

ONLINE IMPOSTERS AND DISINFORMATION

HEARING BEFORE THE SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT OF THE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY HOUSE OF REPRESENTATIVES ONE HUNDRED SIXTEENTH CONGRESS

FIRST SESSION

SEPTEMBER 26, 2019

Serial No. 116–47

Printed for the use of the Committee on Science, Space, and Technology



Available via the World Wide Web: <http://science.house.gov>

U.S. GOVERNMENT PUBLISHING OFFICE

37–739PDF

WASHINGTON : 2020

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

HON. EDDIE BERNICE JOHNSON, Texas, *Chairwoman*

ZOE LOFGREN, California	FRANK D. LUCAS, Oklahoma,
DANIEL LIPINSKI, Illinois	<i>Ranking Member</i>
SUZANNE BONAMICI, Oregon	MO BROOKS, Alabama
AMI BERA, California,	BILL POSEY, Florida
<i>Vice Chair</i>	RANDY WEBER, Texas
CONOR LAMB, Pennsylvania	BRIAN BABIN, Texas
LIZZIE FLETCHER, Texas	ANDY BIGGS, Arizona
HALEY STEVENS, Michigan	ROGER MARSHALL, Kansas
KENDRA HORN, Oklahoma	RALPH NORMAN, South Carolina
MIKIE SHERRILL, New Jersey	MICHAEL CLOUD, Texas
BRAD SHERMAN, California	TROY BALDERSON, Ohio
STEVE COHEN, Tennessee	PETE OLSON, Texas
JERRY McNERNEY, California	ANTHONY GONZALEZ, Ohio
ED PERLMUTTER, Colorado	MICHAEL WALTZ, Florida
PAUL TONKO, New York	JIM BAIRD, Indiana
BILL FOSTER, Illinois	JAIME HERRERA BEUTLER, Washington
DON BEYER, Virginia	JENNIFFER GONZALEZ-COLÓN, Puerto Rico
CHARLIE CRIST, Florida	VACANCY
SEAN CASTEN, Illinois	
KATIE HILL, California	
BEN McADAMS, Utah	
JENNIFER WEXTON, Virginia	

SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT

HON. MIKIE SHERRILL, New Jersey, *Chairwoman*

SUZANNE BONAMICI, Oregon	RALPH NORMAN, South Carolina, <i>Ranking</i>
STEVE COHEN, Tennessee	<i>Member</i>
DON BEYER, Virginia	ANDY BIGGS, Arizona
JENNIFER WEXTON, Virginia	MICHAEL WALTZ, Florida

C O N T E N T S

September 26, 2019

	Page
Hearing Charter	2
Opening Statements	
Statement by Representative Mikie Sherrill, Chairwoman, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, U.S. House of Representatives	10
Written Statement	11
Statement by Representative Frank Lucas, Ranking Member, Committee on Science, Space, and Technology, U.S. House of Representatives	12
Written statement	12
Statement by Representative Don Beyer, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, U.S. House of Representatives	13
Statement by Representative Michael Waltz, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, U.S. House of Representatives	14
Written statement	14
Written statement by Representative Eddie Bernice Johnson, Chairwoman, Committee on Science, Space, and Technology, U.S. House of Representatives	15
Written statement by Representative Ralph Norman, Ranking Member, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, U.S. House of Representatives	16
Witnesses:	
Dr. Siwei Lyu, Director, Computer Vision and Machine Learning Lab, SUNY - Albany	
Oral Statement	17
Written Statement	19
Dr. Hany Farid, Professor of Electrical Engineering and Computer Science and the School of Information, UC, Berkeley	
Oral Statement	24
Written Statement	26
Ms. Camille Francois, Chief Innovation Officer, Graphika	
Oral Statement	31
Written Statement	33
Discussion	38
Appendix: Additional Material for the Record	
Report submitted by Ms. Camille Francois, Chief Innovation Officer, Graphika	58

ONLINE IMPOSTERS AND DISINFORMATION

THURSDAY, SEPTEMBER 26, 2019

HOUSE OF REPRESENTATIVES,
SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT,
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY,
Washington, D.C.

The Subcommittee met, pursuant to notice, at 2:01 p.m., in room 2318 of the Rayburn House Office Building, Hon. Mikie Sherrill [Chairwoman of the Subcommittee] presiding.

**COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT
U.S. HOUSE OF REPRESENTATIVES**

HEARING CHARTER

Online Imposters and Disinformation

Thursday, September 26, 2019

2:00 p.m.

2318 Rayburn House Office Building

PURPOSE

The purpose of the hearing is to explore the enabling technologies for disinformation online, including deep fakes, explore trends and emerging technology in the field, and consider research strategies that can help stem the tide of malicious inauthentic behavior.

WITNESSES

- **Dr. Siwei Lyu**, Director, Computer Vision and Machine Learning Lab, SUNY - Albany
- **Dr. Hany Farid**, Professor of Electrical Engineering & Computer Science and the School of Information, UC-Berkeley
- **Ms. Camille Francois**, Chief Innovation Officer, Graphika

KEY QUESTIONS

- How extensive is online disinformation and how does it affect Americans?
- What are some emerging expressions of online disinformation and online imposters?
- Why is social media so readily exploited for spreading disinformation?
- What are some of the technology and social solutions we have today to combat malicious online imposters, deep fakes and disinformation?
- What are the research and workforce needs associated with these challenges?

BACKGROUND

Researchers generally define *misinformation* as information that is false or misleading but promulgated with sincerity by a person who believes it is true. *Disinformation* is shared with the deliberate intent to deceive. Disinformation represents an intentional effort to shape or distort public perceptions around a particular issue through the dissemination of false information – or opinions or true information being delivered by a false messenger. Global governments and domestic agitators have long used disinformation in its many forms as a weapon against their adversaries. The problem has become far more pervasive in the past decade with the explosive growth of social media, which provides an opportunity to hostile actors to project disinformation directly into the popular discourse at little cost. The natural anonymity of the internet also makes

it cheap and easy for bad actors to impersonate public figures or create fake personalities that don't exist in order to proliferate content.

The shape of disinformation is also changing because fake content and videos can now be created automatically through the use of artificial intelligence. AI programs can write convincing articles and blog posts that seem to be written by real humans, and AI can create fake videos featuring "people" who do not really exist. While audio has traditionally been the least advanced aspect of deep fake technology, recent breakthroughs have occurred that have substantially closed the gap between fake audio content and its parallel video content. AI achievements in this space are exploding, with new advances occurring every 2-3 months.

The phenomena of online imposters and disinformation are virtually inextricable. To the extent that all information, true or false, is received differently according to what messenger delivered it, perhaps all information proliferated by online imposters is disinformation. A relatively low percentage of what might be classified as "disinformation" online is counterfactual in the strictest sense of the word (e.g. "the Moon landing was faked"). A more frequent expression might be when a foreign-based troll or bot retweets or shares a political statement that itself would be either factually true or a matter of opinion (e.g. "Congressional Candidate X cares about family values"), but is revealed to be disingenuous/of false pretenses when the real author is revealed.

Online imposters and their activities can be classified into three baskets:

- **Digital astroturfing**, in which bots and trolls create multiple fake personalities to artificially boost an influence campaign, spread lies and/or enflame public discourse.
- **Digital imposters**, in which bad actors create bogus profiles for real public figures.¹
- **Deep fakes**, in which bad actors manipulate a public figure's physical likeness to create highly-realistic videos and audio clips.

Recent Episodes

Malicious online imposters and disinformation intended to cause harm or artificially influence social discourse take a wide variety of expressions that diversifies every year:

Astroturfing during political content: On September 21, Twitter announced it was suspending 4,258 accounts operating from the UAE that were proliferating a coordinated disinformation campaign targeting Qatar and Yemen.²

Astroturfing during political conflict: In August 2019, Twitter announced that it had identified and taken down 200,000 accounts linked to the Chinese government that sought to present an artificial groundswell of public opposition to the pro-democracy campaign in Hong Kong.³

¹ https://www.vice.com/en_us/article/ae5m7z/meet-the-people-pretending-to-be-celebrities-on-social-media

² <https://www.hongkongfp.com/2019/09/21/twitter-closes-thousands-fake-news-accounts-including-4302-chinese-accounts-targeting-hong-kong-protests/>

³ <https://www.reuters.com/article/us-hongkong-protests-twitter/twitter-facebook-accuse-china-of-using-fake-accounts-to-undermine-hong-kong-protests-idUSKCN1V91NX>

Twitter shared information about the accounts it had identified with Facebook, which was able to execute its own takedown of Facebook accounts.

Audio deep fakes: In August 2019, news media reported on the first known major incidence of an audio-only deep fake being used in a crime. Thieves used sophisticated voice-mimicking software to imitate a senior executive at a British energy company in order to convince a managing director to wire \$240,000 to an account in Hungary.⁴

Political impersonation: The intelligence firm FireEye reported that Iranians posing as Americans set up fake social media accounts in 2018 to interfere in U.S. Congressional races. Some of the fake accounts impersonated the candidates themselves. The researchers found that the impersonators were a part of a broader campaign by Iranian actors, which included entire fake news websites and letters and op-eds in a handful of U.S. newspapers.⁵

Nonconsensual pornography: In June 2019, a programmer created an app called DeepNude which allowed users to create extremely realistic nude renderings from images of clothed women – allowing the subjects to be victims of revenge porn without ever having taken a nude photo. Paradoxically, revenge porn laws are unlikely to prohibit these renderings because the victim’s actual nude image is not displayed, even while the forgeries can be difficult to detect.⁶

Synthetic political messaging with organic augmentation: In a February 2018 indictment, Special Counsel Robert Mueller described one effort by the Russian government to interfere with American political discourse by creating the @TEN_GOP Twitter handle to pose as the Tennessee Republican party.⁷ The account eventually gathered more than 100,000 followers, but its content was received by a much larger audience as it was retweeted by several prominent figures with larger followings.⁸ Twitter shut down the @TEN_GOP account in August 2017.

Political impersonation: In June 2018, a network of foreign actors created and promoted a fake Tweet from Senator Marco Rubio. The culprits did not make a fake account for Senator Rubio; rather, they Photoshopped one of his real tweets to change its content, then proliferated it as a screenshot across (non-Twitter) discussion forums. The tweet was not widely debunked in the media until the Atlantic Council’s Digital Forensics Lab discovered it a year later.^{9,10}

Fake reviews: A cottage industry dedicated to selling fake reviews on Amazon has developed in recent years. Vendors pay perpetrators to create profiles for “consumers” that seem authentic and then leave five-star reviews with detailed narrative descriptions. A project called ReviewMeta analyzed 203 million Amazon reviews and found 11.3% of them to be inauthentic. A similar

⁴ <https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/>

⁵ <https://www.rollicall.com/news/congress/iranians-may-used-influence-operations-2018-midterms>

⁶ https://www.vice.com/en_au/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman

⁷ <https://www.documentcloud.org/documents/4380502-Indictment.html#document/p15/a404968> P. 15

⁸ Including Eric Trump, Donald Trump Jr., Kellyanne Conway, and Roger Stone.

⁹ <https://medium.com/dfrilab/top-takes-suspected-russian-intelligence-operation-39212367d2f0>

¹⁰ <https://twitter.com/marcorubio/status/114146865660345488>

project called Fakespot argues the figure is closer to 30%. Amazon sued over 1,000 sellers for buying reviews from 2015-2018, but it also claims the incidence of fake reviews is <1%.¹¹¹²

Digital astroturfing: In May 2018, a Facebook page called the “WalkAway Movement” was created, and #WalkAway began to trend on Twitter. The page described #WalkAway as a “grassroots movement” of Democrats fleeing the party – but in fact, it was an astroturf campaign conducted by US-based political trolls, amplified by Russian troll farms.¹³ The Facebook page grew to 19,000 members in just one month. #Walkaway was the Number 1 most-used hashtag in overall use on Twitter by June 30. Not every account that retweeted #WalkAway was a bot, but participation from both bots and American citizens misrepresenting their identities was able to amplify an inauthentic campaign to artificially expand its reach.

Inauthentic representations of public figures: In the first hours after the Santa Fe High School shooting in May 2018, which left 10 dead, online hoaxers created fake Facebook accounts using the shooter’s name and doctored photos that linked him to both 2016 Presidential candidates Trump and Clinton. One disinformation analyst noted it took less than 20 minutes for the first fake Facebook account to be created after the suspect’s name was revealed.¹⁴

Acronyms and vocabulary

- **Bot** – an autonomous program that can “behave” like an authentic human user on social media. The more sophisticated algorithms the bot is following, the more realistic its behavior. Not all bots are malicious – e.g. the USGS Earthquake Notification Service is a bot.
- **Troll farm** – an organization of (human) internet users seeking to create conflict or promote inflammatory content online in some coordinated fashion.
- **Deep neural network** – a multi-layer set of algorithms, modeled loosely after the human brain, that uses sophisticated mathematical modeling to process data and information in complex ways, including recognizing patterns.
- **GAN** – generative adversarial network. A type of machine learning system. GAN is the technique most commonly associated with creating deep fakes, but is not the only method.
- **GPU** – graphical computing unit. An inexpensive, commercial available type of computing hardware that can be incorporated into a regular laptop or desktop computer in order make more sophisticated graphics, including deep fakes.
- **IP Address** – A unique numeric identifier assigned to every computer or smartphone. Most IP addresses are dynamic, i.e. change periodically, but the address will always include a signifier of the rough location of the device. Any user can search an IP address for its details

¹¹ <https://thehustle.co/amazon-fake-reviews>

¹² <https://www.npr.org/2018/07/30/629800775/some-amazon-reviews-are-too-good-to-be-believed-theyre-paid-for>

¹³ <https://arcdigital.media/pro-trump-russian-linked-twitter-accounts-are-posing-as-ex-democrats-in-new-astroturfed-movement-20359c1906d3>

¹⁴ Chris Sampson. <https://twitter.com/TAPSTRIMEDIA/status/997541195114000384>

to get a sense of its location (E.g. “U.S. House of Representatives”) as well as which ISP (below) is serving the connected device. IP addresses can be concealed using a VPN (below), but VPNs must be purchased and requires some extra steps on the part of the user.

