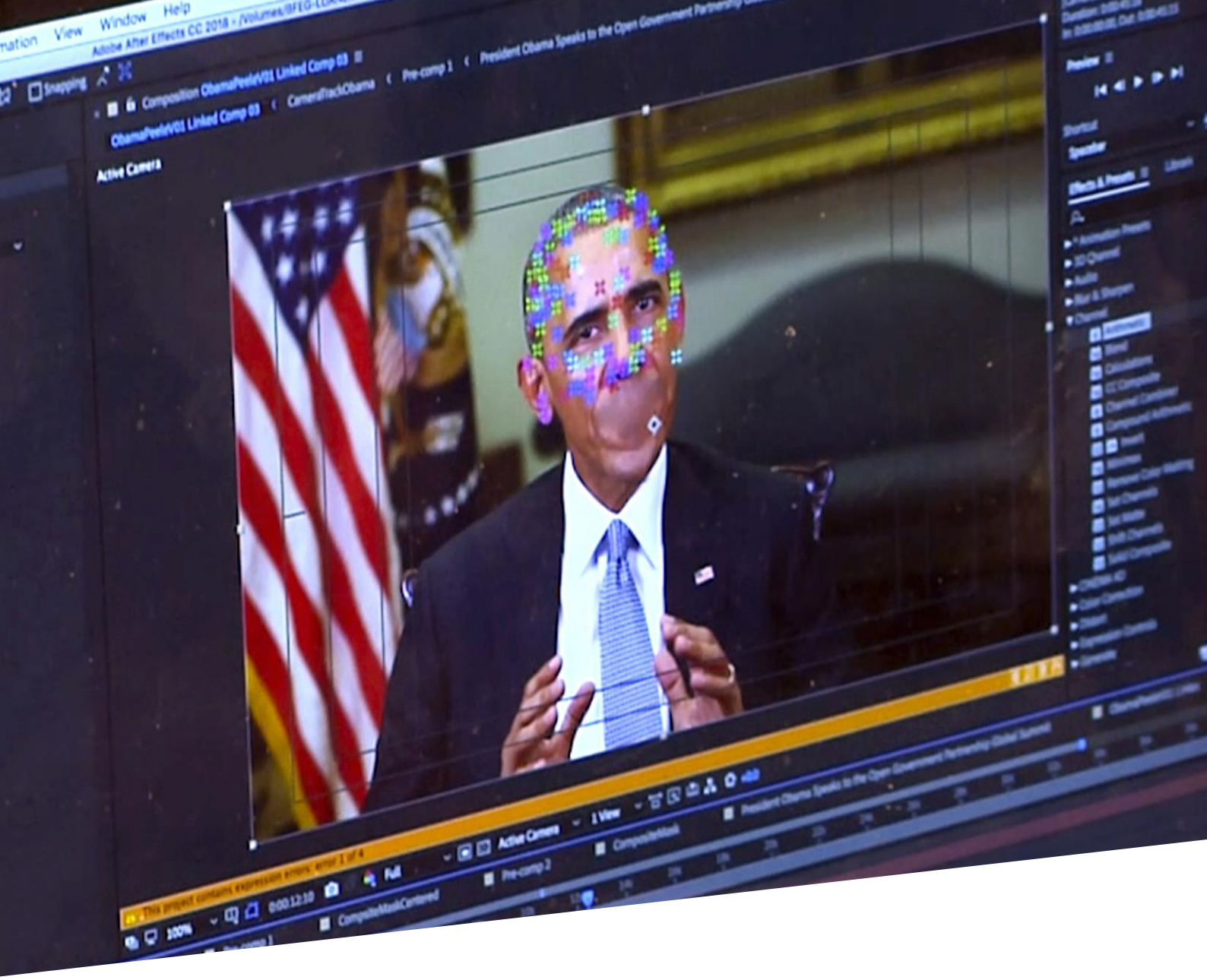

ANALYSIS

SEEING IS NO LONGER BELIEVING

DEEPFAKES, CHEAPFAKES AND THE LIMITS OF DECEPTION

| EMILIA ANNA PORUBCIN |

DECEMBER 2019

RKK ICDS
RAHVUSVAHELINE KAITSEUURINGUTE KESKUS
INTERNATIONAL CENTRE FOR DEFENCE AND SECURITY
EESTI • ESTONIA

Title: Seeing Is No Longer Believing: Deepfakes, Cheapfakes and the Limits of Deception
Author(s): Porubcin, Emilia Anna
Publication date: December 2019
Category: Analysis

Keywords: disinformation, deepfakes, technology, technology governance, social media, freedom of speech, regulation

Cover page photo: This image made from video of a fake video featuring former U.S. President Barack Obama shows elements of facial mapping used in new technology that lets anyone make videos of real people appearing to say things they've never said. There is rising concern that U.S. adversaries will use new technology to make authentic-looking videos to influence political campaigns or jeopardize national security. (AP Photo/Scanpix)

ISSN 2228-2076

63/4 Narva Rd., 10152 Tallinn, Estonia
info@icds.ee, www.icds.ee

Introduction

If 2017 was the year of fake news, summer 2019 was the summer of deepfakes. Following the viral spread of doctored videos earlier in the year, the US House Intelligence Committee held an open hearing on deepfakes and artificial intelligence (AI) in June.[1] Only weeks later, Virginia became the first American state to explicitly ban pornographic deepfakes. Speaking for the largest social media site in the world, Facebook CEO Mark Zuckerberg stated in June that his company would be considering how to manage the spread of deepfakes on its platform. And with the implications of Russian interference in the 2016 elections continuing to weigh heavily on the US national consciousness, the deepfake has been named the newest culprit of the end to truth and democracy.

If 2017 was the year of fake news, summer 2019 was the summer of deepfakes

Though the technology originated in academia and initially proliferated in niche online communities, deepfakes have become the concern of governments—and not just in America, but around the world. When policy and strategy decisions are informed by content that can be manipulated in manners that are increasingly tricky to detect, deepfakes put not only individuals but nations at risk. Efforts have been made in several countries to mitigate the negative effects of deepfakes, but we still have a long way to go before they are effectively managed at a global scale.

In an effort to better understand the individual and societal consequences of deepfakes, and ultimately respond to deepfakes in a more effective manner, this paper surveys the present deepfake environment: how the technology has evolved, the risks and benefits it poses, and how companies and governments have begun responding to the threat.

1. What They Are

Modified images, sounds, and videos, created using Photoshop, vocal effects processors, or simple film editing, have been around for decades. As yet another form of disinformation, deepfakes do not stand alone as a vehicle for falsehoods. What distinguishes deepfakes from other synthetic media is their combination of sophistication and accessibility. Deepfakes are videos manipulated with the help of AI, in the form of deep learning algorithms, to depict events that did not necessarily occur.[2] Deep learning models typically rely on artificial neural networks—nodal networks trained on massive datasets that allow them to build abstract representations of the patterns they recognize within input information.[3] The models can then generate output data, which becomes more accurate with more input data and larger models.

Deepfakes do not stand alone as a vehicle for falsehoods

1.1. Audio and Video Deepfakes

As the algorithms for creating deepfakes have improved to produce nearly undetectable facsimiles, out-of-the-box software used to create these fake videos has diffused online, allowing users with no programming background to generate videos of anyone of whom they have pictures. Before video deepfakes came audio deepfakes. In 2016,

[1] "House Intelligence Committee to Hold Open Hearing on Deepfakes and AI," *U.S. House of Representatives Permanent Select Committee on Intelligence*, 7 June 2019, accessed 26 June 2019.

[2] Deep learning is a subset of machine learning that uses several layers of algorithms to process and simplify data, identifying detectable patterns, or "features," in the data.

