CENTER FOR AN INFORMED PUBLIC
UNIVERSITY of WASHINGTON

DEEPFAKES IN THE 2020 ELECTIONS AND BEYOND

Lessons from the 2020 Workshop Series

CENTER FOR AN INFORMED PUBLIC
UNIVERSITY OF WASHINGTON

OCTOBER 2020

INTRODUCTION

In July 2020, the University of Washington's Center for an Informed Public (CIP) and Microsoft's Defending Democracy Program convened a three-part workshop with experts from the technology industry, media organizations, government, and academia to discuss the state of media manipulated by Artificial Intelligence (AI), also known as deepfakes. The invited participants included representatives from major tech companies and social media platforms, academia and think tanks, major international, national and regional news organizations, fact-checking groups, civil society organizations, and elected officials and government technology professionals.

The workshops' objective was to discuss how to plan for the presence of deepfake technology and guard against it adversely affecting the 2020 U.S. presidential election. Propelled by the staggering spread of less technologically sophisticated misinformation and disinformation during the presidential election in 2016, experts and citizens alike have expressed growing concern about the potential for deepfakes to make it even more challenging to distinguish between authentic and manipulated media.

The three days of virtual roundtable discussions were structured to look at the deepfake issue through three distinct lenses: technology industry, journalism, and law and policy. In each session, participants examined the scope and potential impact of deepfakes on the upcoming election period, identified adverse effects, and discussed potential actions that various stakeholders could take to prevent such adverse effects.

This report highlights four themes that emerged from the discussions and includes supplementary information from other key work that has been done around the topic of synthetic media. In an associated report, we go into greater detail on these four themes and two others.

DETECTING DEEPFAKES IS CHALLENGING.

The technology to detect deepfakes, and synthetic media more broadly, is imperfect, difficult to deliver at scale and speed, and still evolving.

AI-enabled detection can result in numerous false positives, as was demonstrated by the Deepfake Detection Challenge (DFDC) held earlier this year. High rates of false positives could "overflow any human review in the process." The risk of false positives is important to consider because as deepfake technology increases in availability, even relatively small error rates could lead to large amounts of media for human review, something that not every organization has the resources to accomplish.

Automated detection is also hampered by the importance of determining the intent of why a deepfake was originally created and deployed. "Not every deepfake is bad, and not every deepfake is designed to sway the election," one participant said. Since automated detection won't necessarily be able to determine intent, some level of human review will be needed not only to assess intent but also to figure out whether the detected deepfake may be connected to a larger disinformation campaign.

Detection efforts are complicated by the fact that even as detection techniques improve, that improvement doesn't mean the detection problem has been solved. As one organizer said: "Even if you get to a point where detection improves, you have to constantly update it" to react to the continued evolution of deepfake technology. Any detection technology will have only a short shelf life before adversarial algorithms learn how to evade it.

NEWS ORGANIZATIONS NEED RESOURCES AND TOOLS TO SCRUTINIZE IMAGES, VIDEO AND AUDIO RECORDINGS.

Journalists and news organizations need training, support, and resources to better detect and act on identified problematic deepfakes.

Manipulated media poses fundamental challenges to the work of journalists. Currently, cheapfakes are far more prevalent than deepfakes and run the gamut from obviously edited memes to more subtle alterations such as the May 2019 slowed-down Nancy Pelosi video. As synthetic media becomes more widely accessible, newsgathering organizations and journalists will need to be even more careful in scrutinizing and vetting material. Reporters are less and less able to look at a photo or video and determine whether it has been edited, which means investing in training, resources, and partnerships with media forensic experts is the best way to successfully confront the risks posed by harmful synthetic media. Not only do newsrooms and reporters need to vet the media themselves, but they must also maintain public confidence that the information they report on is reliable. As one newsroom manager said during the journalism workshop: "The verification of images, video or audio is a big challenge. … Almost surely, someone in our audience will ask: Is this real or isn't it real?"

This comment poignantly encapsulates not just a need to adequately confront the threat of synthetic media, but the fact that the very real risk of reporting, sharing or otherwise amplifying disinformation, misinformation and false claims comes at a time when public trust in media organizations continues to decline in the U.S. It also comes as news organizations face acute financial instability, budget pressures, strained resources, industry consolidation and a rapid decline in newsroom headcounts thanks to layoffs, furloughs and restructurings.

It is in this environment that many organizations will need to add to their repertoire the ability to identify synthetic media accurately and efficiently. This challenge is difficult because while journalists are trained to seek out the truth, most are not trained in the technically challenging media forensics work needed to identify manipulated content. In order to support journalists, particularly those in newsrooms that are already under resourced, media organizations need relationships with experts who can review and assess what's going on and explore partnerships around how to verify that photos, videos and audio recordings haven't been manipulated, distorted or synthetically generated.