- **ISP** – internet service provider. The utilities that provide internet access but do not adjudicate platforms of content. E.g. Comcast, Verizon, AT&T.
- **Recommendation algorithm** – the automated capability that Youtube and other platforms use to guess what kind of new content a user might be interested viewing based on the content they have already explored. The results are used to suggest new videos to watch, highlight stories to read and otherwise make content more prominent to the user.
- **VPN** – virtual private network. A method of connecting to the internet that (1) conceals a device’s IP address, and thus conceals its general geographic location; and (2) enhances privacy for the user by encrypting data that is transferred over WiFi.

Social media platforms’ approach

Social media platforms, researchers and federal agencies use a variety of terms to describe the challenge of online imposters and disinformation:

- **Facebook**, which is also the parent company of Instagram and Whatsapp, uses the term “coordinated inauthentic behavior.” This includes accounts run by humans impersonating real people; accounts for non-existent people, and users that promote content because they are paid to.
- **Twitter** focuses on “platform manipulation,” defined as “using Twitter to engage in bulk, aggressive or deceptive activity that misleads others and/or disrupts their experience.” This includes “inauthentic engagements” that attempt to make content appear more popular than it is and “coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting.”¹⁵
- **YouTube**, owned by Google, has specific terms of use policies against *Impersonation* and *Spam, deceptive practices & scams*.¹⁶ Under the latter category, Youtube names Voter Suppression (content aiming to mislead voters about the time, place, means or eligibility requirements for voting), incentivization spam (in which engagement metrics, such as likes and positive comments, are sold, and rep), posting of content that is autogenerated by computers in order to post it quickly, and repetitive/excessive posing of the same material in comments or videos. It also names “borderline content,” which seeks to misinform users in harmful ways but does not violate the black letter of its community standards agreement.

The First Amendment does not restrict private companies from moderating content on their own. But social media companies lean heavily on Section 230 of the Communications Decency Act, which became law in the 1990s, to limit their responsibilities to adjudicate disinformation online.

¹⁵ <https://help.twitter.com/en/rules-and-policies/platform-manipulation>

¹⁶ <https://support.google.com/youtube/answer/2801973?hl=en>

Section 230 says that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."¹⁷

Since late 2016, the major platforms have taken steps in fits and starts to address the issues of disinformation. In December 2016, Facebook announced its Third Party Fact-Checking project to help identify and debunk content that is demonstrably false. Facebook has conducted several updates to its News Feed recommendation algorithm to demote various types of false content when it is identified. For example, an update in June 2019 was designed to "reduce the reach" of posts that make sensational and misleading health claims.¹⁸ In September 2019, Facebook introduced a Deepfake Detection Challenge and committed \$10 million to fund the project, which will partner with Microsoft, Massachusetts Institute of Technology (MIT), the University of California-Berkeley, Cornell Tech, Oxford, and others.

Twitter has established a policy of releasing all information about imposter accounts it identifies and removes after the takedown has been executed, including all data on content proliferated by the troll or bot. This practice better allows researchers and cybersecurity professionals to better understand the state of the art in adversarial operations. In January 2019, Twitter also brought back the "chronological timeline," allowing users to read tweets from users that they follow according to when they were written. In 2016 Twitter had reorganized users' feeds with a recommendation algorithm, which tended to highlight more controversial viral content.¹⁹

In January 2019, Youtube announced that while they would continue to host videos were identified for perpetuating falsehoods, its recommendation algorithms would no longer recommend the videos to users.²⁰

Emerging Trends

Malicious actors who seek to misrepresent themselves or spread disinformation online are constantly evolving in their strategies as companies and "white hat" researchers grow more adept at combatting their previous approaches. Trends to anticipate on the five-year horizon:

- **Chinese-sponsored disinformation.**²¹
- **Instagram** will be a more frequent forum for disinformation via viral memes (e.g. photoshopped images) and digital astroturfing.²² Instagram may be more vulnerable because unlike Facebook, it does not have a Real Name Policy.
- **For-profit disinformation** will expand. Social media manipulation is already a more frequent offering in the bundle of services offered by public affairs companies. More hired guns are likely to conduct activities that interfere with Democratic institutions.²³

¹⁷ <https://www.law.cornell.edu/uscode/text/47/230>

¹⁸ <https://www.socialmediatoday.com/news/facebook-updates-news-feed-algorithm-to-demote-misleading-health-claims/558100/>

¹⁹ <https://www.engadget.com/2018/12/18/twitter-chronological-timeline-feature-latest-tweets/>

²⁰ <https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html>

²¹ https://www.npr.org/2019/09/05/757803903/experts-warn-u-s-should-prepare-for-election-interference-from-china?fbclid=IwAR3R_8TPQRWv42ppPZkNXP_o2Vw9BTmfOg9bR4RrZwdagaBktlrl2tbqzka

²² https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_election_2020_report?fr=sY2QzYzlOMjIMwMA

²³ *ibid*

- **Video deepfakes will grow more convincing.**
- Convincing **deepfake audio** renderings will be deployed with greater frequency.
- **Cyborg** accounts will be used to support digital astroturfing. These are accounts where the content is primarily created and posted by bots, but regular engagement from the (human) account owner makes the account's activity look more authentic and makes it harder for algorithms to identify the hallmarks of a bot.
- **Distributed operations.** The U.S. Cyber Command reportedly used offensive hacking to temporarily take down internet access for the Russian Internet Research Agency (IRA) on U.S. election day in November 2018. While this effort may have forestalled some of the activities the IRA had planned to carry out against American voters, Russia and other adversaries are likelier to disperse their activities across various operational centers in order to avoid single points of failure in the future.²⁴

Strategies to detect online imposters

Current strategies to identify deceptive online content fall into two categories: passive forensic detection and active forensic detection.

Passive forensic detection is the process of vetting pieces of content one at a time to determine whether they are authentic. The idea is to identify discrete qualities about the video, tweet, Facebook group or meme that suggest it may be false, misleading or presented by an imposter. As an example, imagine watching a video of a politician that has already been posted on Facebook and looking for clues as to whether it is a deep fake: is the subject blinking normally? Is the format consistent with older (verified) videos of this politician? Tech companies can also run online content through algorithms that automatically “look for” the hallmarks of deception. For example, the GIF-hosting company Gfycat has trained artificial intelligence to spot some types of fraudulent videos as they are uploaded by scouring the internet for other (authentic) versions of the photo or video and running an automatic comparison.²⁵

Active forensic detection is less widely utilized today. It seeks to build unique characteristics into digital content that can be used to proactively affirm that the content is authentic. The focus of these activities is to counter disinformation by identifying content that is real. For example, imagine a link to a video of a politician that is clearly marked as “verified content” because the content has already satisfied an active forensic detection review and has earned a digital signature of authenticity before being proliferated online. For example, a company called Factom uses blockchain technology to affirm the existence of a piece of data or a document at a certain time – assuring that a video was taken at a specific discrete time and that it was created by the specific camera attached to its digital signature.²⁶ An app called TruePic is designed to create a digital signature for visual content, so that information about a photo or video's true origin and metadata is made inextricable from the file itself.²⁷

Research Needs

²⁴ Ibid pg 5

²⁵ <https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/>

²⁶ <https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/>

²⁷ <https://truepic.com/technology/#controlled-capture>

A new research field is emerging around combating of disinformation which we might call *digital information forensics*. Practitioners in the field will need to adopt a holistic perspective on disinformation tools and tactics in an effort to develop and implement technological strategies to counter them. But experts are concerned that digital forensics efforts are currently being limited by the absence of a well-developed digital forensics workforce, which would consist of researchers who could identify disinformation and understand evolving methods to develop and disseminate it. The expansion of the digital forensics workforce through academic and professional programs could be an important foundation for combating disinformation.

Federal agencies also have a role to play in supporting anti-disinformation research efforts. Perhaps the leading federal effort to combat deep fakes is Defense Advanced Research Projects Agency (DARPA) Media Forensics program (MediFor), which is focused on developing applied technologies to detect inauthentic content and improve the integrity of digital content.²⁸ DARPA recently launched a follow-on program called Semantic Forensics – SemaFor – which seeks to develop innovative semantic technologies for automatically analyzing multi-modal media assets (text, audio, image, video) to defend against large-scale, automated disinformation attacks.²⁹

The National Institute of Standards and Technology (NIST) could have a significant impact by supporting more basic research in the field. One example of such research would be the creation of large, controlled, high-quality data sets for the academic sector to use in its research activities. NIST's ability to produce basic data under a broadly-acceptable standard could provide valuable support to researchers seeking applied solutions. The National Science Foundation (NSF) oversees a wealth of research related to cybersecurity and could serve as an important funding source for research on disinformation and online imposters.

²⁸ DARPA MediFor Program, <https://www.darpa.mil/program/media-forensics>.

²⁹ <https://www.darpa.mil/news-events/semantic-forensics-proposers-day>

Chairwoman SHERRILL. The hearing will now come to order. Good afternoon, and welcome to a hearing of the Investigations and Oversight Subcommittee. We're here today to discuss online impostors and disinformation. Researchers generally define misinformation as information that is false, but promulgated with sincerity by a person who believes it is true. Disinformation, on the other hand, is shared with the deliberate intent to deceive. It turns out that these days the concepts of disinformation and online impostors are almost one in the same. We all remember the classic scams and hoaxes from the early days of e-mail—a foreign prince needs help getting money out of the country. But today the more common brand of disinformation is not simply content that is plainly counterfactual, but that is being delivered by someone who is not who they say they are. We are seeing a surge in coordinated disinformation efforts, particularly around politicians, hot-button political issues, and democratic elections.

The 2016 cycle saw Russian troll farms interfering in the American discourse across Facebook, Twitter, and Instagram, trying to sway public opinion for their preferred candidate. But at the same time they were after something else much simpler—to create chaos. By driving a wedge into the social fissures in our society, sowing seeds of mistrust about our friends and neighbors, exploiting social discord, they think they might destabilize our democracy, and allow the oligarchy to look a little more attractive by comparison.

When I was a Russian Policy Officer in the Navy, I learned how central information warfare is in Russia's quest to dominate Western nations. And, unfortunately, modern technology makes information warfare a far easier proposition for antagonists—foreign or domestic. In fact, it's perhaps too easy today to proliferate convincing, harmful disinformation, build realistic renderings of people in videos, and impersonate others online. That's why the incidents of harmful episodes have exploded in the last few years. They range from fake reviewers misleading consumers on Amazon, to impersonating real political candidates, to fake pornography being created with the likenesses of real people. Earlier this year an alleged deep fake of the President of Gabon helped trigger an unsuccessful coup of the incumbent government. Deep fakes are particularly prone to being weaponized, as our very biology tells us that we can trust our eyes and our ears.

There are social science reasons why disinformation and online impostors are such a confounding challenge. Research has shown that online hoaxes spread 6 times as fast as true stories, for example. Maybe human nature just likes a good scandal, and confirmation bias shapes how we receive information every time we log on, or open an app. If we encounter a story, a video, or an influence campaign that seems a little less than authentic, we may still be inclined to believe it if the content supports the political narrative already playing in our own heads. Our digital antagonists, whether the intelligence service of a foreign adversary, or a lone wolf propagandist working from a laptop, know how to exploit all of this.

Our meeting today is the start of a conversation. Before we, as policymakers, can address the threat of fake news and online frauds, we have to understand how they operate, the tools we have today to address them, and where the next generation of bad actors

is headed. We need to know where to commit more resources, in the way of innovation and education. Our distinguished witnesses in today's panel are experts in the technologies that can be used to detect deep fakes and disinformation, and I'm glad they're here to help us explore these important issues. We're especially thankful that all three of you are able to roll with the punches when we had to move the hearing due a change in the congressional schedule, so thank you all. I'd also like to thank my Republican counterparts who have been such great partners in this matter. He will be here shortly, but Mr. Gonzalez of Ohio is joining us today to inform his work on deep fakes, and I'm proud to be a co-sponsor of his bill, H.R. 4355, here he is, and I thank you for being here, Mr. Gonzalez.

[The prepared statement of Chairwoman Sherrill follows:]

Good morning and welcome to a hearing of the Investigations and Oversight Subcommittee.

We're here today to discuss online imposters and disinformation. Researchers generally define *misinformation* as information that is false but promulgated with sincerity by a person who believes it is true. *Disinformation*, on the other hand, is shared with the deliberate intent to deceive.