[3] "A Beginner's Guide to Neural Networks and Deep Learning," *A.I. Wiki*, accessed 1 August 2019.

Adobe and DeepMind began building commercial research products that used deep learning to generate speech from audio recordings.[4] Audio generation research continues today, with research as recent as Facebook's June paper on MelNet demonstrating the social and technical intrigue of synthesizing the unique human voice, but it was quickly followed by the emergence of video generation research.[5]

The first academic deepfake video, published in July 2017, used existing videos of then-American president Barack Obama to train a deep learning model that could make Obama "give" a speech entirely generated from a separate audio recording.[6] "AudioToObama" immediately demonstrated the political appeal of deepfakes. Public figures, particularly politicians who give dozens of recorded speeches daily, provide a great deal of training data that make it all the easier to create deepfakes of their persona. The subsequent diffusion of deepfake videos and their software from academia to the public is further described in Section II.

1.2. Text Deepfakes

In addition to the visual and audio components of deepfakes, now the very content they speak can be generated with deep learning. In February 2019, the research company OpenAI released its GPT-2 language model, which "generates synthetic text samples in response to the model being primed with an arbitrary input."[7] In other words, from only a few paragraphs of sample text, journalists can replicate Donald Trump's and Hilary Clinton's speeches, albeit with mixed results.[8] The multisensory reach of deepfakes across aural and visual mediums, already more convincing to a human audience than fake text alone, has now reached a complexity of range with text generation. Every element of synthetic media can be algorithmically generated, wholesale, from a relatively small amount of data and an even smaller amount of human input.

The release of GPT-2 could change the deepfake playing field with its introduction of algorithmic text generation. But it has already rocked the boat with the manner in which it was released. OpenAI open-sourced only a smaller version of the complete GPT-2 model and sampling code, choosing not to align with standard AI academic practice, as exemplified by Google's fully open-sourced language model BERT, by not releasing the entire model, training code, or full dataset for GPT-2.[9] The company's decision to attempt "responsible disclosure" was meant to minimize malicious applications, including the generation of fake news. But it nonetheless threatened both a nuanced discussion of AI's capabilities and the broader progress of AI research.[10] Journalists who selectively received access to the full GPT-2 model were able to proclaim, without the possibility for external fact-checking, the onset of "AI doom," heightening public fear without engaging in well-rounded efforts at public awareness.[11] Furthermore, OpenAI set a precedent for other companies to limit the openness of their AI research, potentially narrowing the field's development at large.

---

[4] James Vincent, "AI deepfakes are now as simple as typing whatever you want your subject to say," *The Verge*, 10 June 2019, accessed 26 June 2019.

[5] Sean Vasquez and Mike Lewis, "MelNet: A Generative Model for Audio in the Frequency Domain," arXiv:1906.01083, accessed 16 August 2019; Carolyn McGettigan and Nadine Lavan, "Human voices are unique – but our study shows we're not that good at recognizing them," *The Conversation*, 16 June 2017, accessed 16 August 2019.

[6] Supasorn Suwajanakorn et al., "Synthesizing Obama: Learning Lip Sync from Audio," *ACM Transactions on Graphics* 36, no. 4 (July 2017), 95:1-13.

[7] "Better Language Models and Their Implications," *OpenAI*, 14 February 2019, accessed 4 June 2019.

[8] Sean Gallagher, "Twenty minutes into the future with OpenAI's Deep Fake Text AI," *Ars Technica*, 27 February 2019, accessed 1 August 2019. As one example of the model's shortcomings: both politicians' generated speeches devolved into redundancy, with Trump's repeating "GOAT" and Clinton's repeating "And he has said things that are good for America."

[9] Sydney Li and Danny O'Brien, "OpenAI's Recent Announcement: What Went Wrong, and How It Could Be Better," *Electronic Frontier Foundation*, 4 March 2019, accessed 26 June 2019.

[10] Ibid.

[11] Hannah Jane Parkinson, "AI can write just like me. Brace for the robot apocalypse," *The Guardian*, 15 February 2019, accessed 1 August 2019.

2. HOW THEY HAVE BEEN HARMFUL[12]

Despite OpenAI's mixed-review efforts at "responsible disclosure," the general trend of artificial intelligence has been moving in the direction of an entirely open-sourced field of research. This has ensured the possibility of public awareness of deepfake technology, but it has also rapidly shifted the disinformation tool towards the hands of potentially malicious actors.

The general trend of artificial intelligence has been moving in the direction of an entirely open-sourced field of research. This has ensured the possibility of public awareness of deepfake technology, but it has also rapidly shifted the disinformation tool towards the hands of potentially malicious actors

2.1. Pornography and Harassment

The reasons for deepfakes' use have evolved dramatically during their ongoing transition from academia to the public. The open-sourcing of deepfake software made early deepfake generation models accessible to the tech-savvy Internet user, and immediately introduced the complexities of regulation.[13] Reddit user u/deepfakes is credited with publicly popularizing AI-powered video manipulation when they created pornographic videos and animations depicting adult actresses with celebrities' faces as early as December 2017.[14] The user implemented open-source software, including the Keras API and the TensorFlow machine learning library, to build a deep learning model that could be trained with publicly available porn videos and celebrity images to superimpose celebrities' faces into pornography.[15]

Deepfakes became a democratic tool for low-cost, high-sophistication video manipulation

With the emergence of downloadable applications that allowed users to create deepfakes even without programming experience, deepfakes became a democratic tool for low-cost, high-sophistication video manipulation. In January 2018, one such app, called FakeApp, gained particular popularity among subreddits where mostly explicit deepfakes were being shared.[16] The effects of these pornographic creations did not remain online. In April 2018, Indian journalist Rana Ayyub received rape and doxing threats over a deepfake circulated via WhatsApp by her political rivals, a bald attempt at silencing a notably liberal journalist through misogynistic means.[17] A new application known as DeepNude, which used AI expressly for removing the clothing from images of women, emerged in June 2019 and was almost immediately shut down; despite its developer's discontinuation of the product, the software continues to circulate (illegally, and perhaps fraudulently) online.[18] FakeApp, DeepNude, and their counterparts have formed the most important bridge for deepfakes to travel from academia to the public, allowing deepfakes to

[12] In their critical work on deepfakes, Chesney and Citron write about the possible "harmful uses" and "beneficial uses" of deepfake technology. The distinction is not always as black-and-white, but we follow the same division to analyze the real "harmful" and "beneficial" uses of deepfakes that we have witnessed over the past several years.

[13] Parkinson, "AI can write just like me."

[14] Samantha Cole, "AI-Assisted Fake Porn Is Here and We're All ******," *Vice*, 11 December 2017, accessed 26 June 2019.

[15] Ibid.

[16] Samantha Cole, "People Are Using AI to Create Fake Porn of Their Friends and Classmates," *Vice*, 26 January 2018, accessed 2 July 2019.

[17] Rana Ayyub, "In India, Journalists Face Slut-Shaming and Rape Threats," *The New York Times*, 22 May 2018, accessed 5 August 2019.

[18] Katyanna Quach, "DeepNude's makers tried to deep-six their pervy AI app," *The Register*, 2 July 2019, accessed 4 July 2019.

proliferate broadly and rapidly for the primary purpose of harassing women.

