make access to voting more difficult in a number of states.  While a slew of malign actors, foreign and domestic, are prepared to weaponize cheap and easily accessible tools to upend the elections, content moderation and other efforts to understand and mitigate the impact of disinformation, including the role of algorithms, are under relentless pressure and attack.  This means that not only are the platforms less focused on mitigating the threats, but, at the same time, government efforts are constrained, and civil society and academic researchers are cowed.

Some experts believe that the risks of these threats are overstated (*see, e.g.,* "Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown").  These experts posit that voters will have largely made up their minds by the time malign tactics are unleashed at scale, that these tactics are not likely to target large cohorts and that disinformation represents only a small portion of the information that voters would typically consume.

That said, we all know that our elections will produce razor-thin margins and that it might not take much to swing an election, particularly if deepfakes target persuadable voters in the three to six likely battleground states.  Second, studies of disinformation have not tended to factor in the combination of micro-targeting with robo-technology and they have not factored in the quality and sophistication of disinformation generated by large language models.

The 2024 election faces a very different threat landscape than existed in 2016 and 2020, and even 2022. One need only look at a single statistic to understand the magnitude of the threat – approximately 70% of Republicans still believe that Joe Biden is an illegitimate president and that Trump won the 2020 election (*see, e.g.,* "CNN Poll: Percentage of Republicans who think Biden's 2020 win was illegitimate ticks back up near 70%" and "Poll: Majority of Iowa GOP caucus-goers don't believe Biden legitimately won in 2020").  Third, Russia and other malign actors have greater incentive given the current geopolitical crises that the world faces to seek to influence the outcome of our election than was the case in elections past.

Add to that list, as Protect Democracy pointed out ("How generative AI could make existing election threats even worse"), the online ecosystem is are more fragmented today and election administrators are under-resourced while facing threats to their and their families' personal safety (those threats incidentally have caused many to leave their positions, and with new officials working their first election in 2024 there is an increased risk of error that could feed the expected narrative that the elections are rigged).

## Fighting Back

There is a vacuum that needs urgently to be filled.  The vacuum has multiple causes.  First, none of the White House, Office of the Director of National Intelligence or the Department of Homeland Security – Cybersecurity and Infrastructure Security Agency ("CISA") seem to have the resources for attribution, which is tough, and for deterrence, which is impossible. The US government appears to have drawn a line between foreign and domestic actors,

though the tools would suggest that it takes little for foreign actors to appear domestic (*see* "Spate of Mock Sites With Russian Ties Pop Up in the U.S."). Also, it appears that while the government will support election officials, it will not weigh in on veracity. For their part, the platforms have gotten out of the business "of sitting with" the government; as Katie Harbath, a former public-policy director at Facebook, has noted "platforms are exhausted trying to adjudicate issues around political content. There's no clear agreement around exactly what the rules and penalties should be." (*See* "New Era of AI Deepfakes Complicates 2024 Elections.") More troubling, no one seems to be in overall charge.

And as noted above, due to a combination of threats of lawsuits and a wave of subpoenas from the Select Subcommittee on the Weaponization of the Federal Government, civil society and academic researchers have reduced their efforts to identity and track disinformation (*see, e.g.,* March 12th interview with Senator Mark Warner). As I have chronicled in prior briefing notes,[7] courts have drawn a line (in so-called "jawboning cases") beyond which government involvement violates the First Amendment, but I personally fail to see where most reported activity around content moderation rises to the level that that line has been crossed. On Monday, the Supreme Court will hear oral arguments in the most consequential case involving content moderation (*Murthy v. Missouri,* initially *Missouri v. Biden*), in which the court will be considering whether the government compelled social media platforms to remove "conservative free speech" in violation of the First Amendment.

At the Munich Security Conference, Adobe, Amazon, Google, IBM, Meta, Microsoft, OpenAI, TikTok and X announced a voluntary accord, with eight commitments (*see* Press Release). However, the platforms did not commit to ban or remove deepfakes. Rather, they agreed on methods they will use to try to detect and label deceptive AI content when it is created or distributed on their platforms. What is needed are algorithms that do not prioritize engagement above all else.

