TECH FACTSHEETS FOR POLICYMAKERS

SPRING 2020 SERIES

Deepfakes

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs
TECHNOLOGY AND PUBLIC PURPOSE PROJECT

ASH CARTER, TAPP FACULTY DIRECTOR

LAURA MANLEY, TAPP EXECUTIVE DIRECTOR

CONTRIBUTORS

Raina Davis (Harvard)

Chris Wiggins (Columbia)

Joan Donovan (Harvard)

EDITOR

Amritha Jayanti (Harvard)

The Technology Factsheet Series was designed to provide a brief overview of each technology and related policy considerations. These papers are not meant to be exhaustive.

Technology and Public Purpose Project

Belfer Center for Science and International Affairs

Harvard Kennedy School

79 John F. Kennedy Street, Cambridge, MA 02138

www.belfercenter.org/TAPP

Design and layout by Andrew Facini

Executive Summary

Deepfakes can be defined as synthetic auditory or visual media developed using deep learning, a subfield of machine learning (ML), that appear to be authentic and are often created with the intent of deceiving audiences. Synthetically generated media widely varies in technical sophistication and application, ranging from low quality "cheap fakes" to more high quality "deepfakes," and has the ability to challenge and influence perceptions of reality. The development of synthetic audio-visual content is not novel: Hollywood filmmakers have employed computer-generated imagery (CGI) since the 1970s to *temporarily*—for the duration of the film—suspend disbelief among audiences. Advances in ML have made sophisticated synthetic media cheaper and easier to produce (particularly thanks to a proliferation of free and open source software for generating deepfakes. Even technologically unsophisticated actors are now able to create and distribute deepfakes.

Deepfakes have been used to spread disinformation and misinformation about public officials and political issues, non-consensually alter pornographic content, and develop amateur entertainment on social media applications. While experts disagree over the challenges that deep fakes pose for society, there is growing general concern over how deepfakes contribute to the evolution of the internet disinformation age. Certain experts have even sounded the alarm, declaring deepfakes a "looming crisis in national security, democracy, and privacy."[1]

The first United States federal legislation on deepfakes was signed into law in December 2019 as part of the National Defense Authorization Act (NDAA), requiring a comprehensive report of foreign weaponization of deepfakes, among other mandates.[2] Additionally, there are several proposed legislations regarding deepfakes pending on the floor of the House of Representative and the Senate. Virginia, Texas, California, and New York have also targeted deepfakes with recent legislation.[3] Given the existing and anticipated concerns regarding deepfakes, there is a clear need for U.S. legislators and policymakers to continue to deepen engagement with the privacy, safety, and security risks that exist for this technology.

1 Chesney, Robert, and Danielle Citron. "Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?" Lawfare, January 8, 2020. https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy.

2 Chipmon, Jason, and Matthew Ferraro. "First Federal Legislation on Deepfakes Signed Into Law." JD Supra, December 24, 2019. https://www.jdsupra.com/legalnews/first-federal-legislation-on-deepfakes-42346/.

3 Ruiz, David. "Deepfakes Laws and Proposals Flood US." Malwarebytes Labs, January 24, 2020. https://blog.malwarebytes.com/artificial-intelligence/2020/01/deepfakes-laws-and-proposals-flood-us/.

What is a Deepfake?

The term "deepfake"—a hybrid of the terms "deep learning" and "fake"—first appeared on Reddit in 2017 and quickly became part of the technical lexicon, regularly appearing in news and magazine articles by 2018. Deepfakes came to be defined as auditory or visual media that have been manipulated or developed using deep learning, and appear to be authentic.

There are many relevant and related terms in popular media, including cheap fakes, shallow fakes, and synthetic media, that have important distinctions from deepfakes. **Cheap fakes**—also known as **shallow fakes**—are audiovisual (AV) manipulations created with cheaper, more accessible software (or none at all) as compared to deepfakes.[4] Cheap fakes can be rendered through Photoshop, lookalikes, re-contextualizing footage, speeding, or slowing.[5] These videos do not rely on machine learning, but simple techniques available on any video recording or editing software.[6] **Synthetic media** is a catch-all term for media manipulated by software, regardless of use, placement, or intent. Table 1 below demonstrates the spectrum between cheap or shallow fakes and deepfakes.