Recent criminal incidents have also demonstrated the use of AI-powered audio manipulation, or vocal deepfakes, for criminal impersonation. In March 2019, criminals used software to mimic the voice of the CEO of a parent company, who "contacted" the CEO of a UK-based energy firm to request an urgent transfer of €220,000.[19] Head researchers at the UN Interregional Crime and Justice Research Institute argue that the success of voice fraud incidents like this demonstrates how potent video deepfakes could be when used for the same purposes: "Imagine a video call with [a CEO's] voice, the facial expressions you're familiar with. Then you wouldn't have any doubts at all."[20]

The most prominent deepfakes, whether the videos themselves are satirical, educational, or simply entertaining, feature figures with immense political capital

2.2. Politics

Although voice fraud is on the rise and deepfakes are still most popularly used for pornography, one FakeApp creation released in April 2018 introduced another implementation of the deepfake. A Buzzfeed-commissioned deepfake depicted Barack Obama calling Donald Trump "a total and complete dipshit."[21] Although the video was explicitly a PSA warning viewers about fake news, it was widely upheld as an exemplar of how deepfakes could be used for political manipulation. Several similar deepfakes of Obama, Trump, and Vladimir Putin—videos with clearly non-malevolent intentions, but which have nonetheless been accused of propagating distrust—have exploited the entertainment value of celebrity to spread awareness about fake news, to a good deal of negative feedback.

The names listed above demonstrate that the most prominent deepfakes, whether the videos themselves are satirical, educational, or simply entertaining, feature figures with immense political capital. The threat of state actors actually leveraging political deepfakes to influence geopolitics has already manifested internationally. In May 2018, Belgian political party Socialistische Partij Anders (sp.a) created a deepfake they circulated on Facebook and Twitter, which portrayed Trump stating, "As you know, I had the balls to withdraw from the Paris climate agreement. And so should you."[22] Comments on the video, berating Trump and his politics, suggested that viewers did not recognize its inauthenticity, forcing sp.a to conduct damage control by informing their audience that the video was fake.[23] In May 2019, after a political aide in Malaysia "confessed" via video to being caught on film in a tryst with the Economic Affairs Minister, speculation abounded that his confession was deepfaked.[24] The EU-funded East StratCom Task Force has also noted Russian trolls experimenting with the use of deepfakes for disinformation.[25]

Deepfakes are still most popularly used for pornography

[19] Catherine Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case," *The Wall Street Journal*, 30 August 2019, accessed 31 August 2019.
[20] Ibid.
[21] James Vincent, "Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news," *The Verge*, 17 April 2018, accessed 23 July 2019.
[22] Hans von der Burchard, "Belgian socialist party circulates 'deep fake' Donald Trump video," *Politico*, 21 May 2018, accessed 27 June 2019.
[23] Oscar Schwartz, "You thought fake news was bad? Deep fakes are where truth goes to die," *The Guardian*, 12 November 2018, accessed 27 June 2019.
[24] Nic Ker, "Is the political aide viral sex video confession real or a Deepfake?" *Malaymail*, 12 June 2019, accessed 28 June 2019.
[25] Nick Harding, "Video nasties: Russia's faked broadcasts a new threat to West," *The Sunday Telegraph*, 27 May 2018, accessed 16 July 2019.

2.3. Heightening Distrust

Most deepfakes, including those from sp.a, are not entirely convincing. The voice is slightly too deep, or it is mistimed with a face's mouth; a figure blinks too little, producing an "uncanny valley" effect that many claim should make regulating deepfakes a moot point.[26] If the human eye can easily detect deepfakes, why does it matter how we punish them? Furthermore, these instances of deepfakes were created by actors with much greater spending power and social influence than the average netizen, suggesting that the amount of resources required to create convincing deepfakes is large. The problem is, as becomes clear after analyzing the comments left behind on "obvious" deepfakes, that people are fooled even by low-quality creations. In the words of one journalist: "they're convincing enough at a glance, and sometimes that's all that's needed."[27] As deepfake technology is "improving fast," with any "tell" for detecting a fake quickly fixed in the next iteration of a piece of software, and as the underlying social motivations for believing deepfakes persist, learning how to mitigate their damage will grow even more important. [28] With bots aiding a message's spread and the increasing democratization of high-tech capabilities—one can envision a black market for deepfakes or deepfake software akin to the zero-day exploit market facilitated by platforms like Zerodium—it could be only a matter of time before far less sophisticated actors can create deepfakes with similar influence. Their presence will only further muddy the media waters with fake news, driven by new and potentially unpredictable motivations, which will continue to subvert truth and trust alike.

[26] Russell Brandom, "Deepfake Propaganda Is Not a Real Problem," *The Verge*, 5 March 2019, accessed 26 June 2019; Jeffrey Westling, "Deep Fakes: Let's Not Go Off the Deep End," *TechDirt*, 30 January 2019, accessed 29 July 2019.

[27] James Vincent, "Deepfake detection algorithms will never be enough," *The Verge*, 27 June 2019, accessed 28 June 2019.

[28] "Rise of the deepfakes," *The Week*, 9 June 2018, accessed 4 August 2019;
Samantha Cole, "There Is No Tech Solution to Deepfakes," *Vice*, 14 August 2018, accessed 2 July 2019.

3. How They Have Been Beneficial

Despite the legitimate but often overpowering rhetoric of doom surrounding them, deepfakes have a variety of positive applications. The same technology used to non-consensually steal someone's facial likeness, voice, or behaviors can provide anonymity to young workers, reproduce deceased historical figures, or otherwise subvert standard practices of audiovisual self-representation. In theory, deepfakes could most greatly benefit (and even protect) the same populations that have been most harmed by deepfakes to date, and against whom technology in general and machine learning in particular are most likely to hold bias: minority groups by gender, race, disability, and religion, among other traits.

3.1. Awareness (Education)

Chesney & Citron classify three areas—education, art, and autonomy—through which deepfakes can benefit society.[29] Several of the same deepfakes critiqued for spreading disinformation can instead be classified under one or more of these categories. Educational applications of deepfakes have ranged from (highly topical) fake news awareness campaigns to efforts at preserving history. The previously mentioned deepfake of Obama insulting Trump was also used to heighten awareness of fake news, causing the former president to warn, "Moving forward, we need to be more vigilant with what we trust from the Internet."[30] The creator of a viral video superimposing the face of a Chinese actress onto that of another stated, "My initial intention was to help people hear about this technology and protect celebrities from videos of forged, negative content."[31] He seemed to have found success in doing so, as his video prompted discussion

[29] Bobby Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *U of Texas Law, Public Law Research Paper No. 692* (14 July 2018), accessed 26 June 2019, 14-16.

[30] "You Won't Believe What Obama Says In This Video!" YouTube video, 0:43, *Buzzfeed*, 17 April 2018, accessed 2 July 2019.

[31] Liang Chenyu, "Chinese 'Deepfake' Creator Says Videos Meant to Educate Public," *Sixth Tone*, 28 February 2019, accessed 6 August 2019.

online about face-swapping content, infringements of privacy, and the technology's particular potential for harassing women. Merging art with political activism has allowed the deepfake to become a tool for educating viewers about its own dangers.