It turns out that these days, the concepts of disinformation and online imposters are almost one and the same. We all remember the classic scams and hoaxes from the early days of email - a Nigerian Prince needs help getting money out of the country! But today, the more common brand of disinformation is not simply content that is plainly counterfactual, but that it is being delivered by someone who is not who they say they are.

We are seeing a surge in coordinated disinformation efforts particularly around politicians, hotbutton political issues, and democratic elections. The 2016 election cycle saw Russian troll farms interfering in the American discourse across Facebook, Twitter, Instagram, YouTube and beyond, trying to sway public opinion for their preferred candidate. But at the same time, they were after something else much simpler: to create chaos. By driving a wedge into the social fissures in our society, sowing seeds of mistrust about our friends and neighbors, exploiting social discord, they think they might destabilize our democracy and allow the oligarchy to look a little more attractive by comparison. When I was a Russian policy officer in the Navy, I learned how central information warfare is in Russia's quest to dominate western nations. And unfortunately, modern technology makes information warfare a far easier proposition for our antagonists, foreign or domestic.

In fact, its perhaps too easy today to proliferate convincing, harmful disinformation, build realistic renderings of people in videos, and impersonate others online. That's why the incidence of harmful episodes has exploded in the last few years. They range from fake reviewers misleading consumers on Amazon, to impersonating real political candidates, to fake pornography being created with the likenesses of real people. Earlier this year, an alleged deepfake of the President of Gabon helped trigger an unsuccessful coup of the incumbent government. Deep fakes are particularly prone to being weaponized, as our very biology tells us that we can trust our eyes and ears.

There are social science reasons why disinformation and online imposters are such a confounding challenge: research has shown that online hoaxes spread six times as fast as true stories, for example. Maybe human nature just likes a good scandal. And confirmation bias shapes how we receive information every time we log on or open an app. If we encounter a story, a video or an influence campaign that seems a little less than authentic, we may still be inclined to believe it if the content supports the political narrative already playing in our own heads. Our digital antagonists, whether the intelligence service of a foreign adversary or a lone wolf propagandist working from a laptop, know how to exploit all of this.

Our meeting today is the start of a conversation. Before we as policymakers can address the threat of fake news and online frauds, we have to understand how they operate, the tools we have today to address them, and where the next generation of bad actors is headed. We need to know where to commit more resources in the way of innovation and education.

Our distinguished witnesses on today's panel are experts in the technologies that can be used to detect deep fakes and disinformation, and I'm glad they are here to help us explore these important issues. We are especially thankful that all three of

you were able to roll with the punches when we had to move the hearing due to a change in the Congressional schedule.

I'd also like to thank my Republican counterparts who have been such great partners on this matter. Mr. Gonzalez of Ohio is joining us today to inform his work on deep fakes. I'm proud to be a cosponsor of his bill H.R. 4355, and I thank you for being here, Mr. Gonzalez.

Chairwoman SHERRILL. Unfortunately Ranking Member Norman could not be with us today, but we are happy to have the full Committee Ranking Member in his place, so the Chair now recognizes Mr. Lucas for an opening statement. Thank you, Mr. Lucas.

Mr. LUCAS. Thank you, Chairwoman Sherrill, for holding this hearing on the growing problem of disinformation on social media. We all know that photos these days can be digitally altered so easily that it's almost impossible to tell what's real and what's not. Now there's a growing problem where audio and video can be altered so convincingly that it can appear that someone has said or done something that never happened. These deep fakes can be produced more and more easily.

You know, there was once a rumor that I myself was a deep fake, just impersonating the real Frank Lucas. The good news, or, depending on your perspective, perhaps the bad news, is the technology hasn't come quite that far, and I'm the real deal. But once it's on the Internet, it never goes away. But deep fake technology is getting more and more sophisticated, and it's also getting easier to produce. As our witnesses will discuss today, the technology for generating deep fakes is improving at a rapid clip. Soon anyone with a decent computer, and access to training data, will be able to create increasingly convincing deep fakes that are difficult to detect and debunk. False and misleading content like this undermines public trust, and disrupts civil society. Unfortunately, the technology for generating deep fakes is developing at a speed and a scale that dwarfs the technology needed to detect and debunk deep fakes. We must help level the playing field.

This Committee took the first steps to do this yesterday by passing a bipartisan legislation aimed at improving research into the technology to detect deep fakes. I want to commend Representative Anthony Gonzalez for introducing this bill, and his leadership on the issue of technology and security. I often say that one of our most important jobs on the Science Committee is communicating to the American people the value of scientific research and development. Legislation and hearings like this are a great example of how the work we do here can benefit directly people across the country, and I look forward to hearing from our witnesses, and I yield back my time, Madam Chair.

[The prepared statement of Mr. Lucas follows:]

Thank you, Chairwoman Sherrill, for holding this hearing on the growing problem of disinformation on social media.

We all know that photos these days can be digitally altered so easily that it's all but impossible to tell what's real and what's not.

There's now a growing problem where audio and video can be altered so convincingly that it can appear that someone has said or done something that never happened. These deepfakes can be produced more and more easily.

You know, there was once a rumor that I MYSELF was a deepfake, just impersonating the real Frank Lucas. The good news-or maybe the bad news-is that technology hasn't come quite that far and I am the real deal.

But deepfake technology IS getting more sophisticated. And it's also getting easier to produce. As our witnesses will discuss today, the technology for generating

deepfakes is improving at a rapid clip. Soon, anyone with a decent computer and access to training data will be able to create increasingly convincing deepfakes that are difficult to detect and debunk.

False and misleading content like this undermines public trust and disrupts civil society.

Unfortunately, the technology for generating deepfakes is developing at a speed and scale that dwarfs the technology needed to detect and debunk deepfakes. We must help level the playing field.

This Committee took the first step to do that yesterday by passing bipartisan legislation aimed at improving research into the technology to detect deepfakes.

I want to commend Representative Anthony Gonzalez for introducing this bill and for his leadership on the issue of technology and security.

I often say that one of our most important jobs on the Science Committee is communicating to the American people the value of scientific research and development. Legislation and hearings like this are a great example of how the work we do here can directly benefit people across the country.

I look forward to hearing from our witnesses, and I yield back my time.

Chairwoman SHERRILL. Well, thank you, Ranking Member Lucas. And we have an additional opening statement today from my colleague across the aisle, Representative Waltz of Florida. Unfortunately, Mr. Waltz could not make it to the hearing today, but considering his great interest in the issue, I allowed him to submit a video of his opening statement, so we'll now hear from Mr. Waltz.

Mr. WALTZ. Hello, everyone. I'm sorry I can't be in town for the hearing today, but I wanted to make sure to share my concerns about digital impostors. Everyone in this room relies on social media, video messages, and other digital technology to connect with our constituents. We listen to their concerns, we share information about our work in Congress. But deep fake technology, which can literally put words in our mouths, undermines public trust in any digital communication. Today's witnesses will paint a picture of just how sophisticated the technology has become for creating realistic images, videos, and personalities online.

Before I conclude my statement, I want to say a few words about our distinguished Subcommittee Chairwoman, Mikie Sherrill. I think we can all agree that Mikie is one of the most intelligent, accomplished, and persuasive Members of Congress. In fact, she's so persuasive that she convinced me, a Green Beret, to cheer on Navy football in this year's rivalry game. Thanks, Chairwoman Sherrill, for bringing attention to the problems of deep fake technology, and go Navy, beat Army.

Chairwoman SHERRILL. What a pleasure. As you all saw that—thank you so much for your work. That was obviously a deep fake. That is what we're looking at, and that is what we're discussing today. Thank you so—right? How nice is that? And, sadly, knowing how deep the commitment to our respective services' football is, I do know that that was not actually your sentiment, although it should be. So thank you, Mr. Waltz and Mr. Beyer, for your willingness to participate in our deep fake demonstration, and thank you to our distinguished witnesses, Dr. Lyu, for creating this video.

I'll now recognize Mr. Beyer and Mr. Waltz for a few remarks. Mr. Beyer?

Mr. BEYER. Yes. Thank you, Madam Chair, very much. Congressman Waltz and I really had fun making the deep fake video. You can see that it clearly was in jest. As an Army brat, I would never throw a Green Beret under the bus. But you also see how dangerous and misleading it could be. I'm sure we fooled a couple of people. For instance, what if I had said, instead of go Navy, go beat Army, I had said, it's time to impeach the President? Well, that would be viral everywhere. I mean, the things would be ringing off the hook, and the social media—

Mr. WALTZ. Please do not do that to my staff.

Mr. BEYER. No. And Mr. Waltz would be the first to know, so my friends might appreciate it, but I don't think he would at all, so obviously the potential for serious harms with these deep fakes is quite great on elections, international stage for diplomatic purposes, and even for our private lives. That's why we, as a country, need to take swift action and invest in the research and the tools for identifying and combating deep fakes, and create a national strategy immediately, especially for election integrity, and ahead of the 2020 presidential election.

The stakes are high. We've got to act now. We already know of Russia's intentional campaign to spread disinformation throughout the last one, and I don't even want to imagine what Russia, or China, or just private players, the havoc they could wreak on our elections and on our personal lives. So thank you very much to Mikie Sherrill and Frank Lucas for leading this effort. I yield back.

Chairwoman SHERRILL. Thank you very much. Mr. Waltz?

Mr. WALTZ. Thank you, Madam Chairwoman. And while I do certainly hold you in the highest regard, that was not me. But, just to add to my colleagues, that's just an example, and a small example, of what a deep fake synthetic video can do. And we've seen this insidious capability. We're seeing, I think, the birth of it. But I certainly support my colleagues in how we can get our arms around this as a country. I think it's important to note that Mr. Beyer and I both consented to that video, but as, you know, putting words in the mouth of a U.S. Army Green Beret and cheering on for Navy is not the worst application of this technology, and it's certainly not difficult to imagine how our enemies or criminal groups can wreak havoc on governments, on elections, on businesses, on competitors, and the privacy of all Americans. So these videos, and this technology, have the potential to truly be a weapon for our adversaries.

We know that advanced deep fake technology exists within China and Russia. We know that they have the capability, and that both countries have demonstrated a willingness to use asymmetric warfare capabilities. So, as the technology for generating deep fakes improves, we do risk falling behind on the detection front. That's why this hearing is so important, and I certainly commend you for calling it. It will help us examine solutions for both detecting and debunking the deep fakes of the future. And, you know, at the end of the day, I just have to say go Army, beat Navy. I yield back.

[The prepared statement of Mr. Waltz follows:]

What you just saw was an example of a "deepfake," or synthetic video that can be generated thanks to advancements in artificial intelligence and machine learning.

As we have just seen, deepfakes have the ability to make people-myself included-appear as though they have said or done things that they have never said or done. And advancements in the underlying technology, as we will hear today, are making it much more difficult to distinguish an authentic recording from synthetic, deepfake impersonations.

Importantly, Mr. Beyer and I both consented to and participated in the creation of this deepfake. But a Green Beret cheering for Navy is not the worst application of the technology.

It's not difficult to imagine how deepfakes of nonconsenting individuals could be used to wreak havoc on governments, elections, business, and the privacy of individuals.

Deepfakes have the potential to be a weapon for our adversaries and we know that advanced deepfake technology exists in China and Russia and that both countries have asymmetric warfare capabilities.

As the technology for generating deepfakes improves, we risk falling behind on the detection front. That's why today's hearing is so important. It will help us examine solutions for detecting and debunking deepfakes of the future.

Thank you Chairwoman Sherrill and Ranking Member Norman for convening this important hearing.

Yield back.

Chairwoman SHERRILL. I don't know why I let you testify in my—no, thank you very much. Those were really sobering comments, and I appreciate you both for showing us a little bit of what we're contending with.

[The prepared statement of Chairwoman Johnson follows:]

Thank you Madam Chair, and I would like to join you in welcoming our witnesses this morning.

I'm glad we're holding this hearing today. It's worth acknowledging just how deeply the phenomenon of online disinformation affects most of our lives these days. As long as there's been news, there's been fake news. But the American people are far more connected than they used to be. And the new tools that enable fake images, misleading public discourse, even long passages of text are alarming in their sophistication. Maybe we all should have seen this coming, the explosion of disinformation that would accompany the information age.

I suspect my colleagues here in the House are already taking this matter seriously, because in a way, online imposters and twisted facts on the internet present a real and active threat to the way we do our own jobs. We all use social media to connect with our constituents and to hear about their concerns. My staff want to read the comments and the posts from the people in Dallas and hear what they have to say. If I am to believe that a large percentage of the comments on Twitter are coming from "bots" or some other source of disinformation, the waters get muddy very quickly.

We have to acknowledge the serious legacy of disinformation is in this country. In the late 1970s, I was working under President Carter as a Regional Director for the Department of Health. Around that time, the Soviet Union's KGB kicked off a campaign to plant the idea that the United States government invented HIV and AIDS at Fort Detrick. The KGB wrote bogus pamphlets and fake scientific research and distributed them at global conferences. It sold a complex narrative in which the United States military deliberately infected prisoners to create a public health crisis -- biological warfare against our own people. The KGB's efforts were so pervasive that by 1992, 15% of Americans considered it "definitely or probably true" that the AIDS virus was created deliberately in a government laboratory. Decades later, a 2005 study found that a substantial percentage of the African American community believed that AIDS was developed as a form of genocide against black people.

