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.

CENTER FOR AN INFORMED PUBLIC
UNIVERSITY of WASHINGTON

DEEPFAKES IN THE 2020 ELECTIONS AND BEYOND:

Lessons From the 2020 Workshop Series

STEPHEN PROCHASKA, MICHAEL GRASS, JEVIN WEST

CENTER FOR AN INFORMED PUBLIC
UNIVERSITY OF WASHINGTON

OCTOBER 2020

EXECUTIVE SUMMARY

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 AI manipulated media, also known as deepfakes[1]. 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 and prevent deepfake technology from 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 separate fact from fiction[2].

The three days of virtual roundtable discussions were structured to look at the deepfake issue through three distinct lenses: 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 details six themes that emerged from the discussions and includes supplementary information from other key work that has been done around the topic of synthetic media.

1. **Deepfakes are an extension of an already existing issue:** Deepfakes need to be contextualized within the broader framework of malicious manipulated media, computational propaganda and disinformation campaigns.
2. **Deepfakes present a multidimensional problem that demands a collaborative, multi-stakeholder response:** Deepfakes, like other forms of mis- and disinformation, present a multidimensional problem that require experts in every sector to align research and implementation efforts, particularly in finding solutions that can be deployed in real time.
3. **Detecting deepfakes is hard:** The technology to detect deepfakes, and synthetic media more broadly, is imperfect, super hard to deliver at scale and speed, and still evolving.

[1] We use the term "deepfake" throughout the report, but we also recognize the increasing usage of "synthetic media" as the catch-all term for AI-manipulated media and data. This includes deepfakes but could also include text generation, speech, etc.

[2] Jennifer Finney Boylan, "Opinion | Will Deep-Fake Technology Destroy Democracy?," *The New York Times*, October 17, 2018, sec. Opinion, https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html; Samantha Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked," *Vice*, December 11, 2017, https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn; Franklin Foer, "The Era of Fake Video Begins," The Atlantic, April 8, 2018, https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/; Daniel Funke, "A Potential New Marketing Strategy for Political Campaigns: Deepfake Videos," *Poynter* (blog), June 6, 2018, https://www.poynter.org/fact-checking/2018/a-potential-new-marketing-strategy-for-political-campaigns-deepfake-videos/; Britt Paris and Joan Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence" (Data & Society, September 18, 2019), https://datasociety.net/library/deepfakes-and-cheap-fakes/; Joshua Rothman, "In the Age of A.I., Is Seeing Still Believing?," *The New Yorker*, November 5, 2018, https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing; James Vincent, "US Lawmakers Say AI Deepfakes 'Have the Potential to Disrupt Every Facet of Our Society,'" The Verge, September 14, 2018, https://www.theverge.com/2018/9/14/17859188/ai-deepfakes-national-security-threat-lawmakers-letter-intelligence-community; Charlie Warzel, "He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse.," BuzzFeed News, February 11, 2018, https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news.

4. **Journalists and factcheckers need the tools to scrutinize images, video and audio recordings more carefully:** Journalists and news organizations need training, support and resources to better detect and act on identified problematic deepfakes.
5. **Policy needs to evolve in an informed way:** 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.
6. **The Liar's Dividend is challenging:** 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.

I. MALICIOUS DEEPFAKES ARE AN EXTENSION OF AN ALREADY EXISTING ISSUE.

While we gathered to discuss deepfakes in the 2020 election, we needed to take a step back and put deepfakes in context by placing them within the broader framework of manipulated media. Deepfakes pose a unique and intimidating challenge for society. Still, it is essential to remember that it doesn't take a high-quality, hard-to-detect deepfake video to deceive, mislead, and confuse. "I think the 'cheapfakes' that we've seen … can be as damaging as anything that's synthetic," one participant said. "Deepfakes are an important threat, but they've gotten more attention because they're a bit sexier." This is an important recognition, and it is one that many experts and organizations have identified and discussed as synthetic media has evolved[3]; but it bears repeating as we grapple with how best to communicate about deepfakes and the challenges they pose.