Table 1: The Spectrum of Cheap/Shallow fakes and Deepfakes[7]

Classification	Manipulation Technique
Cheap or Shallow fakes  Deepfakes	**Re-contextualizing**: Introducing an existing video/recording under a false pretense or showing an edited clip out of context.
	Lookalikes: Hiring look-alike actors to impersonate target individuals.
	Speeding & slowing: Altering the speed of a video to change the meaning or perception.
	Face swapping - Rotoscope: Swapping facial clips and rotoscoping to composite the two layers into one and create the face-swapping effect.
	Lip-synching: Mapping voice recording from one or multiple contexts to a video recording in another, to make the subject of the video appear to say something authentic.
	Face replacement[8]: Digitally mapping a person's face onto another's), also known as *face-swapping* and *deepfake puppetry*.
	Synthetic speech production: Mimicking the voice of a real person without source content.
	Face reenactment[9] & Audio Modulation: Manipulating the features of a person's face or voice based on pre-defined characteristics.
	Face generation[10]: Generating original images and overlaying them on a person's face using advanced software (no source image needed).

4 Paris, Britt, and Joan Donovan. "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence." Rep. *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence*. Data & Society, 2019.

5 Ibid.

6 Johnson, Bobbie. "Deepfakes Are Solvable-but Don't Forget That 'Shallowfakes' Are Already Pervasive." MIT Technology Review. MIT Technology Review, April 2, 2020. https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/.

7 Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence, 2019.

8 Centre for Data Ethics and Innovation, *CDEI Snapshots*: *Deepfakes and Audio-visual Disinformation* (London, UK: Centre for Data Ethics and Innovation, September 2019) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf.

9 Ibid.

10 Ibid.

Deepfakes can vary in technical sophistication and application. A **Perfect Deepfake**, which is highly sophisticated, will be void of any defects and indistinguishable from real footage by any expert or algorithm. It has yet to be achieved, but some experts predict the technology will be perfected by the end of 2020.[11] Since there is no way to verify through detection techniques that a perfect deepfake has actually been created though, acknowledging the breakthrough relies on the trustworthiness and transparency of the developer.

Technical Overview

Deepfakes rely on **deep learning (DL)**, a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called **artificial neural networks (ANN)**. ANN are capable of learning unsupervised from data that is unstructured or unlabeled, which is advantageous for creation of deepfakes since a significant amount of original AV data pulled from the internet is unstructured or unlabeled. DL capitalizes on very large data sets in order to train networks with high performance – something that has become possible thanks to the generation and capture of enormous amounts of data throughout the internet.

The most popular method for creating deepfakes uses **generative adversarial networks (GANs), first published in 2014.**[12] A GAN is a pair of adversarial ANNs: one network, called the generator, inputs a latent sample and generates an image. This output is then put into a second network, called the discriminator, which has been trained on real data to classify an image as real or fake. The discriminator scores the forgery on a scale from zero to one - one being a high probability of the image being real; zero being a high probability that this image is fake. The generator uses this scoring and feedback to adjust its weights until the discriminator can no longer tell the difference between the forgery (output of the generator) and a real image.[13] The larger the set of training data for the classifier, the more believable the forgery will likely be.

Another, though less common, method of deepfake creation relies on **variational autoencoders (VAEs).**[14] VAEs are generative models that rely on two different networks working together, unlike adversarial networks. The encoder network produces a smaller, dense representation of the input data and the decoder takes this output and attempts to reproduce the original data. These networks are trained as a whole on a single dataset, for example, hundreds of images of a celebrity, until the input and output roughly match. The decoder can then be adjusted to create the desired effect, such as adding glasses to a specific target from the original AV media. A deepfake "face-swap," for example mapping a user's face onto a celebrity's body, can be generated by combining two VAEs. The user's image is encoded by means

11 Mosley, Tonya. "Perfect Deepfake Tech Could Arrive Sooner Than Expected." Perfect Deepfake Tech Could Arrive Sooner Than Expected | Here & Now. WBUR, October 2, 2019. https://www.wbur.org/hereandnow/2019/10/02/deepfake-technology.