How absolutely devastating such disinformation can be. It is clear that information warfare can have such profound, destructive effects. I think it is long past time to recognize how vulnerable we are to the next generation of hostile actors.

As Chairwoman Sherrill said, the first step in addressing a big problem is understanding it. Not every Member of this Committee, myself included, is well-versed in what a "deep neural network" is or how a "GAN" works. However, we have a sense already that the federal government is likely to need to create new tools that address this issue.

We also need to have a serious conversation about what we expect from the social media platforms that so many of us use every day. These companies have enjoyed a level of growth and success that is only possible in the United States. They were created in garages and dorm rooms, but they stand on the shoulders of giants like DARPA, which created the internet, and the National Science Foundation, which developed the backbone of computer networks that allowed the internet to blossom. The American consumer has been overwhelmingly faithful to social media over the past decade. We will need those companies to help combat disinformation. It can no longer be ignored.

I am pleased to welcome our witnesses today, and I'm also pleased that we had bipartisan agreement in yesterday's markup on a bill that would enable more research on deep fakes. These issues require a bold bipartisan response. I thank my colleagues on both sides of the aisle for working together to address these important issues. With that, I yield back.

[The prepared statement of Mr. Norman follows:]

Good afternoon and thank you, Chairwoman Sherrill, for convening this important hearing.

We are here today to explore technologies that enable online disinformation. We'll look at trends and emerging technology in this field, and consider research strategies that can help to detect and combat sophisticated deceptions and so-called "deepfakes."

Disinformation is not new. It has been used throughout history to influence and mislead people.

What is new, however, is how modern technology can create more and more realistic deceptions. Not only that, but modern disinformation can be spread more widely and targeted to intended audiences.

Although media manipulation is nothing new, it has long been limited to altering photos. Altering video footage was traditionally reserved for Hollywood studios and those with access to advanced technological capabilities and financial resources.

But today, progress in artificial intelligence and machine learning have reduced these barriers and made it easier than ever to create digital forgeries.

In 1994, it cost \$55 million to create convincing footage of Forrest Gump meeting JFK. Today, that technology is more sophisticated and widely available.

What's more, these fakes are growing more convincing and therefore more difficult to detect. A major concern is this: as deepfake technology becomes more accessible, the ability to generate deepfakes may outpace our ability to detect them.

Adding to the problem of sophisticated fakes is how easily they can spread. Global interconnectivity and social networking have democratized access to communication.

This means that almost anyone can publish almost anything and can distribute it at lightspeed across the globe.

As the internet and social media have expanded our access to information, technological advancements have also made it easier to push information to specific audiences.

Algorithms used by social media platforms are designed to engage users with content that is most likely to interest them. Bad actors can use this to better target disinformation.

For example, it is difficult to distinguish the techniques used in modern disinformation campaigns from the those used in ordinary online marketing and advertising campaigns.

Deepfakes alone are making online disinformation more problematic. But when combined with novel means for distributing disinformation to ever more targeted audiences, the threat is even greater.

Fortunately, we are here today to discuss these new twists to an old problem and to consider how science and technology can combat these challenges.

I look forward to an engaging discussion with our distinguished panel of witnesses on how we can better address online disinformation.

Thank you again, Chairwoman Sherrill, for holding this important hearing, and thanks to our witnesses for being here today to help us develop solutions to this challenge. I look forward to hearing your testimony.

I yield back.

Chairwoman SHERRILL. At this time I would like to introduce our three witnesses.

First we have Dr. Siwei Lyu. Dr. Lyu is a Professor at the University of Albany's College of Engineering and Applied Sciences. He is an expert in machine learning, and media forensics. Next is Dr. Hany Farid. Dr. Farid is a Professor at the University of California Berkeley School of Electrical Engineering and Computer Science and the School of Information. Dr. Farid's research focuses on digital forensics, image analysis, and human perception. Last we have Ms. Camille Francois. Ms. Francois is the Chief Innovation Officer at Graphika, a company that uses artificial intelligence to analyze online communities and social networks.

As our witnesses should know, you will each have 5 minutes for your spoken testimony. Your written testimony will be included in the record for the hearing. When you all have completed your spoken testimony, we will begin with questions. Each Member will have 5 minutes to question the panel. And we'll start with you, Dr. Lyu.

**TESTIMONY OF DR. SIWEI LYU,
PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE,
DIRECTOR, COMPUTER VISION AND MACHINE LEARNING LAB,
UNIVERSITY AT ALBANY, STATE UNIVERSITY OF NEW YORK**

Dr. LYU. Good afternoon, Chairwoman Sherrill, Ranking Member Lucas, and Members of the Committee. Thank you for inviting me today to discuss the emerging issue of deep fakes. You have just seen a deep fake video we created for this hearing, so let me first briefly describe how this video, and similar fake videos, are made.

Making a deep fake video requires a source and a target. In this case, the source was Representative Beyer, and the target was Representative Waltz. Mr. Beyer's staff was kind enough to prepare a video of the Congressman for this project. While Mr. Waltz's office consented to this video demonstration, it is important to know that we didn't use any video from his office. Instead, we conducted an Internet search for about 30 minutes, and found one suitable minute-long YouTube video of Mr. Waltz, and that's our target video. The next step involves a software tool we developed, which used deep neural networks to create the fake video. It is important to note that our tool does not use a generative adversary network, or GAN.

It first trains the deep neural network models using the source and the target video. It then used the models to extract facial expressions in the source video of Mr. Beyer, and generate a video of Mr. Waltz with the same facial expressions. The audio track is from the original video of Mr. Beyer, and was not modified. The training and the production are performed on a computer equipped with a graphical processing unit, or GPU. The computer and the GPU can be purchased from Amazon for about \$3,000. The overall training and production took about 8 hours, and were completely automated, after setting a few initial parameters.

So a similar process was also used to generate the fake videos that are being displayed on the screen right now. Although we do not distribute this particular software, true, similar software making deep fakes can be found on code-sharing platforms like GitHub, and are free for anyone to download and to use. With the abundance of online media we share, anyone is a potential target of a deep fake attack.

Currently there are active research developments to identify, contain, and obstruct deep fakes before they can inflict damages. The majority of such research is currently sponsored by DARPA (Defense Advanced Research Projects Agency), most notably the MediFor (Media Forensics) program. But it is also important that the Federal Government fund more research, through NSF (National Science Foundation), to combat deep fakes. As an emerging research area that does not fall squarely into existing AI (artificial intelligence) or cybersecurity programs, it may be wise to establish a new functional program at NSF dedicated to similar emerging technologies. It can serve as an initial catch-all for similar high-risk and high-impact research until either an existing program's mission is expanded, or a new dedicated program is established.

We should also examine the approaches we share software code and tools, especially those with potential negative impacts like deep fakes. Therefore, it may be wise to consider requiring NSF to

conduct reviews of sponsored AI research and enforcing controls on the release of software codes or tools with dual use nature. This will help to reduce the potential misuses of such technologies.

Last, but not least, education on responsible research should be an intrinsic part of AI research. Investigators should be fully aware of the potential impact of the sponsored research, and provide corresponding trainings to the graduate students and post-docs working on the project. Again, NSF could enforce such ethics training and best practices through a mandatory requirement to sponsored research projects. The creation of new cross-function NSF programs for emerging technologies, the introduction of controls on the release of NSF-funded AI research with potential dual use, and required ethics training for NSF-funded AI research will go far in defending against the emerging threat posted by deep fakes.

Thank you for having this hearing today, and giving me the opportunity to testify. I'm happy to answer any questions you may have. Thank you.

[The prepared statement of Dr. Lyu follows:]

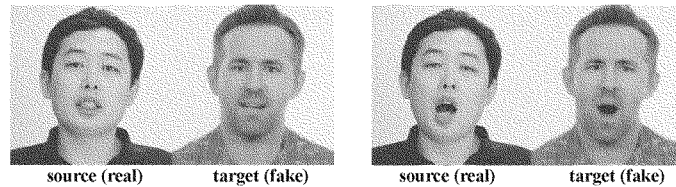
**House Committee on Science, Space, and Technology
Subcommittee on Investigations and Oversight
Online Imposters and Disinformation, 09/26/2019**

Siwei Lyu, Ph.D.
University at Albany, State University of New York

Backgrounds

Deepfakes are the most recent twist to the disconcerting problem of online disinformation. The term *deepfake* first emerged in late 2017 as the name of a Reddit account that began posting synthetic pornographic videos generated using an AI-based face-swapping algorithm. The term has subsequently become synonymous with three types of AI-generated impersonation videos.

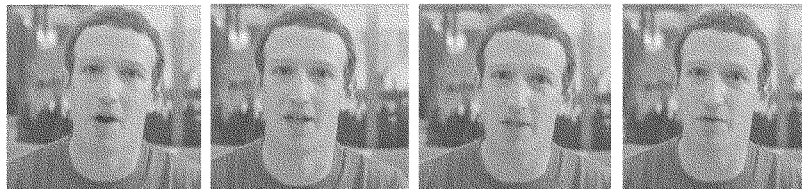
(1) **Head puppetry** entails synthesizing a video of a target person's head using a video of a source person's head, so the synthesized target appears to behave the same way as the source.



(2) **Face swapping** involves generating a video of the target with the faces replaced by synthesized faces of the source while keeping the same facial expressions.



(3) **Lip syncing** is to create a falsified video by only manipulating the lip region so that the target appears to speak something that s/he does not speak in reality.



[Source: Bill Posters and Daniel Howe, *The Spectre Project*]

Photos and videos have been doctored since their nascence. But there are three reasons why the current concerns over deepfakes and other AI-driven audio and visual media manipulations are justified. First, deepfakes can be made more easily, quickly, and with better quality — thanks to the rapid advancement of computer hardware and software technology, in particular those related to AI. Second, the capability to make deepfakes has been democratized through software tools that can be downloaded freely from online code sharing platforms.¹ Third, anyone with an online media presence is a potential target of a deepfake attack. A fake video showing a politician engaged in an inappropriate activity may be enough to sway an election if released close to voting day. A fake video of a falsified recording of a high-level executive commenting on his/her company's financial situation could potentially send the stock market awry. A fake video made by falsely implanting a woman's face in a pornographic video and shared on social-media platforms could tremendously traumatize the victim. The stakes are too high to ignore.

How are deepfakes made?

Deepfakes are created with a type of AI technology commonly known as deep neural networks.² A deep neural network model learns to synthesize realistic faces through *training*, which involves exposing the model to a large number of face images of different people with varying expression, head poses, and lighting conditions. Once the model is properly trained, it is ready to be used to generate deepfakes. Specifically, a face detection method is first run on the input video to locate the target's faces. Then facial landmarks corresponding to distinct locations such as the tips of eyes, eyebrows, nose, mouth, and contour of the face are extracted. Using these landmarks, the detected faces are warped to the same size and in a standard configuration. The standardized faces are fed to the deep neural network model to synthesize a new set of faces of different identity, which are then warped back to match the target's head orientations in the input.

Current computer hardware and AI technology has made it much easier to create deepfakes: a computer that is used to run the generation algorithm with a special computing hardware known as graphical computing unit (GPU) can be readily purchased for an affordable price on Amazon.³ The training videos for the targets can be easily downloaded from social-media platforms such as YouTube, Instagram, and Facebook in large volume and high quality. Convenient software tools have made the whole process automated barring the choice of a few parameters. As a result, a few good-quality, minute-long videos, a commodity computer with a GPU, and several hours of training are sufficient to generate deepfakes with decent visual quality.

¹ e.g., FakeAPP (used to be on Reddit but now defunct), DeepfaceLab (<https://github.com/iperov/DeepFaceLab>), faceswap-GAN (<https://github.com/shaoanlu/faceswap-GAN>), faceswap (<https://github.com/deepfakes/faceswa>), and more recently ZAO (<https://apkproz.com/app/zao>).

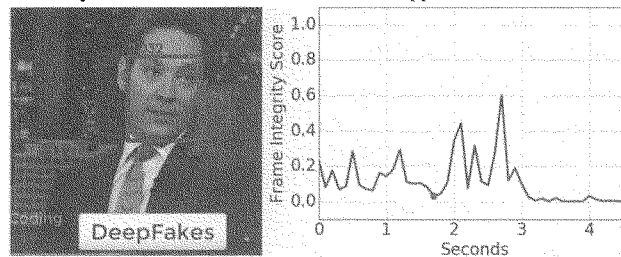
² GANs are only one type of deep neural network model for creating deepfakes. For example, the deepfake created for today's hearing did not use a GAN, but instead a different model known as the variational auto-encoders. This is important because any legislative or rule making effort to address deepfakes should not rely on a single tool. Instead Congress should attempt to future-proof regulation to cover the act instead of the tool.

³ An example of computer configuration for this purpose includes an HP-Z800 workstation (~\$1,000) equipped with an Nvidia 2080Ti GPU (~\$1,200) and other necessary peripherals. Cost effective and large-scale production can also be conducted using cloud platforms such as Amazon AWS or Google Cloud Platform.

How to combat deepfakes with technology development

While sophisticated deepfakes still take time and skill to produce, rough-and-ready fake videos may still cause harm. It is thus important to have effective technologies to identify, contain, and obstruct deepfakes before they can inflict damage. This should be done by focusing on improving forensic capabilities and making it harder to train deepfakes using online videos.