3.2. Art (and Recreation)

Somewhere along the spectrum between education and art lie the work of artists Bill Posters and Daniel Howe, who created deepfakes of Kim Kardashian West and Mark Zuckerberg speaking about big data to advertise their exhibition at the Sheffield Doc Fest in June 2019. Posters and Howe's art installation, titled *Spectre*, "interrogates and reveals many of the common tactics that are used by corporate or political actors to influence people's behaviors and decision making."[32] Video clips of these deepfakes, depicting West, Zuckerberg, and other figures praising big data, quickly went viral after they were shared on platforms like Facebook and Instagram. They served simultaneously (and ironically) as an advertising ploy, a social message, and a test of Facebook's claim that they wouldn't remove a deepfake of the company's frontman.[33]

Merging art with political activism has allowed the deepfake to become a tool for educating viewers about its own dangers

At another museum across the Atlantic, a less sinister deepfake of Salvador Dalí was created to facilitate visitor engagement and experiment with a new form of the very surrealism that its subject pioneered. In May 2019, the Dalí Museum in St. Petersburg, Florida debuted an exhibition guided by a deepfaked recreation of Dalí, compiling thousands of frames from archival footage to train the algorithms that allowed the artist's face to be superimposed onto that of an actor with similar physical proportions.[34] Permission for using Dalí's likeness was granted by the Dalí Foundation in Spain (although technically this was unnecessary, as a 2016 ruling of Spain's Civil Court determined that the artist's image rights died with him in 1989).[35] In a similar instance of closing the gap between artist and art, Chinese search engine Sogou is currently developing technology that will allow authors' avatars to read their works aloud to readers, creating an audiovisual likeness that can allow deeper literary immersion.[36] Technology related to deepfakes was used for the CGI recreations of Peter Cushing and Carrie Fisher on display in 2016's *Rogue One: A Star Wars Story*, to highly polarized mixed reviews that debated the ethics, humanity, and effectiveness of "digital resurrection."[37] Finally, at a lower but much more prolifically met level of entry, Internet users regularly create deepfakes that swap or puppet celebrities' faces for entertainment; one particularly popular video, created with custom open-source software Faceswap, imposed Steve Buscemi's face on Jennifer Lawrence, and gained enough press to enter late-night coverage.[38] Chinese face swap app ZAO skyrocketed up download charts, gaining such immense popularity so rapidly that reigning social media site WeChat has started blocking

Internet users regularly create deepfakes that swap or puppet celebrities' faces for entertainment

[32] "Gallery: 'Spectre' Launches (Press Release)," *Bill Posters*, 29 May 2019, accessed 6 August 2019.
[33] Alex Boutilier, Twitter post, 28 May 2019, accessed 6 August 2019.
[34] Dami Lee, "Deepfake Salvador Dalí takes selfies with museum visitors," *The Verge*, 10 May 2019, accessed 1 July 2019.
[35] Naomi Rea, "Spanish Supreme Court Rules Against Dalí Foundation in Image Rights Dispute," *Artnet News*, 30 June 2016, accessed 7 August 2019.
[36] Jessie Gaynor, "Today in 'AI will replace us all.' Author avatars can now read their books to you," *Literary Hub*, 16 August 2019, accessed 16 August 2019.
[37] Alexi Sargeant, "The Undeath of Cinema," *The New Atlantis*, no. 53 (Summer/Fall 2017): 17-32.
[38] "Jennifer Lawrence-Buscemi on her favorite housewives [Deepfake]," YouTube video, *birbfakes*, 14 January 2019, accessed 2 July 2019.

ZAO videos on their platform.[39] Fun creations like Lawrence-Buscemi demonstrate the recreational potential of deepfakes, with new tools for tech-tinkering akin to Lego or Minecraft, but they also exemplify how even "fun" applications of the technology can (intentionally or otherwise) prompt meaningful discussions about deepfakes to enhance public awareness of their dangers.

Even "fun" applications of the technology can (intentionally or otherwise) prompt meaningful discussions about deepfakes

3.3. Autonomy

Autonomy is the category least represented in existing deepfake implementations, in part for lack of coverage. In their initial paper, Chesney and Citron focus briefly on the potential for self-expression through a deepfake "avatar," which could allow users to present themselves on otherwise impossible physical scenarios. This possibility has become reality, not for the sake of consensual sexual escapades as once predicted, but instead to preserve the voices of individuals with ALS, allowing them to continue speaking with their own voice even if they lose the ability to speak.[40] Project Revoice, conducted in partnership with the ALS Association, recreates voices for use in Augmented / Alternative Communication (AAC) devices. The same tools could theoretically be paired with deepfake video capabilities to allow an individual to "speak" as their own self, expanding the options available to limited-mobility individuals for external self-representation and more personal self-witnessing. Revoice exemplifies how deepfakes could be paired with existing technologies to improve quality of life for disabled persons. Augmented Reality (AR) is another example of technology that could pair well with deepfakes, for example by allowing persons with limited mobility to shop virtually, test out tools, or try on clothing without needing to navigate inaccessible spaces. Deepfake technology could also serve as a meaningful manner of minimizing discrimination online, such as through online interviews in a hiring process. Using real-time facial puppetry during an interview, a company could ensure a candidate's anonymity for the sake of an unprejudiced screening procedure. Projects enabling facial anonymization with deepfakes have already found success.[41]

Extensively hypothesizing every possible positive use for deepfakes lies beyond the scope of this paper.[42] But it is a task that deserves greater effort. Where regulation is meant to mitigate an immediate danger—revenge pornography, or mob violence—understanding and extrapolating the negative applications of deepfakes is an invaluable exercise in thorough policy development. As regulation increasingly becomes about smoothly incorporating and normalizing a new technology into society, however, policymakers, journalists, and scholars need to evaluate

Deepfake technology could also serve as a meaningful manner of minimizing discrimination online, such as through online interviews in a hiring process

[39] Jake Newby, "Viral Deepfake App ZAO Sparks Mass Downloads, Memes and Major Concerns," *Radii China*, 1 September 2019, accessed 1 September 2019.

[40] Allie Volpe, "Deepfake Porn Has Terrifying Implications. But What If It Could Be Used for Good?" *Men's Health*, 13 April 2018, accessed 7 August 2019;
Project Revoice, accessed 7 August 2019.

[41] *DeepPrivacy*, Github repository.

[42] Doing so also becomes exponentially more complex when the proliferation of the component techniques comprising deepfake generation is assessed as well. For example, generative adversarial networks (GANs) pit two competing neural nets against each other to build more accurate models, and are used in deepfake algorithms to test generated videos for detectability. GANs are also being used to generate high-resolution medical images from lower resolution 2D and 3D images, including CT scans and X-rays. Should regulating deepfakes involve regulating the use of GANs, and if so, how would that impact life-saving medical research? These are questions policymakers need to evaluate when preparing to dictate the direction of AI development.

positive use cases of deepfakes more thoroughly. Guidelines for prevention and punishment cannot undermine frameworks for the growth and development of deep learning applications.

4. HOW WE HAVE RESPONDED

Having extensively considered the real and potential applications of deepfakes for both positive and negative purposes, subnational governments, national governments, private companies, international bodies, and civil liberties groups have all taken steps or put forth recommendations for how to regulate deepfakes properly. These responses follow several trends. From the outset, researchers have continued developing tools to better detect deepfakes. Efforts to limit the creation of deepfakes, sponsored by subnational / national governments and private companies alike, have typically opposed the technology's use for the specific purposes of nonconsensual pornography, misuse of likenesses, and election manipulation. Regulations that affect deepfakes, furthermore, have not necessarily been driven by deepfake incidents, but instead by problems—violence, corruption, greater insecurity—driven by cheapfakes or fake news more broadly. Hesitation to regulate has primarily come from private companies and civil liberties organizations like the Electronic Frontier Foundation, which are more concerned with free speech or the appearance that it might be impeded with a blanket ban. They often recommend leveraging existing legal tools to crack down on the negative applications of deepfakes without inhibiting their beneficial uses. Notably, however, these organizations urging caution when it comes to deepfake regulation are typically located in regions (the US) where fake news has not resulted in death or large-scale violence, although it is not difficult to imagine a near future in which deepfakes are used to promote the same racist, white nationalist, or otherwise bigoted views that motivated the string of mass shooting that occurred in the US this summer.[43]

As regulation increasingly becomes about smoothly incorporating and normalizing a new technology into society, policymakers, journalists, and scholars need to evaluate positive use cases of deepfakes more thoroughly. Prevention and punishment cannot undermine frameworks for the growth and development of deep learning applications

4.1. Detection

New tools for deepfake detection, including analyzing facial patterns as sequences over time rather than stand-alone frames, have reached over 90% accuracy.[44] On September 5, 2019, Facebook teamed with partners in industry and academia to launch the Deepfake Detection Challenge, a contest providing useful datasets and grant opportunities to produce better deepfake detection technology.[45] A few weeks later, Google released a large database of visual deepfakes to aid with detection efforts, an act that parallels the release of a synthetic speech database to aid fake audio detection.[46]

[43] Neil Vigdor, "How Many Mass Shootings in 2019? Last Weekend Underscored the Violence," *The New York Times*, 3 August 2019, accessed 6 August 2019. Fake news has resulted in small-scale violence, the best example of which might be 2016's "Pizzagate" scandal. After fake news led him to believe that Hilary Clinton and other high-ranking Democrats were running a human trafficking and child sex ring out of a restaurant in DC, gunman Edgar Welch walked into the restaurant and fired shots with his semi-automatic rifle. Beyond even this example, it is difficult to measure what kinds of hatred and subsequent violence that fake news about different racial, religious, or national groups has produced.