There are, however, many attempting to fill the vacuum, and a list of others who have a role to play – election officials and administrators, civil society groups, start-ups and others in the tech4democracy space that are focused on a range of solutions to detect deepfakes at scale, and democracy donors willing to fund innovative solutions. There will be a range of sports figures/veterans/media celebrities and other trusted figures that can play a role in educating voters. All this, however, has to be undertaken without creating panic in the electorate, which incidentally is one goal of malign actors.

## Legal and Regulatory Responses

Which brings us to legal and regulatory responses.

[7] *See* my December 2023, October 2023 and July 2023 briefing notes. *See also* Justin Hendrix and Ryan Goodman, "A Conspiracy Theory Goes to the Supreme Court: How Did Murthy v Missouri Get This Far?"

In August 2022 and ahead of the 2022 midterms, the House Committee on Oversight and Reform issued a report, "Exhausting and Dangerous": The Dire Problem of Election, that found that:

- Disinformation campaigns carried out by malicious domestic actors are eroding trust in American democracy and disrupting the operation of election offices.
- Election administrators have attempted to counter lies about election practices, but they lack adequate resources and funding.
- Misinformation led to violent death threats against local election officials, often inspired by comments from right-wing politicians and activists, leading many experienced officials to leave their positions.
- Election officials expressed concerns about dangerous, misinformation driven, so-called "election integrity" laws that threaten to undermine the voting process in future elections.
- Disinformation drove reckless and fraudulent audits.

The Majority Staff Report set out a series of recommendations for the Executive Branch and for the Legislative Branch:

- The President should designate a lead federal agency or office to support state and local efforts to counter election misinformation. All relevant federal agencies should use their authorities in coordination with the lead agency to support state and local election officials' efforts to counter misinformation during and after elections.
- CISA should continue to update its "Rumor Control" website to respond to national misinformation narratives. CISA's Mis-, Dis-, and Mal-information Teams should coordinate with state authorities to encourage state-level "Rumor Control" websites to counter misinformation spreading in their communities.
- The Department of Justice ("DoJ") should "aggressively pursue criminal and civil enforcement against those who threaten or harass election administrators." Its Election Threats Task Force should publicly clarify relevant legal definitions to local law enforcement regarding election security.
- Congress should pass legislation to address the funding gap for election officials across the country and to counter threats to election officials.
- Congress should provide emergency funding to the US Election Assistance Commission ("EAC"). The EAC oversees and enforces the Help America Vote Act ("HAVA"), which was passed in 2002 following the 2000 election, to improve voting systems and voter access for federal elections. EAC has authority to issue HAVA Election Security Grants to states.
- Congress should enact statutory penalties for anyone who threatens election officials and administrators.

Not surprisingly, with the exception of charges brought by the DoJ in cases involving threats against election workers, little headway has been made on these recommendations.

- In July, eight House Democrats introduced the Candidate Voice Fraud Prohibitions Act (H.R. 4611) to ban the distribution, with “actual malice,” of certain political communications containing “materially deceptive” audio generated by artificial intelligence that impersonate a candidate’s voice and are intended to injure the candidate’s reputation or to deceive a voter into voting against the candidate.
- In September, 27 Senators (all Democrats and one Independent) reintroduced the Election Worker Protection Act of 2023 (S. 1318). Senator Klobuchar had introduced the original bill in September 2022. The legislation would ban the distribution of materially deceptive AI-generated audio or visual media relating to candidates for federal office.
- In September, Senators Klobuchar and Collins, together with Senators Coons and Hawley, introduced the Protect Elections from Deceptive AI Act (S. 2770) to ban the use of AI to generate materially deceptive content falsely depicting federal election officials, in order to influence federal elections.
- In September, two House Democrats introduced the DEEPFAKES (Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to) Accountability Act (H.R. 5586) to protect national security against threats posed by deepfakes and provide legal recourse to victims of harmful deepfakes.