Effective deepfake forensic detection methods look for traces of the synthesis process to differentiate deepfakes from real videos. For instance, synthesized faces are warped and processed to fit the target's head orientation, such operations leave traces that can be exploited to detect deepfakes. Another type of detection techniques involves examining physiological inconsistencies such as the lack of realistic eye blinking and heart beating. A third approach is to “use AI to fight AI”, using another deep neural networks to detect deepfakes. State-of-the-art detection methods have shown promising accuracy on benchmark datasets, but their actual performance on real life deepfakes have yet to be tested.⁴ Also, due to the complex nature of deepfakes, no single type of technology or specific method will be the *silver bullet*, and an effective solution may come as a combination of all these approaches.



Detection results of a state-of-the-art deepfake detection method over a fake video on [youtube.com](https://www.youtube.com). The lower integrity score (range in $[0,1]$) suggests a video frame more likely to be generated using deepfake algorithms.

In addition to deepfake forensics, there are also technologies to prevent the re-use of online images and videos as training data for the deep neural network generating deepfakes. This would involve inserting imperceptible “adversarial noise” into images and videos before they are uploaded to online social-media platforms. The adversarial noise correspond to subtle perturbations that human eyes cannot see nonetheless can disrupt a face detection algorithm and make it difficult to automate the training process. A dedicated adversary could overcome adversarial noise by painstakingly selecting the target's face in every frame of a training video, but that requires 1,500 hand-marked selections for each 60 second training video.⁵

⁴ One notable effort towards this goal is the upcoming *Deepfake Detection Challenge* (<https://deepfakedetectionchallenge.ai>) sponsored by Facebook, Microsoft and Partnership on AI, to advance the state-of-the-art deepfake detection capacities.

⁵ This is calculated based on a target video quality of 25 frames per second, which is the lowest frame rate for uploaded YouTube videos. YouTube videos are uploaded at 60 frames per second, which would more than double the number of hand-marked selections for a 60 second video and the work to hand select faces.

Perspectives

As the underlying technology continues to develop, the current barriers to making deepfakes will fall and their quality will keep improving. What is also evolving is the quintessential cat-and-mouse game experienced by all attacker-defender relationships, and malicious attackers seem to have an upper-hand — they can adjust the generation algorithm whenever a new detection method is made public. Currently, the majority of research on combating deepfakes is sponsored under DARPA.⁶ But it is important that the federal government also fund more civilian research through NSF. One reason this has not yet happened is because the grant-making capabilities of NSF are focused around existing directorates that are not well equipped to support research into cross-functional emerging technologies. It may be wise to fund the establishment of a new *Emerging Technologies Directorate* at NSF, which can function as a catchall until either an existing directorate's mission is expanded or a new directorate is created. This would create a research home not just for deepfake forensics but also other emerging technologies.

The open-source model of disseminating research code is an enabling factor of the current deepfake problem and requires more scrutiny. The availability of easy-to-use and easy-to-access software tools has significantly lowered the technical threshold for an ordinary user to create deepfakes. A nation state with more manpower and computing resources can build upon them refined and customized versions to make more crafted deepfakes with higher level of realism and use them in a disinformation campaign. It may thus be wise to consider requiring NSF to conduct an ethics review of proposed grants around dual-use technology like deepfakes with mandatory controls on the release of the underlying technology into the proverbial wild.

Last but not least, education on responsible research should be an intrinsic part to the current AI research. Deepfakes add just one more item to the long list of various ethical issues of AI algorithms, such as built-in biases and prejudice, violations of individual privacy and safety, and the lack of accountability and transparency. As academic or industrial researchers working in these areas, we should recognize the potential impact of our research on society, and take them seriously as part of our due responsibilities. We should also provide training to students and post-docs on such issues. These practices could, again, be enforced through requirements from NSF on funded AI research that make such training and compliance mandatory.

Conclusions

It is not an exaggeration to say that we are on the cusp of deepfakes being cheap, easy to produce, indistinguishable from real videos, and ready to cause real damages. We therefore need a comprehensive and robust solution to this problem. The situation calls for continuous investment and perhaps an escalated funding level from the federal government to this strategically important research area. The situation surrounding deepfakes may not turn out to be as severe as we are predicting now. But it is better safe than sorry.

⁶ Most notably, the DARPA Media Forensics (MediFor) program (<https://www.darpa.mil/program/media-forensics>).

Biography

Siwei Lyu is a Professor of Computer Science and the Director of Computer Vision and Machine Learning Lab at University at Albany, State University of New York. Dr. Lyu received his Ph.D. degree in Computer Science in 2005 from Dartmouth College and under the supervision of Prof. Hany Farid, and his M.S. degree in Computer Science in 2000 and B.S. degree in Information Science in 1997, both from Peking University, China. Dr. Lyu's research interests include digital media forensics, computer vision, and machine learning. He is the recipient of the National Science Foundation CAREER Award in 2010, IEEE Signal Processing Best Paper Award, and the Google Faculty Research Award in 2019 for his work on digital media forensics. He has published one book on digital media forensics and over 120 refereed journal and conference papers in relevant research fields.

Chairwoman SHERRILL. Thank you very much. Dr. Farid?

**TESTIMONY OF DR. HANY FARID,
PROFESSOR, ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE AND THE SCHOOL OF INFORMATION,
UNIVERSITY OF CALIFORNIA, BERKELEY**

Dr. FARID. Chairwoman Sherrill, Ranking Member Lucas, and Members of the Committee, thanks for the opportunity to talk with you today on this important topic. Although disinformation is not new, what is new in the digital age is the sophistication with which fake content can be created, the democratization of access to sophisticated tools for manipulating content, and access to the Internet and social media, allowing for the delivery of disinformation with an unprecedented speed and reach.

The latest incarnation in creating fake audio, image, and video, so-called deep fakes, is being fueled by rapid advances in machine learning, and access to large amounts of data. Although there are several variations, the core machinery behind this technology is based on a combination of traditional techniques in computer vision and computer graphics, and more modern techniques from machine learning, namely deep neural networks. These technologies can, for example, from just hundreds of images of the Chairwoman, splice her likeness into a video sequence of someone else. Similar technologies can also be used to alter a video of the Chairwoman to make her mouth consistent with a new audio recording of her saying something that she never said. And, when paired with highly realistic voice synthesis technologies that can synthesize speech in a particular person's voice, these deep fakes can make a, for example, CEO announce that their profits are down, leading to global stock manipulation; a world leader announcing military action, leading to global conflict; or a Presidential candidate confessing complicity to a crime, leading to the disruption of an election.

The past 2 years have seen a remarkable increase in the quality and sophistication of these deep fakes. These technologies are not, however, just relegated to academic circles or Hollywood studios, but are freely available online, and have already been incorporated into commercial applications. The field of digital forensics is focused on developing technologies for detecting manipulated or synthesized audio, images, and video, and within this field there are two broad categories: Proactive and reactive techniques.

Proactive techniques work by using a specialized camera software to extract a digital signature from a recorded image or video. This digital signature can then be used in the future to determine if the content was manipulated from the time of recording. The benefit of this approach is that the technology is well-understood and developed. It's effective, and it is able to work at the scale of analyzing billions of uploads a day. The drawback is that it requires all of us to use specialized camera software, as opposed to the default camera app that we are all used to using, and it requires the collaboration of social media giants to incorporate these signatures and corresponding labels into their systems.

Notice that these proactive techniques tell us what is real, not what is fake. In contrast, reactive techniques are focused on telling us what is fake. These techniques work on the assumption that

digital manipulation leaves behind certain statistical, geometric, or physical traces that, although not necessarily visually obvious, can be modeled and algorithmically detected. The benefit of these techniques is that they don't require any specialized hardware or software. The drawback is that, even despite advances in the field, there are no universal forensic techniques that can operate at the scale and speed needed to analyze billions of uploads a day.

So, where do we go from here? Four points. One, funding agencies should invest at least as much financial support to programs in digital forensics as they are in programs that are fueling advances that are leading to the creation of, for example, deep fakes. Two, researchers that are developing technologies that can be weaponized should give more thought to how they can put proper safeguards in place so that their technologies are not misused. Three, no matter how quickly forensic technology advances, it will be useless without the collaboration of the giants of the technology sector. The major technology companies, including Facebook, Google, YouTube, and Twitter, must more aggressively and proactively develop and deploy technologies to combat disinformation campaigns. And four, we should not ignore the non-technical component of the issue of disinformation, us—the users. We need to better educate the public on how to consume trusted information, and not spread disinformation.

I'll close with two final points. First, although there are serious issues of online privacy, moves by some of the technology giants to transform their platform to an end-to-end encrypted system will make it even more difficult to slow or stop the spread of disinformation. We should find a balance between privacy and security, and not sacrifice one for the other. And, last, I'd like to re-emphasize that disinformation is not new, and deep fakes is only the latest incarnation. We should not lose sight of the fact that more traditional human-generated disinformation campaigns are still highly effective, and we will undoubtedly be contending with yet another technological innovation a few years from now. In responding to deep fakes, therefore, we should consider the past, the present, and the future as we try to navigate the complex interplay of technology, policy, and regulation, and I'm sorry I'm 15 seconds over.

[The prepared statement of Dr. Farid follows:]

Committee on Science, Space, and Technology *Online Imposters and Disinformation*

Hany Farid, Ph.D.

Background

Rumors quickly spread in Trent, Italy that members of the Jewish community murdered a young boy and drained and drank his blood to celebrate Passover. Before long, the city's entire Jewish community is arrested and tortured, fifteen of which are found guilty and executed. The year was 1475.

Fast forward to 2018. Rumors quickly spread in Athimoor-Kaliyam, India that roving gangs are kidnapping children. Over a period of several months, nearly two dozen innocent people are dragged from their vehicles and killed. The rumors this time spread through WhatsApp instead of word of mouth.

Disinformation is not new, nor are its deadly consequences. What is new, thanks to the internet and social media, is its reach and frequency. Today, disinformation propagates around the world at the speed of light. From small- to large-scale fraud, to sowing civil unrest, interfering with democratic elections, and inciting violence, disinformation campaigns today are leading to dangerous and deadly outcomes.

Add to this phenomenon the ability to create increasingly more compelling and sophisticated fake videos of anybody saying and doing anything – popularly referred to as deep fakes – and the threat only increases. This is the landscape that awaits us in 2019 and beyond.

Creating Deep Fakes

Advances in machine learning and access to large and diverse data sets have led to computer systems that are able to synthesize images of people who don't exist, videos of people doing things they never did, and audio recordings of them saying things they never said. These deep fakes are a dangerous

addition to an already volatile on-line world in which rumors, conspiracies, and disinformation spread often and quickly.

By providing millions of images of people to a machine-learning system, the system can learn to synthesize realistic images of people who don't exist. Similar technologies can, in live-stream videos, convert an adult face into a child's face, raising concerns that this technology will be used by child predators.

With just hundreds of images of someone, a machine-learning system can learn to insert them into a video. This face-swap deep fake can be highly entertaining, as in its use to insert Nic Cage into movies in which he never appeared. The same technology, however, can also be used to create non-consensual pornography or to impersonate a world leader. Similar technologies can also be used to alter a video to make a person's mouth consistent with a new audio recording of them saying something that they never said. When paired with highly realistic voice synthesis technologies that can synthesize speech in a particular person's voice, these lip-sync deep fakes can make a CEO announce that their profits are down, leading to global stock manipulation, a world-leader announce military action, leading to global conflict, or a presidential candidate confess complicity in a crime, leading to the disruption of an election.

What is perhaps most alarming about these deep-fake technologies is that they are not only in the hands of sophisticated Hollywood studios. Software to generate fake content is widely and freely available on-line, putting in the hands of many the ability to create increasingly compelling and sophisticated fakes. Coupled with the speed and reach of social media, convincing fake content can instantaneously reach millions.

How do we manage a digital landscape when it becomes increasingly more difficult to believe not just what we read, but also what we see and hear with our own eyes and ears? How do we manage a digital landscape where if anything can be fake, then everyone has plausible deniability to claim that any digital evidence is fake?

Detecting Deep Fakes

Despite efforts by digital forensic researchers, no current technology exists that can contend with the vast array of different types of deep fakes at a speed and accuracy that can be deployed at internet-scale.

There are several challenges that the digital forensic community is facing.

Deep fakes are relatively new and have developed in sophistication much faster than expected. There are significantly more researchers working to develop techniques for synthesizing increasingly more realistic audio, images, and video, than there are those of us trying to detect this content. This means that the nature and quality of deep fakes is developing at an unprecedented rate that is difficult to keep pace with. In addition, the scale and speed of the internet makes deploying effective technology incredibly challenging: Facebook, for examples, sees some one billion daily uploads and YouTube sees some 500 hundred hours of video uploaded every minute. The sheer amount of information uploaded everyday makes any filtering technology incredibly difficult.

There is, however, a family of technologies that could be considered for wide deployment. Control-capture technologies can authenticate content at the point of recording by extracting, at the time of recording, a unique digital signature from any recorded digital content, cryptographically signing this signature, and then placing it on a secure central server or a distributed immutable ledger like the blockchain.¹ This signature can then be compared to any version of the same content found online to determine if the content has been altered from the time of recording. Although this approach tackles disinformation differently than forensic techniques – by telling us what is real instead of what is fake – these technologies are available today and can operate at internet-scale.