Beginning with the initial session and running through each subsequent workshop, the need to integrate solutions for detecting and preventing the spread of synthetic media with strategies preventing the spread of so-called "cheapfakes" (non-AI generated/altered manipulations) was apparent. Just as with the workshops, this theme is evident throughout this report for a reason: as we discuss synthetic media, it is easy to become unnerved by the potential of the technology, especially as more of the public learn about synthetic media who may not fully understand it.

It is, therefore, essential to understand that while synthetic media may be intimidating, non-AI manipulated media is something that many organizations have been trying to address for quite some time[4]. And through their efforts they are making progress in detection, provenance, and policy for both AI and non-AI manipulated media.

[3] Britt Paris and Joan Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence" (Data & Society, September 18, 2019), https://datasociety.net/library/deepfakes-and-cheap-fakes/; Partnership on AI, "The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity," March 12, 2020, https://www.partnershiponai.org/a-report-on-the-deepfake-detection-challenge/.

[4] Britt Paris and Joan Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence" (Data & Society, September 18, 2019), https://datasociety.net/library/deepfakes-and-cheap-fakes/; Reuters, "Identifying and Tackling Manipulated Media," Reuters, accessed August 14, 2020, https://www.reuters.com/manipulatedmedia/en/; Matt Turek, "Media Forensics," DARPA, accessed August 15, 2020, https://www.darpa.mil/program/media-forensics; "Expanding Fact-Checking to Photos and Videos," *About Facebook* (blog), September 13, 2018, https://about.fb.com/news/2018/09/expanding-fact-checking/; "The Washington Post's Guide to Manipulated Video," Washington Post, accessed August 15, 2020, https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/.

II. DEEPFAKES PRESENT A MULTIDIMENSIONAL PROBLEM THAT DEMANDS A COLLABOTIVE, MULTI-STAKEHOLDER RESPONSE.

There was a recognition in all three sessions that there is no single solution, technological or otherwise, to address the complex set of challenges deepfakes, synthetic media, and other forms of manipulated media present. While technology companies play critical roles in mitigating the impacts of synthetic media shared and distributed through their platforms, there is a strong need for policymakers to define legal and regulatory remedies clearly. Media organizations must invest in how best to spot manipulated images, video, and audio meant to deceive or misinform consumers. Whether through investing in training and media forensic resources, or creating partnerships with expert organizations, it is clear that journalists will need to adapt quickly to avoid accidentally amplifying synthetic media.

Perhaps most importantly, there was almost unanimous agreement in all three sessions that improving public awareness of deepfakes and synthetic media is essential. Still, there was no clear consensus about what an effective public outreach and education strategy would look like and what it would need to be successful. The existence of the Liar's Dividend, the idea that the mere presence of deepfakes causes enough distrust that any valid evidence can be dismissed as fake, further complicates the challenges posed by an effective outreach and education strategy. Any attempt to educate the public, then, will need to consider not only how to show people what deepfakes are, but also how they can be critical consumers of information so they are not duped by a public figure’s attempt to deny the veracity of real video, audio, or photos.

While there are already great examples of outreach and awareness around deepfakes and synthetic media, including Reuters’ manipulated media guide and moondisaster.org,[5] there is more work to be done to educate and inform the public not only about deepfakes, but also about information literacy more broadly. To that end, a September 1, 2020, public forum and a quiz developed by UW’s Center for an Informed Public and Microsoft’s Defending Democracy Program titled Spot the Deepfake are just a starting point for education and outreach around this complex issue.

On the technology side, AWS, Facebook, Microsoft, the Partnership on AI’s Media Integrity Steering Committee, and academics came together in September 2019 to build the Deepfake Detection Challenge (DFDC) in order to “accelerate development of new ways to detect deepfake videos.”[6] With this challenge recently concluded, there were a large number of insights discussed during the workshops that were learned from people’s attempts to consistently and efficiently detect deepfakes.[7] While we discuss this challenge in greater detail in Section 3 of this report, one of the primary takeaways was that detection alone is insufficient to solve the problem of synthetic media and problematic information more broadly.