In February, in the aftermath of the New Hampshire robocall incident and prodding by Senators Klobuchar and Collins, the EAC voted unanimously to assist state and local election officials in combating the spread of AI-generated disinformation about federal elections. This would be accomplished by allowing election officials to use HAVA funds to:

- counter foreign influence in elections, election disinformation and potential manipulation of information on voting systems and/or voting procedures disseminated and amplified by artificial intelligence technologies; and
- for voter education and trusted information communications on correct voting procedures, voting rights and voting technology to counter AI-generated disinformation.

For its part, in response to the New Hampshire incident, the Federal Communications Commission (“FCC”) issued a Declaratory Ruling on February 8 to ban AI-generated voices in robocalls (voice cloning) under 1991 legislation, the Telephone Consumer Protection Act (“TCPA”). By bringing voice cloning within the TCPA, the FCC can fine or block companies that carry the calls; and victims can sue robo-callers that use AI. The FCC also is giving AGs new enforcement tools to prosecute. The FCC has Memoranda of Understanding with virtually all state Attorneys General.

This past week, Senators Klobuchar and Collins introduced the bipartisan Preparing Election Administrators for AI Act to require the EAC, in consultation with the National Institute of Standards and Technology, to develop voluntary guidelines for election offices.

Public Citizen’s Tracker of State Legislation on Deepfakes in Elections shows that 44 states have introduced legislation to address deepfakes. California (Oct. 2019), Indiana (Mar. 2024), Michigan (Nov. 2023), Minnesota (May 2013), New Mexico (Mar. 2024), Texas (Jun.

2019) and Washington (May 2023) have passed (in each case, on a bipartisan basis) some form of legislation.  Efforts in South Dakota and Tennessee have failed.  (*See also* Bill Track *50*.)  Legislation largely requires disclosure and speaks of "synthetic media," with definitions that vary, but largely focused on images, audio or video content that has been manipulated by means of generative artificial intelligence or other digital technology so as to create a realistic but false image, audio or video.

A nationwide poll conducted by the University of Chicago Harris School of Public Policy and The Associated Press-NORC Center for Public Affairs Research in October found broad bipartisan support for actions to address the use of AI, including a federal government ban on false or misleading AI images in political ads, though support generally is higher among Democrats than Republicans (Democrats 78%; Republicans 66%).

## Concluding Thoughts

The platforms on which AI tools are accessible, the social media platforms though which deepfakes can be disseminated, lawmakers and regulators will need to move quickly to reduce the likelihood that deepfakes targeting voters can be deployed at scale.  Election officials need the resources to educate voters in their states and these officials should rapidly expand the circle of trusted voices across their communities to heighten awareness of how and when to vote, how and when ballots will be counted, and whose pronouncements regarding elections are to be trusted.

It would be highly unfortunate were the Supreme Court to rule against the administration in the *Murthy* case, and assuming it finds no First Amendment violations, then hopefully the ruling will encourage civil society and academic researchers to resume their efforts to identify and counter all forms of online disinformation.

The good news is that civil society and academia are focused on a range of other solutions, including advocating for platform design modifications (*see, e.g.*, "Introducing the Neely Center Design Code for Social Media"), supporting the design of digital democracy tools (*see, e.g.*, "Defending Democracy with New Deliberative Tools") and advising election administrators on how best to optimize offensive and defensive tactics to tackle deepfake threats (*see, e.g.*, "The AI + Election Security Coalition"), none of which touches on content moderation.  These efforts deserve attention, amplification and financial support.

The media and society at large also have roles to play, the former in educating voters and the latter in taking those messages onboard.

Lastly, on the assumption that the most damaging deepfakes will be deployed 48-72 hours ahead of Election Day, encouraging voters to vote early, by mail or otherwise, will greatly mitigate deepfake efforts to persuade voters to stay home.

*     *     *

**Mark S. Bergman**
**7Pillars Global Insights, LLC**
**Washington, D.C.**
**March 17, 2024**

No portion of this briefing note was generated by AI.