[44] Ekraam Sabir et al., "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," arXiv:1905.00582, accessed 2 September 2019;
Will Knight, "A new deepfake detection tool should keep world leaders safe—for now," *MIT Technology Review*, 21 June 2019, accessed 10 July 2019.

[45] Mike Schroepfer, "Creating a data set and a challenge for deepfakes," *Facebook Artificial Intelligence*, 05 September 2019, accessed 05 September 2019.

[46] Nick Dufour and Andrew Gully, "Contributing Data to Deepfake Detection Research," *Google AI Blog*, 24 September 2019, accessed 24 September 2019.

But despite significant advancements in research and funding, some scholars conducting detection research assert that such efforts are inevitably pointless. "[A]t some point," said Hao Li, an associate professor studying deepfake detection with soft biometrics, "it's likely that it's not going to be possible to detect [AI fakes] at all."[47] Just as the generative adversarial networks (GANs) at the heart of deepfake software bounce between deepfake generation and detection, creators of deepfakes are consistently able to adjust their models to correct for the "tell" that made their earlier deepfakes detectable. In addition to algorithmic detection efforts, some groups have advocated for a watermarking tool or other mechanism by which a video's veracity could be tracked, for example with blockchain.[48] Some scholars have even gone so far as to suggest lifelong "authenticated alibi services" by which politicians and other public figures could prove their absence or presence from deepfake contexts.[49] But besides being laborious, easily avoided, and inaccessible to the majority of the population, such approaches could also become massive invasions of privacy. Data & Society's report "Deepfakes and Cheap Fakes" further emphasizes that regulation needs to account for the larger spectrum of social issues surrounding deepfakes—not only the problems that regulation could pose for free speech and privacy, but also the distrust of media and structural inequalities that fuel fake videos' effectiveness in the first place.[50]

Regulations that affect deepfakes, furthermore, have not necessarily been driven by deepfake incidents, but instead by problems—violence, corruption, greater insecurity—driven by cheapfakes or fake news more broadly

Any method of detecting or marking deepfakes is valuable because it attempts to deal with the problem at its source, but any method for approaching the technology that only involves detecting deepfakes does not do anything for effective management of the technology, including its positive uses. For that, regulatory guidelines are essential.

4.2. Private Regulation

Most recently, private companies have had to confront deepfake apps for invasion of privacy. Particularly in regions of the world where state surveillance and facial recognition are already being used to monitor the population, face swap apps and what they do with the images uploaded to them are a source of great controversy. The conflict between facial recognition technology—used both for deepfake creation and ethnic profiling, as is the case among the Uighur Muslim minority group in China—and privacy rages especially strongly.[51] The Chinese face swap app, ZAO, which began circulating on August 30, initially had a user agreement that allowed the company to use imagery created in the app for any purpose, without an option for users to revoke their permission for that usage.[52] In response to user concerns, on September 1 the company modified their agreement to require users' prior consent for using imagery created within the app, and allowing them to revoke permission as well.[53]

Regulation needs to account for the larger spectrum of social issues

[47] Vincent, "Deepfake detection algorithms."
[48] "Discussion summary: next level disinformation – deepfakes," *Friends of Europe*, 14 June 2019, accessed 2 August 2019.
[49] Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War," *Foreign Affairs*, 11 December 2018, accessed 28 June 2019.
[50] Britt Paris and Joan Donovan, *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence*, 18 September 2019.
[51] Paul Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority," *The New York Times*, 14 April 2019, accessed 12 July 2019.
[52] Newby, "Viral deepfake App ZAO."
[53] Ibid.

Beyond privacy concerns, which often remain intangible, deepfakes have also been condemned for propagating disinformation with serious physical consequences. Political deepfakes, for example, can impact public opinion, action, and ultimately elections. Despite political deepfakes being ostensibly of concern primarily to governments, governmental bodies were not the first to regulate them. Private companies—those on whose platforms deepfakes were initially proliferating—were the first to respond, not entirely by choice, to the threat of deepfakes in real time. The first wave of deepfakes, the explicit variety, was effectively banned in early 2018 by communication services and image hosting platforms such as Discord, Gfycat, Twitter, and Reddit, following extensive reporting on the phenomenon by *Vice*.[54] Existing rules about consent, or modified rules about revenge pornography, allowed these websites to stem the widest proliferation of deepfakes at their source. As the motivations for creating deepfakes shifted from gratuitous to commentative or deceptive, however, several American platforms faced difficulties deciding how to appropriately limit the spread of fake news without also impeding free expression. Various companies and artists—namely those previously mentioned, including BuzzFeed, Posters and Howe, and Xiao—propelled this debate by creating deepfakes warning about deepfakes, transforming threats into lessons.

Deepfakes have also been condemned for propagating disinformation with serious physical consequences

The largest social media platforms grappled with a precursor to deepfakes that demonstrated how complicated regulation would ultimately become. In July 2018 and May 2019, "cheapfakes"—synthetic videos made without the use of AI—of freshman congresswoman Alexandria Ocasio-Cortez and Speaker of the House Nancy Pelosi went viral to the tune of millions of views on Facebook, forcing the almost 2.4 billion-user company to address the scandals swiftly and publicly. In line with the company's tendency not to remove misinformation, Facebook did not remove the "satirical" AOC interview from its platform, and the video was circulated to over 1 million viewers.[55] Trump retweeted the Pelosi videos on Twitter, whose misinformation policies did not necessitate the videos' removal, and YouTube erased the videos entirely, but on the platform which enabled the widest spread of Pelosi's cheapfakes, Facebook hesitated.[56]

The largest social media platforms grappled with a precursor to deepfakes that demonstrated how complicated regulation would ultimately become

Facebook left the Pelosi videos untouched for several days before labeling them as "fake" and algorithmically deprioritizing them, an "execution mistake" that allowed them to get "more distribution than [Facebook's] policies should have allowed."[57] With a slightly shorter reaction time, that same approach was taken when a true deepfake, Posters and Howe's video of Zuckerberg, circulated on Facebook.[58] The social media site's prominence in sharing

[54] Rob Price, "Discord just shut down a chat group dedicated to sharing porn videos edited with AI to include celebrities," *Business Insider*, 26 January 2018, accessed 2 July 2019;
Samantha Cole, "AI-Generated Fake Porn Makers Have Been Kicked Off Their Favorite Host," *Vice*, 31 January 2018, accessed 2 July 2019;
Samantha Cole, "Reddit Just Shut Down the Deepfakes Subreddit," *Vice*, 7 February 2018, accessed 2 July 2019;
Samantha Cole, "Twitter Is the Latest Platform to Ban AI-Generated Porn," *Vice*, 7 February 2018, accessed 2 July 2019.