Even just within the realm of detection it isn’t always a simple exercise in determining whether something is "fake" or "real." The truth of a piece of media, synthetic or otherwise, often falls somewhere in between those two

[5] Other resources include MIT Media Lab’s DetectFakes website as well as a collection of materials from WITNESS

[6] “Deepfake Detection Challenge Dataset,” Facebook AI, accessed August 15, 2020, https://ai.facebook.com/datasets/dfdc.

[7] For an in depth breakdown of specific insights and recommendations from the DFDC see the Partnership on AI’s report: Partnership on AI, “The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity,” March 12, 2020, https://www.partnershiponai.org/a-report-on-the-deepfake-detection-challenge/.

choices. “We see this in the fact-checking industry with scales like ‘number of Pinocchios’ or ‘somewhat false,’” one participant said. “It's tough enough to explain what a deepfake is to the public, let alone keep their attention long enough to explain the nuances.” In other words, the truth can be complicated, so when it comes to detecting and preventing the spread of harmful synthetic media organizations have to consider a lot more than just whether a video was synthetically generated or altered. Instead they have to consider factors such as the context and intent of deepfakes before being able to take any action to remove potentially harmful or misleading videos or audio.

More broadly speaking, the workshops highlighted steps that can be taken to establish and solidify partnerships and opportunities for collaboration. Namely the need to bring experts and organizations together in informal settings like the workshops as well as collaborations like the DFDC. These kinds of events and partnerships will be essential in creating established networks to better detect and remove harmful synthetic media across multiple platforms.

Perhaps one of the most important outcomes that could come from more vital collaboration within and across industries is an increase in the speed with which deepfakes and other examples of synthetic media can be detected and removed from multiple platforms if determined to be harmful. Frequently harmful information can travel exceptionally fast, and the damage is done before fact-checkers can debunk the information and distribute a correction.[8] By the time a video is factchecked, the damage has been done. Time is therefore of the essence in not just detecting deepfakes, but also taking action when a deepfake could cause harm. This is complicated by the fact that the media can be reuploaded to numerous platforms, be reposted, captured in a screenshot, etc. Because of the numerous paths available to post deepfakes, solutions and preventative measures need to be fast and coordinated or they risk always lagging behind the false information.

Overall, the workshops made it clear that the only way to successfully confront the spread of malicious synthetic media, and disinformation more broadly, is to bring together people and organizations from numerous stakeholder groups so that they can successfully and meaningfully collaborate in the detection and prevention of the spread of harmful synthetic media. This means not just involving the technology industry, media organizations, and policy makers, but also international stakeholders, including citizen and civil society groups, that could be affected by the spread of harmful synthetic media.

While much of our discussion was in the context of the 2020 U.S. presidential election, it is important to remember that in the information age we are all connected and susceptible to the same misleading information in more than just elections.

III. DETECTING DEEPFAKES IS HARD.

During the workshops, there was an in-depth presentation and discussion of the Deepfake Detection Challenge (DFDC) held earlier this year. As touched on in Section 2, the DFDC was a collaborative effort between industry (Microsoft, Facebook, and Amazon), non-profits (Partnership on AI), and academic researchers. The goal was to speed up the development of deepfake detection. The DFDC provided a cash prize to the winners to recruit more researchers to detection efforts. Kaggle hosted the public dataset provided by the partners. The competitors trained their algorithms against this data set, but the evaluation of the algorithms was conducted on an unseen,

[8] D. N. N. Media, “Fact Checking Is Important, and Here’s Why,” Medium, February 24, 2018, https://medium.com/@dnnmedia/fact-checking-is-important-and-heres-why-66bfe76c8e55.

black box dataset to simulate the unpredictable nature of media in the real world. Overall, 2,265 competitors applied their methods to a large face-swap dataset — 130 terabytes of data including more than 100,000 total clips sourced from 3,426 paid actors who consented to have their likenesses modified[9] — to help refine their detection models. The Partnership on AI's Media Integrity Steering Committee released a report in March 2020 detailing insights and recommendations based on their experience with the DFDC.[10]

According to one of the industry participants, AI-enabled detection in the DFDC resulted in numerous false positives – high rates of which 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. By the end of the challenge, the top-performing model against the public data set had an 82.56% average precision rate – but when tested against the black box set the top-performing model (ranked 4th against the public data set) achieved only 65.18% average precision.[11] This discrepancy points to one of the primary problems with detection discussed in the workshops (and subsequent papers/reports[12]): the difficulty of generalizing detection techniques to unknown and unforeseen synthetic media examples.