We should be exploring the further development and deployment of both control-capture and forensic technologies.

The Future

Despite the challenges, I propose several calls to action.

1. Funding agencies have to invest at least as much financial support to programs that seek to build systems to detect fake content as they do to programs in computer vision and computer graphics that are giving rise to the sophisticated synthesis technologies described above.
2. Researchers that are developing technologies that we now know can be weaponized should give more thought to how they can put proper

¹For full disclosure, I am a paid advisor to a company, Truepic, that develops control-capture technology.

safeguards in place so that their technologies are not misused.

3. No matter how quickly forensic technology advances, it will be useless without the collaboration of the giants of the technology sector. The major technology companies (including, Facebook, Google/YouTube, and Twitter) must more aggressively and proactively deploy technologies to combat disinformation campaigns, and more aggressively and consistently enforce their policies. For example, Facebook's terms of service state that users may not use their products to share anything that is "unlawful, misleading, discriminatory or fraudulent". This is a sensible policy — Facebook should enforce their rules.
4. Lastly, we should not ignore the non-technological component to the issue of disinformation: us the users. We need to educate the public on how to consume trusted information, we need to educate the public on how to be better digital citizens, and we need to educate the public on how not to fall victim to scams, fraud, and disinformation.

Conclusions

I will end where I began. Disinformation is not new. Deep fakes is only the latest incarnation. We should not lose sight of the fact that more traditional human-generated disinformation campaigns are still highly effective, and we will undoubtedly be contending with yet another technological innovation a few years from now. In responding to deep fakes, therefore, we should make every effort to consider the past, present and future as we try to navigate the complex interplay of technology, policy, regulation, and human nature.

Lastly, I would be remiss in not mentioning that although there are serious issues of on-line privacy, moves by some of the technology giants to transform their platform to an end-to-end encrypted system will only make the problem of disinformation worse. Such end-to-end encrypted systems will make it even more difficult to understand and slow or stop the spread of disinformation. We should make every effort to consider the balance between privacy and safety and how these can be best accomplished.

Committee on Science, Space, and Technology
Online Imposters and Disinformation

Hany Farid, Ph.D.

Biography

Hany Farid is a Professor at the University of California, Berkeley with a joint appointment in Electrical Engineering Computer Science and the School of Information. His research focuses on digital forensics, image analysis, and human perception. He received his undergraduate degree in Computer Science and Applied Mathematics from the University of Rochester in 1989, his M.S. in Computer Science from SUNY Albany, and his Ph.D. in Computer Science from the University of Pennsylvania in 1997. Following a two-year post-doctoral fellowship in Brain and Cognitive Sciences at MIT, he joined the faculty at Dartmouth College in 1999 where he remained until 2019. He is the recipient of an Alfred P. Sloan Fellowship, a John Simon Guggenheim Fellowship, and is a Fellow of the National Academy of Inventors.

Chairwoman SHERRILL. Thank you very much. Ms. Francois?

**TESTIMONY OF MS. CAMILLE FRANCOIS,
CHIEF INNOVATION OFFICER, GRAPHIKA**

Ms. FRANCOIS. Chairwoman Sherrill, and Ranking Member Lucas, Members of the Committee, thank you for having me here today. We're here to discuss the growing issue of online imposters and disinformation. As you know, this problem is nuanced and complex. I've been looking at disinformation campaigns for many years, and I have seen great diversity in the types of actors, techniques, and impacts that those disinformation campaigns can have. I want to highlight that, while we tend to focus on fake content, the most sophisticated actors I have seen operate online actually tend to use authentic content weaponized against their targets. This is what I want to talk about a little bit more.

It's really hard to give a sense of the growing and global scale of the issue, but here are a few recent examples. Today a report by my colleagues over at the Oxford Internet Institute highlighted that more than 70 countries currently use computational propaganda techniques to manipulate public opinion online. Since October 2018, Twitter has disclosed information around more than 25,000 accounts associated with information operations in 10 different countries.

Twitter is one thing. On Facebook, over 40 million users have followed pages that Facebook has taken down for being involved in what they call coordinated inauthentic behavior. Those may seem like huge numbers, but, in fact, they represent a needle in a haystack, and the danger of this particular needle is its sharpness. Targeting specific communities at the right time, and with the right tactics, can have a catastrophic impact on society, or on an election. That impact remains very difficult to rigorously quantify. For instance, if you take a fake account, what matters is not just the number of followers it has, but who those followers are, how they have engaged with the campaign, and how they have engaged both online and offline. Similarly, for a piece of content, it's not often the payload that matters, but really the delivery system, and the targeted system.

We are finding more and more state and non-state actors producing disinformation. What keeps me awake at night on this issue is also the booming market of disinformation for hire. That means troll farms that one can rent, bot networks that one can purchase, for instance. These tools are increasingly attractive to domestic political actors, who also use them to manipulate American audiences online. I see that you discovered how easy it was to make a deep fake, and I encourage you to also discover how easy it is to buy a set of fake accounts online, or, frankly, to purchase a full blown disinformation campaign.

The good news here, if there is any, is that, as a society, and as a professional field, we've come a long way since 2016. These problems began long before 2016, but it really took the major Russian interference in the U.S. election to force us toward a collective reckoning. In 2016 the top platforms, law enforcement, and democratic institution sleepwalked through the Russian assault on American democratic processes. Those who raised the alarm were, at best, ig-

nored. Today we're in a better place. We have rules, definition, and emerging processes to tackle these campaigns. Coordination between researchers, platforms, and public agencies have proven successful, for instance, in protecting the U.S. 2018 midterms from Russian disinformation efforts. Then, those actors worked hand in hand to detect, take down, and, to a certain extent, document the Russian attempts to deceive and manipulate voters.

We still have a long way to go, but the scale of the problem is staggering. Sophisticated state actors, and, again, a growing army of hired guns, are manipulating vast networks' interactions among billions of people on dozens of platforms, and in hundreds of countries. This manipulation is discoverable, but almost in a way that a submarine is discoverable under the ocean. What you really need is sophisticated sensors that must evolve as rapidly as the methods of evasion. That requires a serious investment in the development of analytic models, computational tools, and domain expertise on adversary trade crafts. We need better technology, but also more people able to develop and adopt rapidly evolving methods.

Accomplishing this also requires access to data, and that is currently the hardest conversation on this topic. The task at hand is to design a system that guarantees user security and privacy, while ensuring that the corps of scientists, researchers, and analysts can access the data they need to unlock the understanding of the threats, and harness innovative ways to tackle the issue. Today we're very far from having such a system in place. We critically need not just the data, but the community of scholars and practitioners to make sense of it. That emerging field of people dedicating to ensuring the integrity of online conversation needs support, funding, and a shared infrastructure.

[The prepared statement of Ms. Francois follows:]



**BRIEFING FOR THE UNITED STATES HOUSE OF REPRESENTATIVES COMMITTEE ON
SCIENCE, SPACE, AND TECHNOLOGY**

Investigations and Oversight Subcommittee Hearing on Online Imposters and Disinformation

Statement of Camille Francois, Chief Innovation Officer, Graphika, and Affiliate at the Berkman Klein Center for Internet & Society, and Ben Nimmo, Director of Investigations, Graphika

Washington, DC
September 26, 2019

The threat represented by the proliferation of information operations designed to deceive and manipulate users on social media demands a unified, forceful response by the whole of society.

The problem is nuanced and complex. There is enormous diversity in the types of actors, techniques, and impacts across the many campaigns our team has analyzed over the past few years.

To date, Facebook, Twitter, and Google alone have detected and taken down information operations emanating from at least 25 different countries, some of them designed to control domestic politics, others targeting geopolitical rivals. The countries named as points of origin for these operations, most of which were not directly attributed to state actors, include Russia, Iran, Bangladesh, Venezuela, Spain, China, Saudi Arabia, Ecuador, United Arab Emirates, Egypt, Myanmar, Iraq, Ukraine, Israel, Thailand, the Philippines, Honduras, India, Indonesia, Pakistan, the United Kingdom, Romania, Moldova, Macedonia, and Kosovo.

Some of these operations were directly coordinated by state actors. The best known are Russia and Iran, but China,¹ Honduras, and others also belong on the list.² Campaigns in Spain³ and India⁴ were linked to political parties; others were run by shadowy marketing companies, mercenary firms that execute influence operations on behalf of their clients, as witnessed in Israel,⁵ Egypt, and the United Arab Emirates.⁶ Some appeared linked to individual media outlets, as in the case of Kremlin-sponsored outlet Sputnik,⁷ small groups of activists, as in the United Kingdom,⁸ or even to specific individuals with political agendas, as in the Philippines during the recent senatorial election.⁹

¹ <https://newsroom.fb.com/news/2019/08/removing-cib-china/>

² <https://newsroom.fb.com/news/2019/07/removing-cib-thailand-russia-ukraine-honduras/>

³ <https://newsroom.fb.com/news/2019/09/removing-coordinated-inauthentic-behavior-in-spain/>

⁴ <https://newsroom.fb.com/news/2019/04/cib-and-spam-from-india-pakistan/>

⁵ <https://newsroom.fb.com/news/2019/05/removing-coordinated-inauthentic-behavior-from-israel/>

⁶ <https://newsroom.fb.com/news/2019/08/cib-uae-egypt-saudi-arabia/>

⁷ <https://newsroom.fb.com/news/2019/01/removing-cib-from-russia/>

⁸ <https://newsroom.fb.com/news/2019/03/removing-cib-uk-and-romania/>

⁹ <https://newsroom.fb.com/news/2019/03/cib-from-the-philippines/>

The global impact of deliberate manipulations of political conversations on social media is difficult to quantify, but a few data points can help us grasp orders of magnitude.¹⁰ Since October 2018, Twitter has published 25,084 accounts associated with information operations in ten different countries and has confirmed that tens of thousands more low-grade spam accounts were also involved in similar behavior.¹¹ The full archive of information operation¹² posts shared by Twitter encompasses over 65 million tweets, spanning more than five years of posting.

Over the past two years, Facebook has announced taking down 12,085 accounts, pages, groups, and Instagram accounts for engaging in what it calls “coordinated inauthentic behavior.” Just over 40 million other accounts followed one or more of these assets. The least-followed operation gathered under 1,000 followers, while the most followed, run by commercial companies in the United Arab Emirates and Egypt, gathered over 13 million.¹³ Reach measured in number of followers here is a very imperfect proxy for impact, though: *who* those followers are and *how* they are engaged often matter more than *how many* they are.

For the individual user, 50 million tweets or 40 million followers are almost too many to visualize, but for the platforms themselves, with hundreds of millions of active users, they represent only a fraction of daily activity. For the operators, meanwhile, what matters most is often a small group of deliberately chosen targets: a protest community, a politically influential group, or even an individual journalist who might unwittingly spread the desired narratives and alter their behavior based on anything from an artificially boosted trend to the release of hacked¹⁴ materials.¹⁵ The impact of these operations, from ruined reputations, to gaming the journalistic agenda, to election dynamics, are very real. The increase in information operations since 2016, and the range of actors carrying them out, should be ample evidence of the effectiveness of these methods.

It is critical to understand that these types of operations long predate 2016. Iran’s known operations targeted US audiences with fake social media profiles as early as 2013. Russia’s Internet Research Agency began attacking domestic opposition on Russian-language channels as far back as 2010 and further developed these methods in the 2014 Ukrainian conflict while ramping up US involvement in the same year. As early as 2012, the campaign of Mexican presidential challenger (and eventual president) Enrique Peña Nieto was accused of benefiting from large-scale amplification by Twitter bots (automated accounts), nicknamed Penabots. This problem has been with us for a while.

Unfortunately, it took until the Russian interference in the 2016 US election to force us toward a collective reckoning. In 2016, the major platforms, law enforcement, and democratic institutions sleepwalked through the Russian assault on US democratic processes, and those in the open-source community who raised the alarm were, at best, ignored. As just one example, Russian operators ran a Twitter account that claimed to be the unofficial outlet of the Republican Party in Tennessee and registered it to a Russian mobile phone number, yet the account survived three

¹⁰ These figures are based on the platforms’ public announcements, made intermittently through their blog posts. As a result, they represent voluntary disclosures on incidents that have been investigated to date, and are therefore not fully representative of the scale of the problem.

¹¹ https://about.twitter.com/en_us/values/elections-integrity.html#data

¹² For a searchable archive, see our efforts to make this data more available to the public on www.io-archive.org

¹³ <https://newsroom.fb.com/news/2019/08/cib-uae-egypt-saudi-arabia/>

¹⁴ On hack-to-leaks operations, see our work on *False Leaks* at CYBERWARCON 2018: <https://www.youtube.com/watch?v=P8iXN8j4gMk>

¹⁵ <https://www.nature.com/articles/d41586-019-02235-x>

complaints to Twitter from the actual Tennessee Republican Party.¹⁶ Even when disinformation became a national security issue for American democracy, we collectively failed to properly recognize and address it. It's fair to say that back then, most platforms were unaware of the scale or seriousness of this type of activity, did not have applicable rules against it, and weren't actively looking to protect their users from it.

We've come a long way since then. The main platforms, but also investigative journalists, government actors, and a network of skilled researchers are now actively looking for, investigating, and coordinating to take down influence operations across a wide range of online environments. They have begun working with external researchers, both to expose more operations and to explain what they have found. These green shoots are promising and should be commended. But unfortunately, we are not the only ones making progress.