[55] Adi Robertson, "A million Facebook users watched a video that blurs the line between bad satire and 'fake news,'" *The Verge*, 24 July 2018, accessed 24 July 2019.

[56] Makena Kelly, "Distorted Nancy Pelosi videos show platforms aren't ready to fight dirty campaign tricks," *The Verge*, 24 May 2019, accessed 25 July 2019.

[57] Kurt Wagner, "Facebook CEO: Company Was Too Slow to Respond to Pelosi Deepfake," *Bloomberg*, 27 June 2019, accessed 24 July 2019.

[58] Samantha Cole, "This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies," *Vice*, 11 June 2019, accessed 26 June 2019.

cheapfakes and deepfakes, as well as its history of failing to detect Russian misinformation in other forms, prompted a discussion of deepfakes at the 2019 Aspen Ideas Festival.[59] Zuckerberg noted the possibility that the company would reevaluate its approach to regulating deepfakes.

In early August, Facebook-owned messaging service WhatsApp updated the latest version of their app to include a "frequently forwarded" feature that notifies users when a message they receive has been forwarded multiple times before.[60] This change holds significant consequence for users in India, who constitute WhatsApp's largest user base, because since at least 2017 the chain message format has been used to rapidly spread fake news and sow social disorder throughout the country.[61] The deepfake of journalist Rana Ayyub that circulated in April this year is one sophisticated example, but cheapfakes are in fact a more common—and more widely damaging—form of fake news in India than deepfakes. Videos of men harvesting organs and images of child kidnapping victims are circulated among villages, overlaid by local-language voice or text messages warning users that they could be harmed, robbed, or killed if they stray outdoors or don't stop criminals.[62] These warnings have led to mob lynchings and other forms of group violence against often innocent individuals, even though the images and videos in question are not of local crimes; one video of organ harvesters featured Spanish-language speakers, and several images of children's corpses were of children killed in Syria in 2013.[63] WhatsApp's newest update to aid users in identifying spam demonstrates an important instance in which regulations not specifically targeting deepfakes—in this case, the "frequently forwarded" feature that targets all forms of fake news—can still be helpful for limiting deepfakes' spread.

4.3. Non-Technical Responses

The leadership of one woman in the southern Indian state of Telangana also reveals how tackling fake news, in her case via non-technological means, could limit the effects of deepfakes. Police officer Rema Rajeshwari led her officers in personally visiting families in the 400 villages under her force's jurisdiction to inform them of the fake news epidemic, making use of the traditional skit-style storytelling format of *Janapadam* to communicate their warning message.[64] This personal, human, and highly local approach to combatting fake news has successfully prevented violence in Rajeshwari's villages. One can quickly extrapolate how the same approach of one-on-one awareness efforts could prevent equivalent violence sowed by more sophisticated fake news in the form of deepfakes. Her approach and that of WhatsApp both have benefits and drawbacks—the former is immediate, but tedious and hyperlocal; the latter is widespread, but it took a long time to roll out and is less direct (and thus less individually impactful) in its warning message. But both exemplify a means of keeping individuals' free speech intact while combatting the spread of fake news via education and awareness, an approach which can easily be transposed from fake news to deepfakes specifically.

Cheapfakes are in fact a more common—and more widely damaging—form of fake news in India than deepfakes

[59] Adam Satariano, "Facebook Identifies Russia-Linked Misinformation Campaign," *New York Times*, 17 January 2019, accessed 20 August 2019;
Alexis Madrigal, "Mark Zuckerberg Is Rethinking Deepfakes," *The Atlantic*, 26 June 2019, accessed 28 June 2019.
[60] Shweta Ganjoo, "WhatsApp starts rolling out frequently forwarded messages feature," *India Today*, 2 August 2019, accessed 5 August 2019.
[61] Maroosha Muzaffar, "She Keeps Fake News from Getting Deadly—Because WhatsApp Won't," *OZY*, 5 August 2019, accessed 5 August 2019.
[62] Timothy McLaughlin, "How WhatsApp Fuels Fake News and Violence in India," *Wired*, 12 December 2018, accessed 5 August 2019.
[63] Muzaffar, "She Keeps Fake News from Getting Deadly"; McLaughlin, "How WhatsApp Fuels Fake News."
[64] Ibid.

4.4. State and National Regulation

4.4.1. Within the US

Within the United States, where most of the social media platforms in question are headquartered (including Facebook, Twitter, and YouTube), several attempts at the state and federal level have been made to regulate deepfakes. Bills in California, New York, and Texas have sought to ban the technology on the grounds of mitigating electoral impropriety, nonconsensual use of likenesses, or disinformation at large.[65] In June 2019, Virginia became the first state to explicitly and officially ban deepfakes through an amendment to a law banning nonconsensual pornography; in September 2019, Texas became the second to do so as part of broader electoral integrity efforts; and in October 2019, California passed two pieces of legislation banning deepfakes on both counts.[66] Federal bills introduced in December 2018, June 2019, and September 2019 maintained similar lines of reasoning to promote regulation in an effort to limit the application of deepfakes for illegal activity.[67]

Some of the greatest long-term consequences of deepfakes centered on "truth decay," "long-term apathy," and a general erosion of the truth or any interest in identifying it

The nature of the national government's interest in deepfakes was best distilled in the House Intelligence Committee's June 13 hearing earlier this summer, an indictment of the emerging technology that emphasized its societal and national security implications.[68] Although short-term consequences could involve election insecurity, individual blackmail, or market manipulation, the experts argued that some of the greatest long-term consequences of deepfakes centered on "truth decay," "long-term apathy," and a general erosion of the truth or any interest in identifying it.[69] To solve the deepfake problem, Citron suggested a combination of "law, markets and social resiliences" to mitigate their spread.[70] The hearing emphasized that there is the need for collaboration between the government and social media companies, markets, and society at large; whatever form of cooperation takes place, it needs to span the sectors impacted by fake news in any form. The experts present noted the myriad difficulties of regulation.

There is the need for collaboration between the government and social media companies, markets, and society at large; whatever form of cooperation takes place, it needs to span the sectors impacted by fake news in any form

In broader anti-disinformation efforts which will encompass deepfake regulation, the US government has also mobilized military forces through the Defense Advanced Research Projects Agency (DARPA), the same organization responsible for "paving the way to the modern internet."[71] On August 28, DARPA

[65] Katyanna Quach, "New York State is trying to ban 'deepfakes' and Hollywood isn't happy," *The Register*, 12 June 2018, accessed 2 July 2019;
"Bill Targets 'Deepfake' Videos Before Election," *Techwire*, 26 June 2019, accessed 18 July 2019.
Kristin Houser, "New Law Makes It Illegal to Distribute Political Deepfakes," *Futurism*, 7 October 2019, accessed 1 November 2019.

[66] Adi Robertson, "Virginia's 'revenge porn' laws now officially cover deepfakes," *The Verge*, 1 July 2019, accessed 18 July 2019;
Lucas Ropek, "Handful of States Begin Legislating 'Deepfake' Videos," *Government Technology*, 30 April 2019, accessed 18 July 2019.

[67] Hayley Tsukayama, India McKinney, and Jamie Williams, "Congress Should Not Rush to Regulate Deepfakes," *Electronic Frontier Foundation*, 24 June 2019, accessed 26 June 2019;
Brandi Vincent, "Bill to Combat Deepfakes Passes House Committee," *Nextgov*, 26 September 2019, accessed 28 September 2019.

[68] "House Intelligence Committee."

[69] Ibid.

[70] Ibid.