12 Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in neural information processing systems, pp. 2672-2680. 2014.

13 Naoki Shibuya, "Understanding Generative Adversarial Networks," *Medium*, November 3, 2017, https://medium.com/activating-robotic-minds/understanding-generative-adversarial-networks-4dafc963f2ef.

14 "Artificial Intelligence: GANs and Autoencoders Applied to CyberSecurity." Artificial Intelligence: GANs and Autoencoders applied to CyberSecurity. Eleven Paths, May 3, 2019. https://www.elevenpaths.com/wp-content/uploads/2019/06/whitepaper-artificial-intelligence-gans-and-autoencoders-applied-to-cybersecurity.pdf.

of the user encoder but then decoded using the celebrity decoder, creating a convincing recreation of the original video but with a new face attached to the target actor.

In comparison to video, audio manipulation is less sophisticated and has been the "weakest link"[15] in the development of a perfect deepfake. Current techniques include modulation and synthesis. Modulation uses GANs to copy, model, and transform a voice to fit certain characteristics: e.g., young, high-pitched and female, British aristocrat, southern deep male, etc.[16]

Detection & Mitigation

Similar techniques for creating deep fakes can be used to identify them. Detection algorithms pick up subtle imperfections in the altered content invisible to the naked eye. For audio deep fakes, spectrograms of audio clips that sound similar are visually distinct, enabling detection with even rudimentary technology. Cutting edge detection techniques identify patterns of sound and inflection points in the audio recording that humans do not normally or cannot physically make.[17]

However, detection of deepfakes can be extremely difficult and time intensive, and technical limitations have kept the burden for detection on end-point content consumers. Detecting deepfakes at a consumer-level can be tricky, though, especially when a majority of content today is consumed on mobile devices. Common technical glitches often mask the sub-par sophistication of manipulated media. Viewers dismiss glitches in low-quality videos as a streaming error or poor internet/cellular connection. Further, low quality videos appear high-quality on a small screen and can easily fool the mobile-viewer.

Government agencies and private companies alike have adopted programs to begin to overcome these challenges in the interest of the public good. For example, U.S. Defense Department's **Defense Advanced Research Project Agency (DARPA)** backs a large research project into media forensics, MediFor. It also launched its own Semantic Forensics program, or SemaFor, where researchers aim to help computers use common sense logical reasoning to detect manipulated media.[18] Technology companies such as Facebook and Google have launched in-house and crowd-sourced initiatives to better identify deepfakes circulating on their platforms.[19]

15 The Weekly, "Deepfakes – Believe a Your Own Risk," *The New York Times*, 26:38, Nov. 22, 2019, https://www.nytimes.com/2019/11/22/the-weekly/deepfake-joe-rogan.html.

16 Knight, Will. "This AI Lets You Deepfake Your Voice to Speak like Barack Obama." MIT Technology Review. MIT Technology Review, April 2, 2020. https://www.technologyreview.com/2019/02/27/66005/this-ai-lets-you-deepfake-your-voice-to-speak-like-barack-obama/.

17 Waddell, Kaveh. "Researchers Are Figuring out How to Detect Audio Deepfakes before It's Too Late." Axios, April 3, 2019. https://www.axios.com/deepfake-audio-ai-impersonators-f736a8fc-162e-47f0-a582-e5eb8b8262ff.html.

18 Turek, Matt. "Defense Advanced Research Projects Agency." Defense Advanced Research Projects Agency. Accessed 2020. https://www.darpa.mil/program/semantic-forensics.

19 Mccabe, David, and Davey Alba. "Facebook Says It Will Ban 'Deepfakes'." The New York Times. The New York Times, January 7, 2020. https://www.nytimes.com/2020/01/07/technology/facebook-says-it-will-ban-deepfakes.html.