There are now more actors perpetrating information operations, and primary adversaries are better resourced and more sophisticated every day. Facebook confirmed in July 2018 that the Internet Research Agency's operators were taking ever more effective steps to mask their presence, including using internet phone numbers to register accounts and proxy servers in third countries to mask their origins. They also paid third parties to run ads on their behalf.¹⁷ A separate Russian operation, exposed in June 2019 and suspected of being run by an intelligence service, went a step further by creating hundreds of blogs and social media accounts to post forged documents and divisive content and then abandoned most of the accounts after they had posted just once.¹⁸ A government-linked operation in China used large numbers of hijacked and repurposed accounts to spread its message.¹⁹ An operation emerging from Iraq used stolen official personal identification documents in an attempt to avoid systems in place to detect false accounts.²⁰ And as the mainstream platforms crack down on information operations, we also see operators invest in alternative and more marginalized platforms to run operations in more permissive environments. We tend to focus on the major platforms, but as their efforts to combat bad actors become more effective, the problem is migrating to smaller platforms that lack the capabilities and, sometimes, the will to fight back. We have documented one Russian operation alone that worked across more than 30 different platforms to spread false narratives.

So how do we tackle this issue together? Disinformation and information operations present a multi-faceted problem requiring technical, methodological, and policy solutions borrowed from disciplines as diverse as cybersecurity, data science, and consumer protection and privacy law. We need to understand all vectors critical to a disinformation campaign's impact: *Manipulative actors*, *deceptive behaviors*, and *harmful content*. These three vectors are what we call the "ABC's of disinformation."²¹ Content elements, like "deep fakes," get the most public attention, but their impact depends on more hidden but critical matters of *how* that content is being disseminated and *who* is hiding behind a campaign.

¹⁶ <https://www.buzzfeednews.com/article/kevincollier/twitter-was-warned-repeatedly-about-this-fake-account-run#tjENWBAQlv>

¹⁷ <https://newsroom.fb.com/news/2018/07/removing-bad-actors-on-facebook/>

¹⁸ <https://medium.com/dfrlab/top-takes-suspected-russian-intelligence-operation-39212367d2f0>

¹⁹ <https://www.aspi.org.au/report/tweeting-through-great-firewall>

²⁰ <https://newsroom.fb.com/news/2019/09/removing-coordinated-inauthentic-behavior-from-iraq-and-ukraine>

²¹ See attached paper: Francois, Camille. "Actors, Behaviors, Content: A Disinformation ABC. Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses." A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression: <https://www.ivir.nl/twg/>.

Science and technology have a crucial role to play in tackling this problem. The sheer volume of information on these platforms, and the speed with which it is shared, require new methods for campaign detection that can scale beyond our current capabilities. As our opponents become more effective at concealing their identities, we need to continuously innovate by creating forensic approaches that will be both accurate and difficult to undermine. And for us to make real, measurable progress on these fronts, we need to address the thorny but essential problem of data availability.

The task at hand is to design a system that guarantees user security and privacy while ensuring that academic researchers, cybersecurity professionals, and human rights investigators can access the data they need to unlock our understanding of these threats and harness innovative ways to tackle the issue. Today, we're very far from such a system.

Let me illustrate: The most well-understood campaign ever, for which the most data to date has been made available by all platforms, is the Russian campaign targeting the American public around the 2016 election. The trove of data released by US platforms and institutions has enabled superb academic work by our colleagues.²² The Graphika team, along with our colleagues at the Oxford Internet Institute, spent seven months investigating additional, non-public data on behalf of the Senate Intelligence Committee.²³

Our confidence in the completeness of this picture is false. There remain critical data blind spots. For instance, while platforms released a trove of data regarding the Internet Research Agency's public posts on social media, little to nothing has been shared regarding the GRU's campaigns, when in reality the GRU is the better funded and more persistent actor. It is also inherently more threatening, given its advanced hacking capabilities and readiness to leak apparently compromising material. Similarly, we know that Russian operators used private messaging to target and cultivate relationships with activists, campaign staffers, and journalists, but there is no data available anywhere to indicate a sense of scale and no public records to learn from to determine how best to immunize future targets against these types of threats. Finally, we know that the Russian operators designed their messages to be inflammatory and sometimes overtly hateful: how many of these posts have been moderated by platforms long before anyone cared about Russian trolls? Are we missing a large chunk of content from the public record? We believe that these data blind spots undermine our preparation for the threats ahead. We can, and must, do better.

Data availability is a major part of the solution. Another is ensuring that a community of scholars and practitioners exists to leverage it. At present, the study of these kinds of information operations on social media is a nascent discipline. We need help to turn it into a comprehensive and cooperative field that brings together experts who span the social and data sciences under a common framework, and with common goals.

²² See for instance: Stewart, Leo G., Ahmer Arif, and Kate Starbird. "Examining trolls and polarization with a retweet network." In Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web. 2018. ; Boatwright, Brandon C., Darren L. Linvill, and Patrick L. Warren. "Troll factories: The internet research agency and state-sponsored agenda building." Resource Centre on Media Freedom in Europe (2018) ; Benkler, Yochai, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018. (Chapter 8: *Are the Russians coming?*).

²³ Howard, Philip N., Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. *The IRA, social media and political polarization in the United States, 2012-2018*. University of Oxford, 2018.

In summary, the emerging field of interdisciplinary scholars and practitioners dedicated to ensuring the integrity of online conversations needs support, funding, and shared infrastructure that allow effective and collaborative innovation. We need this field to keep maturing and growing, to blossom into a community of ethnographers, historians, data scientists, cognitive psychologists, computer scientists, political scientists, sociologists, and many others. The diversity of this community will enable us to address our biggest challenges with a variety of informative perspectives. In this way, we will continue to collaboratively build upon a robust set of interdisciplinary methods, scientific rigor, and shared ethical principles.

Chairwoman SHERRILL. Thank you, Ms. Francois. We'll have to get to the rest of it as we go through the questions, but thank you very much. At this point we'll begin our first round of questions, and I'm going to recognize myself for 5 minutes.

I'd just like to start with Dr. Farid and Dr. Lyu, because we read a lot about the potential for deep fakes to be used on political candidates, and we watched Dr. Lyu's very compelling example here in this room, so thank you for that brilliant demonstration. I hope my fellow Members of Congress who aren't in the room today will actually get a chance to see for themselves, and hear just how limitless the potential impacts of deep fakes can be.

Let's talk about some hard truths. On a scale of 1 to 10, what do you think are the chances of a convincing video deep fake of a political candidate, someone running for Congress, or President, or Governor, emerging during the 2020 election cycle, and why do you think that?

Dr. FARID. I'm going to save five, to minimize my chances of being wrong. I am—and for another reason too, that I think we shouldn't—despite the sophistication of deep fakes, we shouldn't overlook that traditional disinformation works really well, and it's easy, right? Teenagers in Macedonia were responsible for a lot of the disinformation campaigns we saw in 2016. So I think it's coming. I don't know whether it'll be in 2020, or 2022, or 2024, but largely because the cheap stuff still works, and it's going to work for a while. I think we'll eventually get out ahead of that, and then this will be the new front.

But I think it is just a matter of time. We've already seen nefarious uses of deep fakes for cases of fraud, and I think the bigger threat here is not going to be—the first threat I predict is not going to be an actual deep fake, but the plausible deniability argument, that a real video will come out, and somebody will be able to say, that's a deep fake. And that, in some ways, is the larger threat that I see coming down the road, is once anything can be faked, nothing is real anymore. And I think that's probably more likely to happen before the first real deep fake comes out.

Chairwoman SHERRILL. That's interesting. Dr. Lyu?

Dr. LYU. Yes. Thank you for the question. As, actually, I mentioned in the opening remarks, the technical capability of making high-quality deep fakes is already at the disposal of whoever wants to make it. As I mentioned, for the deep fake videos we made, we have a specially made software, but anybody can potentially also develop similar softwares based on the open-source software on the GitHub, and then they can just buy a computer for about, you know, a couple thousand dollars, and then run this for a couple hours. Everything is automatic. So this is really the reality that, you know, people, whoever want to make these kind of videos, they have that capacity.

However, the question whether we will see such a video in a coming election really—as Professor Farid mentioned, depends on a lot of other factors, especially, you know, deep fake is not the only approach for disinformation. So it is kind of difficult to come up with a precise number there, but the possibility is certainly substantial. Thank you.

Chairwoman SHERRILL. Thank you. And then, Ms. Francois, you have a lot of experience observing how trolls and bots behave when they identify a hoax they might want to spread. If a convincing deep fake of a politico emerges next year, what do you expect the bot and troll participation to look like in proliferating the video? In other words, will we see this sort of erupt all at once, or does it percolate in dark areas of the Internet for a short-period of time before it emerges? How does that work?

Ms. FRANCOIS. All of the above are possible. I will say that, if we are facing a sophisticated actor able to do a convincing deep fake, they will be able to do a convincing false amplification campaign, too.

Chairwoman SHERRILL. Thank you very much. And then, Dr. Farid, you said in your testimony that researchers working on technologies to detect disinformation should give more thought to proper safeguards so their work cannot be misused or weaponized. What kind of safeguards do you believe could be adopted voluntarily by the research community to protect against the spread of disinformation?

Dr. FARID. Good. So I think there's two things that can be done. So, first, you have to understand in computer science we have an open source culture, which means we publish work, and we put it out there. That's been the culture, and it's wonderful. It's a wonderful culture. But when that technology can be weaponized, maybe we should think about putting the data, and the code, and the GitHub repository, where anybody can download it, as Professor Lyu was saying. So that's number one, is just think about how you disseminate. We can still publish and not put the details of it out so that anybody can grab it, No. 1.

No. 2 is, there are mechanisms by which we can incorporate, into synthetic media, watermarks that will make it easier for us to identify that. That can become a standard. We can say academic publishers who are going to post code should incorporate into the result of their technology a distinct watermark. That is not bullet-proof, it's not that it can't be attacked, but it's at least a first line of defense. So those are the two obvious things that I can see.

Chairwoman SHERRILL. That was perfect timing. Thank you very much, I appreciate it. I would now like to recognize Mr. Lucas for 5 minutes.

Mr. LUCAS. Thank you, Madam Chair. Dr. Farid, following up on what the Chair was discussing, in your written statement you say that no matter how quickly forensic technology for detecting deep fakes develops, it'll be useless without the cooperation of the technology giants like Google and Facebook. How do we bring those people to the table to begin this collaboration?

Dr. FARID. Yes. So the bad news is they have been slow to respond, for decades, really. It's not just disinformation. This is the latest, from child sexual abuse, to terrorism, to conspiracy theories, to illegal drugs, illegal weapons. The technology sector has been very slow to respond. That's the bad news. The good news is I think a combination of pressure from here on Capitol Hill, from Brussels, from the UK, and from the public, and from advertisers, there is now an acknowledgement that we have a problem, step number one.

Step number two is, what are we going to do about it? And I still think we are very slow here, and what you should understand is we are fighting against business interests, right? The business model of Facebook, Google, YouTube, Twitter is data, it's content. Taking down content is bad for business. And so we have to find mechanisms and either through regulatory pressure, advertising pressure, public pressure, bring them to the table. I will say the good news is, in the last 6 months, at least the language coming out of the technology sector is encouraging. I don't know that there's a lot of action yet.

So I will give you an example. We all saw a few months ago an altered video of Speaker Pelosi. This was not a confusing video, we all knew it was fake, and yet Facebook gleefully let it on their platform. In fact, defended the decision to leave it on their platform, saying, we are not the arbiters of truth, OK? So we have two problems now. We have a policy problem, and we have a technology problem. I can help with the technology problem. I don't know what I can do about the policy problem, when you say, we are not the arbiters of truth. So I think we have to have a serious look at how to put more pressure on the technology sector, whether that's regulatory, or legislative, or advertising, or public pressure, and they have to start getting serious as to how their platforms are being weaponized to great effect in disrupting elections, and inciting violence, and sowing civil unrest. I don't think they've quite come to grips with that reality.

Mr. LUCAS. Well, when that moment comes, and inevitably it will, in your opinion, what will that collaboration look like? There's a government element, there's an academic element, there's a public-private partnership element.

Dr. FARID. Yes.

Mr. LUCAS. Can you just—

Dr. FARID. Sure.

Mr. LUCAS [continuing]. Daydream for a moment here with me?

Dr. FARID. So I think the good news is the Facebooks and the Googles of the world have started to reach out to academics, myself included, Professor Lyu included. We now receive research funding to help them develop technology. That's good. I think the role of the government is to coax them along with regulatory pressure. I think what we've noticed over the last 20 years of self-regulation is not working. I'd like it to work, but it doesn't work in this particular space.

So I think the role of the government can be through oversight, it can be regulatory, it can be through a cyber ethics panel that is convened to talk about the serious issues of how technology is being weaponized in society. But very much I think the academic/industry model has to work, because most of the research that we are talking about is happening at the academic side of things, and obviously the industry has different incentives than we do in the academy, so I think there is room for everybody.

I'll also mention this is not bounded by U.S. borders. This is very much an international problem, so we should be looking across the pond to our friends in the UK, in the EU, and New Zealand, and Australia, and Canada, and bringing everybody on board because this is a problem for not just us, but for the whole world.