[71] "Paving the Way to the Modern Internet," *DARPA*, accessed 31 August 2019.

began seeking developers for its Semantic Forensics (SemaFor) program, which plans to develop massive algorithmic systems for detecting, preempting, and defending against large-scale and automated disinformation attacks.[72] Although the organization and other scholars hold high hopes for the program's success, there is also consensus that it will not be effective without regulation in the form of legislation as well.[73]

4.4.2. Beyond the US

As regulatory powers move up the chain to broader governing bodies whose jurisdictions affect a greater number of people, deepfake regulations need to take into account the various competing interests of the groups they affect, the number of which typically increases relative to the breadth of influence. This is clearly evident in the US, but the conflict between self-interest and true protection of free speech or individual privacy is also proven in other countries. China has passed a law, which will go into effect January 2020, making it illegal not to disclose that a deepfake is not real, citing the technology's ability to "endanger national security, disrupt social stability, disrupt social order and infringe upon the legitimate rights and interests of others."[74] The state is also likely concerned, however, with deepfakes' potential for political activism against the CCP, making the technology's regulation in fact an effort against, rather than in support of, democracy.

Though their fears have yet to materialize, African research centers also express concern over the complications of progressing AI research, considering deepfakes could be mobilized by rogue non-governmental actors to inflame existing religious and political tensions in a manner similar to what is already occurring in India.[75] In Singapore, the Protection from Online Falsehoods and Manipulation bill claims to target potentially dangerous fake news, and bans the spread of falsehoods through online or other mobile network mediums, especially if that process is facilitated by bots or fake accounts.[76] Citizens also fear, however, that it allows the government to monitor private and encrypted communications.

4.4.3. International Bodies

The need to accommodate—and critically evaluate—diverse interests is most potently true on the international stage, where different economic and political motives could drive not just organizations but nations into conflict over the governance of an Internet that traverses geographic boundaries. In an EU strategy for online disinformation that explicitly noted the emergence of deepfakes, Brussels presented guidelines for enhancing the transparency, diversity, credibility, and inclusivity of the online information environment, sustainable with the possible creation of institutions like an EU-wide network of fact-checkers.[77] A video released by NATO StratCom COE in July 2019 discussed potential efforts to raise public awareness and resilience against deepfakes by engaging in educational efforts, such as creating a deepfake of Margaret Thatcher or Ronald Reagan complimenting their audience. For the most part, global efforts to regulate deepfakes have been subsumed within less granular debates about AI and disinformation at large.

4.5. How and Why We Have Not Responded

4.5.1. "Liars' Dividend"

One of the many reasons that a global response to deepfakes has yet to be sustained is due to what Chesney and Citron term the "liars'

[72] "Semantic Forensics (SemaFor) Proposers Day (Archived)," *DARPA*, 31 August 2019, accessed 2 September 2019.
[73] Pete Norman, "U.S. Unleashes Military to Fight Fake News, Disinformation," *Bloomberg*, 31 August 2019, accessed 31 August 2019.
[74] Yingzhi Yang and Brenda Goh, "China seeks to root out fake news and deepfakes with new online content rules," *Reuters*, 29 November 2019, accessed 3 December 2019.
[75] Clayton Besaw and John Filitz, "AI & Global Governance: AI in Africa is a Double-Edged Sword," *UNU-CPR Centre for Policy Research*, 16 January 2019, accessed 2 August 2019.
[76] Tessa Wong, "Singapore fake news law polices chats and online platforms," *BBC News*, 9 May 2019, accessed 14 May 2019.
[77] Charlotte Stanton, "How Should Countries Tackle Deepfakes?" *Carnegie Endowment for International Peace*, 28 January 2019, accessed 23 July 2019.

dividend" for raising awareness about deepfakes.[78] As the public becomes better educated on what deepfakes can be used for, they warn, it will become easier for deceptive subjects accused of wrongdoing to claim that their actions, as "caught" on film or photograph, are in fact only deepfakes. Educational efforts could backfire, then, by only further subverting the basis of trust necessary for the stability of any media environment.

4.5.2. Competing Interests

The Electronic Frontier Foundation has elaborated at length why a blanket ban on deepfakes, as recommended in Yvette Clarke's proposed DEEPFAKES Accountability Act (the federal bill put forth in June 2019), is so problematic: in addition to being overbroad and ineffective, the bill would trigger First Amendment complications within the US and exempt federal employees from its jurisdiction, upholding public safety in name but undermining it in practice.[79] Writers at the EFF stress that any form of congressional regulation would need to narrowly target only malicious deepfakes, and more generally remain cautious of government regulation of speech.

Any form of congressional regulation would need to narrowly target only malicious deepfakes

The EFF and congressional debates crystallize the manner in which, though they have not found success across the board, the plethora of bills in the US attempting to regulate deepfakes captures an essential microcosm of the debates surrounding the process at a grander scale. Within California, for example, the Screen Actors Guild-American Federation of Television and Radio Artists (SAG-AFTRA) is backing regulatory bill SB 564 on deepfakes, which organizations like Disney and the Motion Picture Association of America (MPAA) oppose.[80] The union primarily argues that deepfakes could violate artists' rights by enabling pornography with their likenesses, while the MPAA argues that banning deepfakes would cripple future attempts at creations such as biopics. Both sides, however, also serve to gain financially from their stances on deepfakes, with artists receiving paid acting work from the ban and film studios benefitting from the ability to include low-cost, high-tech human replicas in their productions. The competing interests of artists and companies, or of the state and the citizen, explain some of the stagnation that has been witnessed around developing a robust and unified response to deepfakes.

4.5.3. Internet Infrastructure

Besides the absence of money, political will, or fear about the "liars' dividend," yet another important argument has been posed against regulating deepfakes. A final category of actors has notably and intentionally remained reticent in addressing deepfake regulation. Web infrastructure and content delivery network (CDN) companies such as Akamai, Amazon CloudFront, and Google Cloud CDN provide globally distributed web server networks that allow users to access content efficiently wherever in the world they may try to do so.[81] As private companies, CDN operators can technically terminate services for websites distributing content that they do not support. Cloudflare, a San Francisco-based web infrastructure and security company, as well as one of the largest companies in the CDN space, has faced a great deal of controversy for its decisions both to continue and to stop offering support for online hate and terrorist groups. In 2017, partially in response to mounting public pressure, Cloudflare stopped providing cybersecurity services for neo-Nazi and white supremacist website *Daily Stormer*.[82] Their decision, though admittedly within their legal rights as a company, prompted significant backlash for surpassing the content-neutral bounds of

[78] Chesney and Citron, "Deep Fakes," 28-29.
[79] Tsukayama, McKinney, and Williams, "Congress Should Not Rush."
[80] Ed Targett, "California Moves Closer to Making Deepfake Pornography Illegal," *Computer Business Review*, 16 May 2019, accessed 18 July 2019.
[81] "Content Delivery Network Explained," *GlobalDots*, accessed 7 August 2019.
[82] Matthew Prince, "Why We Terminated Daily Stormer," *The Cloudflare Blog*, 16 August 2017, accessed 7 August 2019.

networks and setting a dangerous precedent for limiting free speech.[83]

Cloudflare's decision to openly discuss, and even memorialize, the criticism it received in 2017 mirrored the very debate that preceded the company's recent decision to terminate service for forum website 8chan.[84] Following the mass shooting that occurred in El Paso, Texas on August 3, 2019, investigations revealed that the gunman posted a racist manifesto on 8chan prior to his act of terrorism. His document explicitly referenced the mass shootings that targeted mosques in Christchurch, New Zealand in March 2019, which were themselves announced and subsequently glorified on 8chan; the white supremacist responsible in Christchurch also posted a screed to the website prior to his attacks.[85] Cloudflare cofounder and CEO Matthew Prince cited these incidents and others in his statement about his company's decision to terminate service to 8chan, calling the website a "cesspool of hate" driven by "lawlessness."[86] The oxymoron implied is that, in the end, "[h]is decision was lawless too."[87] Whether his decision would best serve the public interest—keeping 8chan and its brethren online, besides upholding the right to free speech that should always guide content moderation on the Internet, also keeps these fora off the dark web and in the eyes of law enforcement—did not rely on rigorous legal procedure as much as an arbitrary scale of pros and cons. Regardless, it was only a matter of time before 8chan came back online, and since being booted by Cloudflare it has shifted the debate over its right to exist onto multiple other platforms in an attempt to survive.[88]