Applications and Market Development

Deepfakes gained popularity through satirical videos of political figures, celebrities rapping, and popular social media applications that allowed the user to "swap" faces with a friend. Once viewed as an innocuous tool, deepfakes have since generated widespread condemnation for their use in celebrity sex-tapes, revenge porn, incriminating videos of politicians, and financial fraud.[20] Leading up to the 2020 presidential election in the United States, policymakers and journalists have turned their attention to this technology as the latest evolution of disinformation and express concern for its corrosive effects on U.S. democracy.[21]

Like all technologies though, the potential for malicious use lies in the application of the tool. In this way, deepfakes present both opportunities and risks through positive and negative applications. Common use cases, which illuminate both types of application, include:

- **Political and Social Satire:** Deepfakes have the ability to enhance entertainment media by allowing for the low-cost production of high-end political and social satire content by capitalizing on authentic-seeming AV. Satirical content produced with deepfake technology is not meant to spread misinformation or disinformation, but to elevate the message behind the satire. With pressure to remove deepfake content on social media platforms, companies such as Facebook have attempted to make distinctions between deepfakes for satire and deepfakes for mis- and disinformation (emphasizing that intent matters).[22]

- **Advertising Visuals:** Companies, such as Rosebud AI, are applying deepfake technology to empower businesses to create high-quality, low-cost advertising visuals and assets.[23] Rosebud AI claims to be "democratizing the creation of visuals, so that brands big and small can tell their story compellingly." The application of deepfakes to allow companies to make more on-brand and cost-effective marketing campaigns could present shifts for traditional marketing and modeling agencies.

- **Voice Reconstruction:** Audio construction with deepfakes has presented opportunities for the construction and reconstruction of voices, especially for those who have lost the ability to speak due to health conditions, such as amyotrophic lateral sclerosis (ALS). Project Revoice is an initiative that aims to "fully recreate the unique essence of any voice and build a complete digital voice clone for everyday use with Augmented/Alternative Communication (AAC) devices."[24] This project targets those with ALS who have lost their speech abilities, but demonstrates the possibility for wider applications.

- **Political Disinformation:** Deepfakes have been used to intentionally spread disinformation and misinformation about public officials and divisive political issues. Distinct from political satire, these videos deliberately

20 Chen, Angela. "Three Threats Posed by Deepfakes That Technology Won't Solve." MIT Technology Review. MIT Technology Review, April 2, 2020. https://www.technologyreview.com/2019/10/02/75400/deepfake-technology-detection-disinformation-harassment-revenge-porn-law/.

21 Ibid.

22 Bloomberg. "Facebook Bans 'Deepfake' Videos, but Policy Doesn't Cover Parody or Satire." Los Angeles Times, January 7, 2020. https://www.latimes.com/business/technology/story/2020-01-07/facebook-deepfakes-policy.

23 "Rosebud AI." Rosebud AI. Accessed 2020. https://www.rosebud.ai/.

24 "We Can Stop ALS from Stealing More Voices." Home - Project Revoice. Accessed 2020. https://www.projectrevoice.org/.

aim to influence public opinion through false stories. A well circulated example of synthetic media targeting politicians includes a doctored video from 2018 of Barack Obama badmouthing Donald Trump.[25]

- **Non-Consensual Pornography:** There is growing concern about the use of deepfakes to non-consensually alter pornography, especially in its implications for targeted individuals. Women and minorities have been disproportionately targeted by this "image-based sexual abuse."[26] A 2019 report by DeepTrace found that 96 percent of the estimated 15,000 deepfakes circulating online are pornographic, and 99 percent of those pornographic deepfakes mapped faces from female celebrities onto pornographic actresses.[27]

- **Blackmail and Extortion:** For blackmail and extortion, the sophistication of the media becomes secondary to the potential consequence of its release. Often, extortion overlaps with non-consensual pornography. Past victims include political leaders and detractors: the imprisoned Philippines Senator Leila De Lima in 2016; an Indian journalist critical of the nationalist Bharatiya Janata Party (BJP) in 2018;[28] and Malaysian cabinets minister in 2019.[29]