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 the intent definitively, 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. For example, early on, one of the strategies for identifying deepfakes was to watch how often the person blinked, as people in deepfakes tended not to blink as often as a normal person.[13] Consequently, shortly after identifying the "tell," deepfake videos included more natural blinking. The technology has advanced dramatically since then, but the underlying adversarial nature of production vs. detection remains the same. Any detection algorithm can be used to improve the creation of deepfakes. Ultimately, the difficulties in detection, coupled with the fact that there is no apparent end in sight for detection strategy updates, reinforces the need for solutions outside of just detection.

To that end, determining the provenance of a deepfake, and manipulated media more generally, is a strategy that is being tested as a complement to standard detection techniques. One of the participants noted provenance – the origins and authenticity of a particular photo, video, or audio file – is difficult; it requires skilled experts and "a lot of things have to be aligned" to happen. In order to align as many resources as possible to this end, multiple initiatives are underway. One of the initiatives discussed in the workshops was ProjectOrigin,[14] a collaborative undertaking by *The New York Times*, BBC, CBC, and Microsoft to place a digital "watermark" on media originating

[9] Kaggle, "Deepfake Detection Challenge," Deepfake Detection Challenge, accessed August 14, 2020, https://kaggle.com/c/deepfake-detection-challenge; "Deepfake Detection Challenge Dataset," Facebook AI, accessed August 15, 2020, https://ai.facebook.com/datasets/dfdc.

[10] Partnership on AI, "The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity," March 12, 2020, https://www.partnershiponai.org/a-report-on-the-deepfake-detection-challenge/.

[11] "Deepfake Detection Challenge Dataset," Facebook AI, accessed August 15, 2020, https://ai.facebook.com/datasets/dfdc.

[12] Brian Dolhansky et al., "The DeepFake Detection Challenge Dataset," *ArXiv:2006.07397 [Cs]*, June 24, 2020, http://arxiv.org/abs/2006.07397.

[13] Siwei Lyu and Siwei Lyu, "The Best Defense against Deepfake AI Might Be . . . Blinking," Fast Company, August 31, 2018, https://www.fastcompany.com/90230076/the-best-defense-against-deepfakes-ai-might-be-blinking.

[14] Sara Fischer, "Project Origin: News and Tech Coalitions Partner to Fight Fake News in U.S. Election," *Axios*, July 14, 2020, https://www.axios.com/project-origin-election-fake-news-06fe6c7b-3a5e-48da-ab30-5f1f938706d3.html.

from authentic content creators that degrades if the content is manipulated. Project Origin complements current detection techniques for synthetic media and is one of several broader initiatives working against misinformation and disinformation, including the News Provenance Project[15] and Trusted News Initiative.[16] Collectively, they are poised to play an important role in preventing the spread of harmful deepfakes, and harmful information more generally, in the lead up to the 2020 election.

It is clear from the work in detection that while researchers are making progress, there is still a long way to go. Perhaps more importantly than simple detection, there needs to be progress in developing rapid mitigation strategies to respond to detected deepfakes, better understand their provenance, and prevent manipulated media reuploading across different platforms.

IV. JOURNALISTS AND FACTCHECKERS WILL NEED THE ABILITY TO SCRUTINIZE IMAGES, VIDEOS AND AUDIO RECORDINGS MORE CAREFULLY.

Manipulated media poses fundamental challenges to the work of journalists and newsgathering organizations. 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 convince the public 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.[17] 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.[18] 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

[15] “The News Provenance Project,” The News Provenance Project, accessed August 15, 2020, https://www.newsprovenanceproject.com.