ACTIONABLE POLICY IS NEEDED AT THE STATE AND LOCAL LEVELS WHERE LEGISLATION AND REGULATION HAS BEEN LIMITED.

The U.S. legal and regulatory landscape regarding deepfakes may look very different in the not-so-distant future at both the state and federal levels as policymakers gain a better understanding of what deepfakes are and what threats they pose. organizations need training, support, and resources to better detect and act on identified problematic deepfakes.

In addition to technical and educational strategies, there is opportunity to better use legislation as a tool, both at the state and federal level. According to a Washington, D.C.-based legal analyst who tracks deepfake-related legislation,

currently, there aren't a lot of examples of state or federal laws and regulations addressing deepfakes, but the prediction is that will likely change in the coming years.

Three states — California, Texas, and Virginia — have deepfake-related laws on the books, and approximately a dozen more have legislation pending. Nearly all of these laws and legislation are civil in nature and address "actual malice related to the intent to deceive and the knowledge of deceivery," according to the analyst. In other words, to successfully prosecute a creator under the law, the deepfake would not only have to be deceptive, but you would have to prove that it was intended to deceive.

What many of the current laws don't specifically address is the fact that the vast majority of deepfakes exist in the form of non-consensual pornography that disproportionately impacts women, a finding identified by Sensity in their *State of Deepfakes 2019* report. According to Sensity, 96% of deepfakes in 2019 were pornographic in nature. In recognition of this, Virginia lawmakers expanded an existing ban on non-consensual pornography to images of people "whose image was used in creating, adapting, or modifying a videographic or still image with the intent to depict an actual person and who is recognizable as an actual person by the person's face, likeness, or other distinguishing characteristic."

At the federal level, changes are being made as well. The most recent National Defense Authorization Act, the defense appropriations omnibus bill covering Fiscal Year 2020, signed by President Trump in December 2019, includes provisions that mandate the federal government to create a comprehensive report on the foreign weaponization of deepfakes. The act requires the federal government to "notify Congress of foreign deepfake-disinformation activities" targeting U.S. elections and establishes a "Deepfakes Prize" competition to encourage research and development of deepfake-detection technology.

THE 'LIAR'S DIVIDEND' IS JUST AS FORMIDABLE AS DEEPFAKES.

The idea that the mere existence of deepfakes causes enough distrust that any true evidence can be dismissed as fake is a major concern that needs to be addressed.

One of the most concerning and challenging problems associated with deepfakes is the "Liar's Dividend," the idea that the mere existence of a deepfake video causes enough distrust that any actual evidence can be dismissed as fake, either by merely calling it a deepfake or releasing synthetically manipulated content that is claimed to be real instead of the actual footage. This problem is troubling, and in some ways, unavoidable – after all, we cannot stop the creation of synthetic media altogether, even if that were desirable, which it is not. The challenge is to prepare for synthetic media without allowing unethical actors to sow doubt and distrust, especially in a time when distrust in media is already so high.

Preparing for the Liar's Dividend's consequences is not just a theoretical exercise: in June 2019 compromising footage of the Malaysian Minister of Economic Affairs Azmin Ali was allegedly captured on video. In response to the accusations, Ali claimed that the video was a deepfake, something that was not able to be confirmed by experts who examined the footage.

It is not difficult to imagine similar scenarios playing out in the 2020 U.S. election. During one of the workshops, participants engaged in the following thought experiment: What would happen if something like the infamous "Access Hollywood" tape, where Donald Trump is heard making disparaging and offensive comments about women in an audio recording, were to happen in the 2020 elections? The person leading the experiment observed that "In the moment in the past, the very media artifact was the truth," but with the Liar's Dividend, "I get to claim any video is fake," even if it's

been authenticated as real, the newsroom manager said. "The next 'Access Hollywood' tape will be challenged as a deepfake" even if it hasn't been manipulated or distorted.

CONCLUSION

Because information consumers – and voters in the upcoming elections – are their own last line of defense against the forces of disinformation and misinformation, including synthetic media, educational outreach is both a short-term need and a long-term necessity for tech-sector stakeholders, news organizations and policymakers.

It is important to remember, though, that just as it is technically challenging to detect a deepfake video and determine its provenance and intent, educating the public to be aware of deepfakes, the Liar's Dividend, and disinformation and misinformation dynamics that impact beliefs and actions is a tall order that won't be accomplished overnight.