- **Amateur Entertainment:** Deepfakes have become a popular feature on social applications and video-sharing services. Snapchat filters and dedicated applications like "FaceSwap" mimic the animation technique known as rotoscoping to map objects (like glasses or bunny ears) or another person's face onto the user.[30] Face-swaps and other deepfakes of celebrities and politicians have reached millions of viewers on video-sharing websites like YouTube and Vimeo. A YouTube channel TheFakening uses "AI to make goofy memes to highlight, educate and entertain with deepfakes. All this is worth it if you're laughing."[31]

- **Financial Fraud and Cybercrime**: Synthetic audio is a growing concern in financial fraud and white-collar cybercrime. In August 2019, an audio deepfake of a European company's chief executive resulted in the fraudulent transfer of $243,000.[32] ML-assisted financial scams rose significantly in 2019 as speech synthesis technology improved. If applied at scale or used to imitate the voice of a military leader or top US official, this technique could have serious national security implications.

- **Professional Achievement:** In addition to videos posted for entertainment and satire, video-sharing services feature channels by professionals to post their latest achievement and advertise services. For example, Shamook boasts "best worst deepfakes on the internet" and offers "commissions and collaborations."[33] A number of startups and tech companies actively compete to produce the most authentic deepfakes, not only for profit but also for the feat of "being the first" to engineer a perfect deepfake.[34]

---

25 Vincent, James. "Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA about Fake News." The Verge. The Verge, April 17, 2018. https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed.

26 McGlynn, C., Rackley, E. & Houghton, R. Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. *Fem Leg Stud* 25, 25–46 (2017).

27 Patrini, Giorgio. "Mapping the Deepfake Landscape." Deeptrace, October 20, 2019. https://deeptracelabs.com/mapping-the-deepfake-landscape/.

28 Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence, 2019.

29 Golingai, Philip. "Is It Azmin or a Deepfake?" The Star Online, December 31, 1969. https://www.thestar.com.my/opinion/columnists/one-mans-meat/2019/06/15/is-it-azmin-or-a-deepfake.

30 O'Kane, Sean. "Snapchat Now Lets You Face Swap with Pictures from Your Camera Roll." The Verge. The Verge, April 22, 2016. https://www.theverge.com/2016/4/22/11486630/snapchat-update-free-replays-face-swap-photos.

31 *YouTube*. TheFakening. Accessed 2020. https://www.youtube.com/channel/UC5D-8hVVwLB0DNrcSBqoVxg/about.

32 Statt, Nick. "Thieves Are Now Using AI Deepfakes to Trick Companies into Sending Them Money." The Verge, September 5, 2019. https://www.theverge.com/2019/9/5/20851248/deepfakes-ai-fake-audio-phone-calls-thieves-trick-companies-stealing-money.

33 *YouTube*. Shamook. Accessed 2020. https://www.youtube.com/channel/UCZXbWcv7fSZFTAZV4beckyw/about.

34 The Weekly, "Deepfakes – Believe a Your Own Risk," *The New York Times*, 26:38, Nov. 22, 2019, https://www.nytimes.com/2019/11/22/the-weekly/deepfake-joe-rogan.html.

Additionally, there are various anticipated applications to the gaming industry, such as mapping players' faces onto characters in video games to create immersive engagement; and to the movie industry, such as AV lip-syncing to make high-quality and cost-effective translations of movies into any language.[35]

Market Landscape

There are many organizations, from startups to larger firms, commercializing deepfakes, such as **Faceswap**, **Topaz Labs**, **Rosebud AI**, **Modulate**, **Meo, Project Revoice,** and the Chinese company, **Zao**. Beyond companies looking to monetize deepfake technology, there are many open source toolkits by individuals and organizations, such as **Avatarity** and **OpenFaceSwap**. The code for these tools is available on Github for anyone to use.[36]

Current Governance and Regulation

Existing regulation in the U.S. for deepfakes has been concentrated at the state-level. There is currently only one federal statute prohibiting the creation or distribution of deepfakes, though there are at least four bills in both the U.S. House of Representatives and the Senate. Internationally, there is limited government regulation targeted specifically at deepfakes, though has been increased interest in exploring regulatory options.