[16] “BBC - Trusted News Initiative (TNI) Steps up Global Fight against Disinformation with New Focus on US Presidential Election - Media Centre,” accessed August 15, 2020, https://www.bbc.co.uk/mediacentre/latestnews/2020/trusted-news-initiative.

[17] Gallup/Knight Foundation, “American Views 2020: Trust, Media and Democracy” (Gallup/Knight Foundation), accessed August 14, 2020, https://knightfoundation.org/reports/american-views-2020-trust-media-and-democracy/.

[18] 1615 L. St NW, Suite 800Washington, and DC 20036USA202-419-4300 | Main202-857-8562 | Fax202-419-4372 | Media Inquiries, “U.S. Newspapers Have Shed Half of Their Newsroom Employees since 2008,” *Pew Research Center* (blog), accessed August 15, 2020, https://www.pewresearch.org/fact-tank/2020/04/20/u-s-newsroom-employment-has-dropped-by-a-quarter-since-2008/; “Above & Beyond - Looking at the Future of Journalism Education by Dianne Lynch,” accessed August 15, 2020, https://knightfoundation.org/features/journalism-education/.

content. In order to support reporters, 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.

As is exemplified by the Trusted News Initiative and the News Provenance Project, several of the larger international and national news organizations have already made newsroom investments in resources and partnerships needed for media forensics work. While these larger news outlets might have the staff resources and tools to do that type of work, most local news organizations don't. So deepfakes may be more difficult to detect in local news ecosystems where there is less scrutiny. It is that much more important that relationships exist[19] between the tech industry and academic communities and resources are available that allow local newsrooms to successfully combat the spread of harmful synthetic media. Some proposals from the workshops were that any media forensic tools developed to examine potential deepfakes and other synthetic media content should be available for free or have a low price point as well as be easy to use so that reporters can use them with only limited training.

While the tools for detection, authentication, and provenance are not perfect, they do exist and are improving. Some examples include Sensity's "Sensity Detection API" and work being done by Owen Myer and Matthew Stamm on forensic similarity graphs.[20] It remains to be seen what will be the most successful or widely adopted, but researchers and organizations have invested heavily in finding ways to detect synthetic media, and in media forensics more generally.[21] The challenge moving forward will be in getting the tools into the right hands and making them easy to use for people unfamiliar with technology, particularly in the case of organizations where resources are limited.

With the November 2020 elections rapidly approaching, news organizations don't have much time to ramp up training programs and establish the teams and partnerships needed to identify synthetic media and other manipulated content. The risks posed from deepfakes and other content meant to confuse and distort the truth and feed into a longer-term conversation about how news organizations need to adapt their newsgathering, vetting and fact-checking processes a future where it will be more difficult to know what is real and what is not. In particular, access to effective tools for detection and provenance will be essential. Whether that happens through collaboration with other organizations or becomes a new part of what it means to be a journalist remains to be seen.

V. POLICY NEEDS TO EVOLVE IN AN INFORMED WAY.

In addition to technical and educational strategies, one of the essential tools that needs to be effectively utilized is legislation, 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. What became clear during

[19] In a general sense, this is already being done; for example, the Associated Press has programs specifically tailored to help members with smaller newsrooms.

[20] Owen Mayer and Matthew Stamm, "Exposing Fake Images with Forensic Similarity Graphs," accessed August 15, 2020, https://omayer.gitlab.io/forensicgraph/.

[21] Stephanie Kampf and Mark Kelley, "A New 'Arms Race': How the U.S. Military Is Spending Millions to Fight Fake Images | CBC News," *CBC*, November 18, 2018, https://www.cbc.ca/news/technology/fighting-fake-images-military-1.4905775.

the workshop is how unfamiliar legislators are with the nuances of deepfakes if the legislators have not come into contact with deepfakes previously.

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.[22] 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."[23]

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[24] 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.