As the internet becomes a, if not the, primary means of facilitating free speech, control over its governance will continue to transition into the hands of private companies that lack the oversight of governance laws originally intended to constrain governments

CDN companies are not the only organizations that can regulate content on the Internet.[89] Cloudflare is just one example of a company that, though once a neutral utility provider, has become a content moderator—a transition that has occurred without precedent, protection, or consequence (beyond public lashings). As the internet becomes a, if not the, primary means of facilitating free speech, control over its governance will continue to transition into the hands of private companies that lack the oversight of governance laws originally intended to constrain governments, although it should be promising that they have historically demonstrated restraint in regulating content.

Deepfakes are one of many viable vehicles for hate speech online. They will thus fall under the umbrella of under-the-radar, global content

[83] "Criticism Received by Cloudflare for Content Censorship," *Cloudflare*, accessed 7 August 2019.
[84] Matthew Prince, "Terminating Service for 8Chan," *The Cloudflare Blog*, 5 August 2019, accessed 6 August 2019.
[85] Queenie Wong, Richard Nieva, "El Paso massacre shines light on 8chan, a racist troll haven," *CNET*, 6 August 2019, accessed 7 August 2019.
The Christchurch shootings posed a unique problem for online platforms because the killer livestreamed almost 20 minutes of his actions on Facebook, clips from which were amplified on YouTube and by news outlets in the immediate aftermath of the attack. The terrorism that New Zealand experienced in March 2019 exemplified the once-distant difficulty that private companies face when making decisions about content regulation online. One expert on the far-right described regulating right-wing extremists as a "whack-a-mole" exercise, in that as soon as a website is taken down, it will find another platform on which to re-emerge. Even extremist content, however, is arguably more straightforward to regulate because some of it is objectively illegal content that violates preexisting codes or terms of service on media platforms; for these black-and-white distinctions, the debate largely focuses on whose role it is to lay down the punishing fist. Not only do debates about deepfakes address responsibility, but deepfakes themselves constitute such a broad category of content that blanket regulations or blanket bans on the technology are likely to overstep the bounds of free expression.
[86] Prince, "Terminating Service."

[87] Evelyn Douek, "The Lawless Way to Disable 8chan," *The Atlantic*, 6 August 2019, accessed 6 August 2019.
[88] Kate Conger and Nathaniel Popper, "Behind the Scenes, 8chan Scrambles to Get Back Online," *New York Times*, 5 August 2019, accessed 20 August 2019.
[89] In his 2017 blog post, Prince provides an extensive list of the organizations that are technically capable of intervening, which also includes Internet Service Providers (ISPs), registries, registrars, and more. The complete list is available here.

regulation performed by the companies comprising the skeletal structure of the Internet. When tracing the regulation of deepfakes and all other forms of communication online, scholars and journalists need to regularly evaluate not only what groups *are* responding to volatile content, but also what groups are *not*. This self-restraint might not indicate private companies' sincere belief in protecting free speech. But it does demonstrate that, at least in the present media environment, it is in these companies' self-interest to exhibit such an opinion to the public. That will be a short-term safeguard against the private sector's intrusion on the public's right to free speech.

Scholars and journalists need to regularly evaluate not only what groups are *responding to volatile content, but also what groups are* not

CONCLUSION

The spate of discussion about deepfakes to which 2019 has been witness points regulation in no obvious direction. It will take many years, and likely some consequential deepfake incidents, for policymakers or companies to determine any consistent approach to mitigating their negative consequences without stifling free speech, creative expression, or self-representation. There are, however, a few steps that can be taken to facilitate progress toward appropriate regulatory frameworks.

The first step—step zero, perhaps—is to understand that the root problem is human, not technological. Cybersecurity compromisers, like deepfakes, are able to operate as harmful elements because of the fragility of the media environment they infiltrate. Viewers are capable and often even willing to believe fake news, particularly multisensory information such as deepfakes, because of a base distrust, which itself might stem from a variety of sources: dislike of a deepfaked subject, perceived media bias, or public attacks on the media and other institutions that underpin democracy.[90] The diversity of both the Internet-using population and the reasons for fake news' effectiveness is important to remember when combatting deepfakes, but it should not be an excuse to pause the creation of true safeguards against their negative applications.

To that end, researchers and companies need to continue what they are already doing to combat specifically harmful uses of deepfakes. Regardless of its inevitable "uselessness," research into deepfake detection methods can help stall the immediate effects of their believability before they spread further than they would otherwise. Even if their responses are short-sighted or overly narrow, private companies working *ex post facto* are sure to target real, and not just imagined, harms. Equivocating in responses to deepfakes only gives them more time to propagate; immediate action in the form of regulatory bans and detective work must continue.

As ongoing work conducts damage control at the level of freer-acting private companies, legal regulations need to be deliberated and slowly put in place. Protecting electoral integrity, information security, and women from harassment are all valuable reasons to do so that span societal and international scales.

Finally, to prevent this regulatory step from overwhelming the positive forms of deepfakes, governments and companies need to increase funding into research and development for the beneficial applications of deepfakes. The more we can build an understanding of this technology as a positive and sophisticated outlet for art, politics, entertainment, the quicker we can diminish the inordinate fear surrounding its use. Deepfakes are more than just weaponized masks; they are also art, outlet, self-expression, and both a creator and creation of immense technological progress.

The root problem is human, not technological

[90] "Indicators of News Media Trust," *Gallup* and *Knight Foundation*, 2018, accessed 29 August 2019.

Like much of internet-related policy, current standards for dealing with deepfakes have primarily been reactionary changes to singular incidents or persistent trends, rather than preemptive or preventative guidelines. This "call and response" style of policy construction is limited and limiting in its scope. It evokes excessively cautious restrictions and permits malicious actors to restrict individual freedoms by capitalizing on a hyper-wary environment. Future deepfake regulations, if grounded in optimism rather than fear about the technology's potential, can avoid these pitfalls. Encouraging hope about deepfakes may be difficult as the technology continues to be used as a vehicle for personal and political attacks. It is also, however, an invaluable step toward eliminating the mythos that surrounds this and other disruptive technologies—and, ultimately, confronting the complexity of the cutting-edge technology that is coming to define our life and times.

"Call and response" style of policy construction is limited and limiting in its scope. It evokes excessively cautious restrictions and permits malicious actors to restrict individual freedoms by capitalizing on a hyper-wary environment. Future deepfake regulations, if grounded in optimism rather than fear about the technology's potential, can avoid these pitfalls

About the Author

Emilia Anna Porubcin was a research intern at the International Centre for Defence and Security in the summer of 2019. She is currently a third-year BA student at Stanford University, where she is studying history and computer science, specializing in Russia and Eastern Europe. Her internship was part of the long-standing cooperation between the ICDS and Freeman Spogli Institute for International Studies of Stanford University. Since 2018, she has worked as a research assistant at the Center for Internet and Society, where she focuses on smart tech and consumer privacy.

Follow us on:

ICDS.Tallinn
@ICDS_Tallinn
icds-tallinn

International Centre for Defence and Security
63/4 Narva Rd., 10152 Tallinn, Estonia
info@icds.ee, www.icds.ee

Issn 2228-2076