U.S. Federal Regulation

- **National Defense Authorization Act (NDAA) for Fiscal Year 2020**: This law (1) requires a comprehensive report of foreign weaponization of deepfakes; (2) requires the government to notify Congress of foreign deepfake-disinformation activities targeting U.S. elections; and (3) establishes a "Deepfakes Prize" competition to encourage the research or commercialization of deepfake-detection technologies.[37]

Pending Legislation

- **The Defending Each and Every Person from False Appearances by Keeping Exploitation Subject (DEEPFAKES) to Accountability Act:** This bill mandates that all deepfakes are accompanied by a textual disclosure explaining that the video

35 Lee, Linda W, Jan Kietzmann, and Tim C Kietzmann. "Deepfakes: Five Ways in Which They Are Brilliant Business Opportunities." The Conversation, May 1, 2020. https://theconversation.com/deepfakes-five-ways-in-which-they-are-brilliant-business-opportunities-131591.

36 Cole, Samantha. "This Open-Source Program Deepfakes You During Zoom Meetings, in Real Time." Vice, April 16, 2020. https://www.vice.com/en_us/article/g5xagy/this-open-source-program-deepfakes-you-during-zoom-meetings-in-real-time.

37 Bill (2020). https://www.congress.gov/bill/116th-congress/senate-bill/1790

was modified and irremovable digital watermarks; criminalizes the creation of synthetic media imitating a person that is not identified as such; and establishes right of the victim to sue the creator and "vindicate their reputations" in court.[38]

- **The Identifying Outputs of Generative Adversarial Networks (IOGAN) Act:** The IOGAN bill directs the National Science Foundation (NSF) to support research on manipulated or synthesized content and information authenticity, and the National Institute of Standards (NIST) to support research for the development of measurements and standards necessary to accelerate the development of the technological tools to examine the function and outputs of GANs or other technologies that synthesize or manipulate content.[39]

- **The Deepfake Report Act of 2019:** This bill requires the Science and Technology Directorate in the Department of Homeland Security to report at specified intervals on the state of digital content forgery technology.[40]

- **S. 1348 (2019):** This bill requires the Secretary of Defense to conduct a study on cyberexploitation of members of the Armed Forces and their families.[41]

Additionally, though not specific to deepfakes, the **Section 230 of the Communications Decency Act (CDA)** protects social media platforms from liability for content published by their users,[42] but several companies have adapted terms of service agreements to combat the growing rise of deepfakes on their platforms. In 2019, Facebook introduced a ban on "AI-developed deepfakes that can reasonably fool a user."[43] It is worth noting that Facebook made an exception to this rule for satire pieces, and it is now up to the company to dedicate and discriminate accordingly.

U.S. State Regulation

- **California AB 730**: This law makes it "illegal to create or distribute videos, images, or audio of politicians doctored to resemble real footage within 60 days of an election."[44]

- **Virginia 18.2-386.2**: This law amends current criminal law on revenge porn, making it a crime to, for example, insert a woman's digital presence into a pornographic video without her consent.[45]

38 Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019. Bill (2019). https://www.congress.gov/bill/116th-congress/house-bill/3230

39 Identifying Outputs of Generative Adversarial Networks Act. Bill (2019). https://www.congress.gov/bill/116th-congress/house-bill/4355

40 Deepfake Report Act of 2019. Bill (2019). https://www.congress.gov/bill/116th-congress/senate-bill/2065

41 A bill to require the Secretary of Defense to conduct a study on cyberexploitation of members of the Armed Forces and their families, and for other purposes. Bill (2019). https://www.congress.gov/bill/116th-congress/senate-bill/1348

42 "47 U.S. Code § 230 - Protection for Private Blocking and Screening of Offensive Material." Legal Information Institute. Legal Information Institute, n.d. https://www.law.cornell.edu/uscode/text/47/230.

43 Edelman, Gilad. "Facebook's Deepfake Ban Is a Solution to a Distant Problem." Wired. Conde Nast, January 7, 2020. https://www.wired.com/story/facebook-deepfake-ban-disinformation/.