In order to ensure that future legislation is written in such a way that it takes into account the nuances of synthetic media, collaborative communication needs to be established with experts in academia, the technology industry, and journalism to educate lawmakers. A lawmaker present in the workshop stated that there "absolutely needs to be an effort to educate state legislators," but that can be challenging in states with short legislative sessions where "we don't have time to dive into" all the policies. In addition to more outreach to legislators, it was noted that regulators responsible for enforcement would also need to be better informed, otherwise the laws will be ineffective.

[22] Sensity, "Mapping the Deepfake Landscape," October 7, 2019, https://sensity.ai/mapping-the-deepfake-landscape/.

[23] Marcus Simon, "HB 2678 Unlawful Dissemination or Sale of Images of Another; Penalty.," Pub. L. No. 2678, accessed August 14, 2020, https://lis.virginia.gov/cgi-bin/legp604.exe?191+sum+HB2678.

[24] Matthew Ferraro, Jason Chipman, and Stephen Preston, "First Federal Legislation on Deepfakes Signed Into Law," Wilmer Hale, accessed August 14, 2020, https://www.wilmerhale.com/en/insights/client-alerts/20191223-first-federal-legislation-on-deepfakes-signed-into-law.

VI. THE LIAR'S DIVIDEND IS CHALLENGING.

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.[25] 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.[26]

It is not difficult to imagine similar scenarios playing out in the 2020 U.S. election. 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 election?[27]

During one of the workshops, participants engaged in the above thought experiment and after describing the scenario, the person leading the experiment observed that "In that 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.

It is important to remember that not everyone has to believe a deepfake is real for it to have an effect. After all, we know that creating doubt is a major strategy of disinformation campaigns,[28] and now, with the mere existence of synthetic media, it is even more difficult to know what and whom to trust. Malicious actors can capitalize on the Liar's Dividend by creating deepfakes that are just convincing enough to make people question whether it is real. Furthermore, for a malicious deepfake to have an effect, it doesn't need to reach massive audiences. As one of the participants in the workshops noted, "When I say viral, it doesn't need to be millions of views."

The collaborative work discussed in this report on detection and provenance (among others) is vital not just in identifying malicious deepfakes, but in establishing a trusted relationship with the public. It is critical for

[25] Kelly McBride, "The 'Liar's Dividend' Is Dangerous for Journalists. Here's How to Fight It.," *Poynter*, May 17, 2019, https://www.poynter.org/ethics-trust/2019/the-liars-dividend-is-dangerous-for-journalists-heres-how-to-fight-it/.

[26] Sensity, "Mapping the Deepfake Landscape," October 7, 2019, https://sensity.ai/mapping-the-deepfake-landscape/.

[27] Jane Timm, "Trump Caught on Hot Mic Making Lewd Comments about Women in 2005," *NBC News*, October 7, 2016, https://www.nbcnews.com/politics/2016-election/trump-hot-mic-when-you-re-star-you-can-do-n662116.

[28] Kate Starbird, "Disinformation Campaigns Are Murky Blends of Truth, Lies and Sincere Beliefs – Lessons from the Pandemic," *The Conversation*, July 23, 2020, http://theconversation.com/disinformation-campaigns-are-murky-blends-of-truth-lies-and-sincere-beliefs-lessons-from-the-pandemic-140677; Kate Starbird, Ahmer Arif, and Tom Wilson, "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 1–26, https://doi.org/10.1145/3359229.

technology companies and news organizations to say with as much certainty as possible what is a deepfake and what is not, or the public will not be able to trust the information they provide.

In addition to working together to verify media authenticity, educating the public about synthetic media and its ramifications, including the Liar's Dividend, is vital. As one participant identified, "We have to teach people to be cynical" about the information they consume before they share and amplify it, "We should start focusing on education and media literacy … for consumers." The challenge with teaching consumers about synthetic media, and mis/disinformation more broadly, is ensuring that they have a healthy level of skepticism, but not so much that they are unwilling to trust or believe anything that doesn't align with their world view, or worst case scenario, they become too cynical and lose trust in everything.

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 and the Liar’s Dividend is a tall order that won’t be accomplished overnight.