44 AB-730 Elections: deceptive audio or visual media., AB-730 Elections: deceptive audio or visual media. § (2019).

45 "Deepfakes Laws and Proposals Flood US," (2020).

- **Texas SB 751**: This law prohibits the use of deepfakes for election interference, such as making a video that inaccurately shows a political candidate at a Neo-Nazi rally the month leading up to an election.[46]

In early 2020, Massachusetts, New York, and Maryland introduced their own versions of these bills and the legislation is currently pending.[47]

International Regulation

- **Cyberspace Administration of China (CAC) deepfakes policy**: In effect as of January 1, 2020, this CAC regulation requires publishers of deepfake content to disclose that a piece of content is in fact a deepfakes. It also requires content providers to detect deepfake content themselves.

- Though international regulation of deepfakes directly is limited, countries and international bodies, such as India and the European Union, are currently exploring ways to combat deepfakes through targeted legislation.

Public Purpose Considerations

There are many public purpose considerations around privacy, safety, security, and trust that arise from the widescale use of deepfake technology. Some of these considerations include:

- **Evidence and Truth:** Deepfakes, like other historical media technologies, have shifted the way we think about evidence and truth. When the integrity and authenticity of AV content—something that often has seemed to be the arbiter of truth—comes into question, and on a wide scale, it becomes hard to understand what is believable.[48] Though this issue is not necessarily novel to deepfakes, the introduction of low-barrier technology to spread synthetic media presents added challenges. It is important to consider how trusted institutions can play a role in redefining the lines of evidence and truth in our society, so that there is a standard that we can employ for use cases as important as our justice systems.

- **Liar's Dividend:** The rise of media fact-checking to address mis- and disinformation produced an unintended consequence, known as the liar's dividend. The term describes a phenomenon in which debunking fake or manipulated material not only gives the content a longer lifetime but actually "legitimize[s] the debate over the veracity."[49] Detractors can dismiss the authentic as false and the public skeptical of any reports related to that topic.[50] When thinking about mitigation strategies for deepfakes, it is important to consider how liar's dividend may play a role in the effectiveness of those strategies.

46 Ibid.

47 Ibid.

48 Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence, 2019.

49 Benz, Kevin. "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/.

50 Chadwick, Paul. "The Liar's Dividend, and Other Challenges of Deep-Fake News | Paul Chadwick." The Guardian. Guardian News and Media, July 22, 2018. https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin.

- **Denial of Authenticity:** Related to lines of evidence and truth, and liar's dividend is the ability for deepfakes to seed enough public doubt in AV content writ large that people can claim real content is a deepfake. In late 2016, Donald Trump publicly acknowledged his infamous "Access Hollywood" video was authentic when it was first released,[51] but shortly after his presidential victory a few months later, he shifted his statement.[52] He claimed that there was audio manipulation, and that he never said what the video depicted him saying. Due to the high-quality alterations offered by deepfake technology, claims that deny authenticity have the ability to generate effective public doubt.

- **Consumer Protection & Privacy**: The popularity of deepfake social media filters and applications raises privacy concerns over how companies protect, and in some cases, exploit user data. The Viral Chinese Zao face-swapping app allowed users to star in their favorite movies by uploading personal photos. Its terms of service granted the company "perpetual and transferable rights to uploaded data," raising major privacy concerns and prompting WeChat from banning Zao-generated content on its platform.[53] A second area of privacy debate has emerged over the use of public images and social media photos in non-consensual pornography. Privacy advocates argue that an image shared publicly becomes private when manipulated "because the person has lost any agency and freedom to choose the portrayal of their sexual self."[54] The training datasets that drive ML algorithms deserve more scrutiny. Policy makers and technologists alike must consider the ethical and legal boundaries for audio-visual manipulation.

- **Digital Security:** More and more digital services are relying on biometric data, such as facial recognition and voice recognition, as security barriers. With the introduction of highly sophisticated face-swap technology and voice alteration techniques, the effectiveness of these security barriers could be compromised.[55] Policymakers and companies should consider how deepfakes could impact the integrity of digital security infrastructure.

- **Delayed Action of Content Detection:** Related to liar's dividend is the idea that detection will always trail behind the production and distribution of deepfakes. This means that by the time content is removed, it has already been consumed by some portion of the population. Policymakers and companies must consider how to be proactive in mitigation strategies to minimize the possible harm caused by even limited circulation of synthetic content. This is especially true since copies of internet content are easily stored and reuploaded for distribution. Managing multiple sources of distribution can then become difficult.

- **Role of Media:** The media can play an important role in how the public perceives deepfakes. A specialist in disinformation and counter-messaging, emphasized that "No-one likes to be fooled." If you get the message out early and frame it narrowly, audiences will dismiss specific incidents of deepfakes without losing trust in digital media altogether. The current media hype around deepfake videos is eroding this trust, even though most deepfakes will not fool the average viewer. As gener-

51 Burns, Alexander, Maggie Haberman, and Jonathan Martin. "Donald Trump Apology Caps Day of Outrage Over Lewd Tape." The New York Times. The New York Times, October 7, 2016. https://www.nytimes.com/2016/10/08/us/politics/donald-trump-women.html?_r=0.

52 Haberman, Maggie, and Jonathan Martin. "Trump Once Said the 'Access Hollywood' Tape Was Real. Now He's Not Sure." The New York Times. The New York Times, November 29, 2017. https://www.nytimes.com/2017/11/28/us/politics/trump-access-hollywood-tape.html.

53 Grace Shao and Evelyn Chen, "The Chinese face-swapping app that went viral is taking the danger of 'deepfake' to the masses," *CNBC*, September 4, 2019, https://www.cnbc.com/2019/09/04/chinese-face-swapping-app-zao-takes-dangers-of-deepfake-to-the-masses.html.

54 Clare McGlynn, Erika Rackley & Ruth Houghton, "Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse," *Fem Leg Stud* 25 (2017): 25–46.

55 Coleman, Alistair. "'Deepfake' App Causes Fraud and Privacy Fears in China." BBC News. BBC, September 4, 2019. https://www.bbc.com/news/technology-49570418.

ation techniques become more sophisticated, the media will have an important role to play in authenticating true footage and exposing forgeries. Explaining the reasoning behind these decisions in a clear and accessible manner will be critical.

- **Allocation of Resources:** While the most advanced deepfakes are indistinguishable from real content to the average audience, forensic experts at the FBI and other branches of the US justice system have the tools to discern forgeries. For audio files, experts can turn to spectrograms, visual representations of the sound waves, which can pick up on different inflections and changes in volume too subtle to hear. For digital images and videos, experts turn to the RAW file, similar to the negatives of traditional film, to identify footprints left behind in post-processing (alteration). But the most sophisticated detection tools require significant time and resources, that may not be worth the effort. Policy makers and practitioners must consider guidelines for when and where such resources should be allocated.

Appendix: Key Questions for Policymakers

Trust & Truth

- How is the role of establishing and reassuring users of truthful content shared between the private and public sector? What is the delineation of roles?
- What novel risks, if any, do deepfakes pose to the public's use of evidence and truth within our formal and informal justice and legal systems?
- How do deepfakes fit into the large landscape of misinformation and disinformation with the internet era? What does this mean for the way we approach the delineation and governance of evidence and truth overall?

Consumer Protection & Privacy

- What techniques can be used to signal deepfake content to consumers?
- How do we reconsider user data rights, especially data that exists openly online? Should there be measures to limit the use of open personal data for the generation of deepfakes writ large? What cascading effects could this cause?
- What measures should be taken to ensure that digital security infrastructure at an individual level, such as facial recognition and voice recognition technology, maintains integrity?

National Security & Democracy

- How should policymakers consider improving deepfake regulation while upholding first amendment rights?
- How can deepfake technology comprise national security and democratic foundations, especially as they relate to local and federal elections?
- In efforts to detect and mitigate threats from deepfakes, how important is attribution? How can the concept of attribution, especially at either an individual or national level, be integrated into mitigation techniques?