Mr. LUCAS. One last question. In your written testimony you suggest there's a non-technological component to solving the problem related to deep fakes and disinformation. Specifically, you wrote that we need to educate the public on how to consume trusted information, and how to be better digital citizens. What should this public education initiative—

Dr. FARID. Yes.

Mr. LUCAS [continuing]. Look like?

Dr. FARID. I'm always reluctant to say this, because I know how taxed our schools are in this country, but at some point this is an educational issue, starting from grade school on the way up. And, as an educator, I think this is our role. We have to have digital citizenry classes. Some of the European countries have done this. France is starting to do this, the UK is starting to do it. Public service announcements (PSAs) explaining to people how information can be trusted, what disinformation is, but we've got to start taking more seriously how we educate the next generation, and the current generation. And whether that's through the schools, through PSAs, through industry sponsored PSAs, you know, I think all of those are going to be needed.

Mr. LUCAS. And you would agree that our technology giant friends have a role in that education process?

Dr. FARID. They absolutely have a role. They made this mess, they need to help fix it.

Mr. LUCAS. Very concise. Thank you, Doctor. I yield back, Madam Chair.

Chairwoman SHERRILL. Thank you, Mr. Lucas. And now, Ms. Wexton, I recognize you for 5 minutes.

Ms. WEXTON. Thank you, Madam Chair, and thank you to the panelists for appearing today. I want to speak a little bit about the explosive growth that the major social platforms have experienced over the past few years, because I'm worried that these companies are more focused on growth, and getting more users, than they are about essential oversight and user support functions. And, in fact, as has been noted, they disclaim responsibility for any information that goes out onto the web by the users. And, in fact, it seems to me that they have a disincentive to purge suspicious, or fake, or bot accounts.

You know, I have here an article from July of last year, where Twitter's stock price went down by about eight and a half percent after they purged, over the course of two months, 70 million suspicious accounts. Now, don't feel too bad for Twitter, because their stock price went up 75 percent over that six month period, but, you know, by being socially responsible, or by being responsible, it hurt their bottom line.

Now, the platforms are incredibly powerful. We have already seen the power that they have here in the capitol, not just because of the lobbyists and everything, but because we all use them. We all have those platforms on our phones, and on our various devices. And, Dr. Farid, you spoke a little bit about how the basic features of the technology and the business model at social media companies kind of help exacerbate the proliferation of disinformation. Can you explain, from a business perspective, what benefit a bot account or a fake account might represent for a social media company?

Dr. FARID. Sure. So, first of all, I think you're absolutely right that growth has been priority No. 1. And because the metrics of Silicon Valley are number of users, number of minutes online, it's because that's what eventually leads to advertising dollars. What we have to understand is that Silicon Valley, for better or worse, today is driven by ad revenue, and ad revenue is optimized by having more engagement, OK? So fake account, real account, don't care. Fake like, real like, fake tweet, doesn't matter, right, because at the end of the day, you get to report big numbers to the advertisers who are going to pay more money. Whether 50 percent of those accounts are fake or not, who's to know?

So that's the underlying poison, if you will, of Silicon Valley, I think, and is the reason why the system is entirely frictionless, by design. There's no friction to creating an account on Twitter, or on Facebook, or on YouTube, because they want that to be easy. They want bots to be able to create these things because that is what elevates the numbers. And I think this is sort of our core problem that we have here.

Ms. WEXTON. So, related to that, why would social media companies allow, or even encourage, their recommendation algorithms to—

Dr. FARID. Good.

Ms. WEXTON [continuing]. Put people, you know, to direct users to—

Dr. FARID. Good.

Ms. WEXTON [continuing]. Suggested videos, or things like that, that are sensational, or even false? Why would they do that?

Dr. FARID. The metric on YouTube is engagement, how long do you stay on the platform? And so what the algorithms learn is that, if I show you a video that is conspiratorial, or outrageous, you are more likely to click on it and watch it. If you are more likely to click or watch, you're going to stay on the platform longer, right? So the algorithms are not trying to radicalize you. What they are trying to do is to keep you on the platform for as long as possible. And it turns out, in the same way that people will eat candy all day long instead of broccoli, people will watch crazy videos all day long instead of PBS. I don't think this is surprising. And so the underlying algorithms, what they are being optimized for, in part, is exactly this.

And we have been studying the nature of these conspiracy videos for over a year now, and I will tell you that, despite claims to the contrary, there is a rabbit-hole effect, that once you start watching the slightly crazy conspiratorial videos, you will get more and more and more of that because you are more likely to click, you are more likely to view, they're going to get more data, and they're going to sell more advertising. That's the underlying business model, is how long do you stay on my platform? And, in that regard, the quality of the information is utterly unimportant to the platforms. It is what keeps you there.

Ms. WEXTON. So maybe we should all have more cats and kittens, and less conspiracy?

Dr. FARID. I'm all for cat videos.

Ms. WEXTON. So, switching gears a little bit, yesterday this Committee—we marked up a bill, it was Congressman Gonzalez's bill,

that would expand research into technologies to help us better identify deep fake videos. And I had an amendment which was made in order, and approved by the Committee, to help increase education to help people identify deep fake videos, and so I was encouraged to hear you talk about that. So I would inquire of the panel, do you have any advice on what the most important elements of a public education campaign on deep fake videos should be?

Dr. FARID. Again, you know, I am reluctant to put this on our public schools. I think they are overtaxed, and overworked, and underfunded. But at the end of the day, this is sort of where it belongs. And I think if we can do this, not as an unfunded mandate, but actually give them the resources to create courses of digital citizenry, of how you are a better digital citizen, how you can trust information and not trust information.

I'll point out too, though, by the way, it's not just the young people. The senior citizens among us are more likely to share fake news than the young people, so this is across the spectrum. So I'm more—this—for me, the education level is more about the next 20, 30, 40 years than necessarily today. So I think a combination of PSAs, about returning to trusted sources, and about educating kids not just, by the way, about trusted information, but how to be a better digital citizen, how to interact with each other. The vitriol that we see online is absolutely horrific, and the things that we accept online we would never accept in a room like this, and I think we have to start teaching the next generation that this is not a way that we interact with each other. We need a more civil discourse.

Chairwoman SHERRILL. Thank you, Dr. Farid. And I'd now like to recognize Mr. Biggs for 5 minutes.

Mr. BIGGS. Thank you, Madam Chair, and I appreciate each of the witnesses for being here. It's a very, very interesting hearing, and appreciate the Chair for convening this hearing.

So one of the main things I'm worried about is the de facto gray area between misinformation and disinformation, despite the seemingly clear definitional difference between these concepts. While disinformation may be defined in terms of the malicious intent on the part of the sender, such intent, as we've seen today, can at times be very difficult to identify. And then, on top of that, we need to make sure the gatekeepers, themselves trying to police content, are objective. Objective enough to identify potential misinformation, and able to do so as expeditiously as possible.

It seems to me that, even if we have the technological anti-disinformation tools that we've learned about in this discussion, and that we anticipate seeing developed over time, human judgment will always be a key component of any anti-deep fakes effort, and human judgment can, of course, be fallible. In short, the difficulties and nuances of the battle pile up the deeper we delve into this topic. Maybe that's why I find it so interesting to hear what you all have to say today.

But I want to just get back to something, and I would say I feel like we've been doing what I would call an endogenous look, and that is what's the technology here? And you mentioned it, Dr. Farid, in item four on page four of your recommendations in your written report, but it really gets to what I think is a real-world

problem I'd like all of you to respond to, and the last questioner just kind of touched on it a bit as well. What do you tell a 13- or 14-year-old that you're trying to warn of potential disinformation, misinformation? How do you do it as a parent, as a grandparent, as someone who cares for, loves, an individual. I mean, that really becomes a part of the equation as much as anything else on the technological side.

Dr. LYU. Well, thank you for asking the question, because the nature of my work, I usually show a lot of fake videos to my 12-year-old daughter, and she actually grow the habit of distrust for any video I showed to her. So I think this may be a very effective way to actually tell them—to show them that the existence of fake videos will make them aware that those are something they should be careful about.

Ms. FRANCOIS. I can take the question on, you know, what goes beyond technology, and I want to talk about one specific example. I think, when you look at the most sophisticated campaigns that have leveraged disinformation, and we're talking about actors who are willingly doing this, there's still a lot that we don't know. So, back to the Russian example, for instance, which is largely seen as the best-documented campaign, right, on which the platforms have shared a lot of data. I have myself worked with the Senate Select Intelligence Committee to document what happened. There are still essential pieces of that campaign that we know nothing about, and on which there's no data, in the eye of the public, to really understand how that technology was leveraged to manipulate audiences, direct messages, and how the Russians used to target deliberately specific journalists to feed them sources. We don't know anything about the scale of how much of that was going on.

Similarly, what the GRU was doing, alongside the IRA, is something that there's zero available data on. So I would go back to those important and large-scale campaigns that we know have really disrupted society and interrogate, where are our blind spots? How can we do better? How can we produce this data so that we actually are able to fully understand those tactics? And then, of course, to build the tools to detect it, but also to train people to understand it, and to build defense.

Mr. BIGGS. Thank you. Dr. Farid? What are you going to tell your kid?

Dr. FARID. I, fortunately, don't have kids, so I don't have to struggle with this problem.

Mr. BIGGS. They're a blessing and a curse.

Dr. FARID. I think this is difficult, because the fact is this generation is growing up on social media—

Mr. BIGGS. Yes.

Dr. FARID [continuing]. And they are not reading *The Washington Post*, and *The New York Times*, and MSNBC, and Fox News. They think about information very differently. And I can tell you what I tell my students, which is, do not confuse information with knowledge. Those are very different things. And I think there is this tendency that it's online, therefore it must be true. And so my job as an educator is to make you critically think about what you are reading. And I don't know how to do that on a sort of day-to-day basis, but I do that every day with my students, which is crit-

ical reasoning. And with critical reasoning, I think everything comes.

And, if I may, I wanted to touch one—because I think you made a good point about the—sort of the nuance between mis- and disinformation, and we should acknowledge that there are going to be difficult calls. There is going to be content online that falls into this gray area that it's not clear what it is, but there is black and white things out there, and we should start dealing with that right now, and then we'll deal with that gray area when we need to, but let's not get confounded with that gray area, and not deal with the real clear cut harmful content.

Mr. BIGGS. Right. So information's not knowledge. I'd like to tell people in Congress, activity is not progress either, so, I mean, we—

Dr. FARID. We agree on that.

Chairwoman SHERRILL. Thank you, Mr. Biggs. And next I would like to recognize Mr. Beyer for 5 minutes.

Mr. BEYER. Madam Chair, thank you very much. Dr.—Ms. Francois—so Dr. Lyu talked about funding more civilian research through the National Science Foundation, and setting up an emerging technologies directorate, and you spoke about this emerging field of interdisciplinary scholars, practitioners, that needed support, funding, and shared infrastructure. How best do you see us making that happen? Do we need congressional legislation? How big a budget does it have to be? Is it only NSF, or NIST (National Institutes of Standards and Technology), or—

Ms. FRANCOIS. That's a great question, thank you. I think it can be a whole of government effort, and I do think that a series of institutions have to get involved, because indeed, as I say, it's very interdisciplinary. I do think that regulation has to play a role too, not only to address those critical and complex questions, like the one of data access that I discussed.

I want to build on a point that Dr. Farid made about the algorithmic reinforcement, as an example. This is something that we know is impacting society. People watch one video, and seem to end up in a filter bubble of conspirational video. But, unfortunately, we have very little serious research on the matter. We are making those observations on a purely empirical basis out of, you know, people who let their computers run. We can't afford to be in the dark on the impact of technology on society like this. And in order to do serious scientific research on those impacts at scale, we need data, and we need the infrastructure to systematically measure and assess how this technology is impacting our society.

Mr. BEYER. Thank you very much. Dr. Farid, I was fascinated you talked about determining what's real, rather than what's fake, and specifically talking about the control capture technologies. We've had a number of Science Committee hearings on blockchain technology, which inevitably lead into quantum computing (QC) technology. Is blockchain, and ultimately QC, the right way to deal with this?

Dr. FARID. I think blockchain can play a role here. So the basic idea, for those who don't know, blockchain—basically all you have to know is that it's an immutable distributed ledger. So immutable, when you put information on there, it doesn't change. Distributed

as it's not stored on one central server, but on millions of computers, so you don't have to rely on trust of one individual.

So one version of control capture is, at the point of capture, you extract that unique signature, cryptographically sign it, and you put that signature on the blockchain for public viewing of it, and public access to it. It's a very nice application of blockchain. I don't think it's critical to the solution. If you have a trusted central server, I think that would work well, but the reason why people like the blockchain is that I don't have to trust a Facebook, or an Apple, or a Microsoft, I can trust the globe. So I do see that as being part of the control capture environment, and being part of the solution of a universal standard that says, if you want your content to be trusted, take it with this control capture, and then we can trust that going down the line. I think we're eventually going to get there. I think it's just a matter of time.

Mr. BEYER. And, Dr. Lyu, how would you contrast watermarking technology with the blockchain, with the control capture? And is one better than the other, or do you need both, or—

Dr. LYU. I think these technologies are somehow complementary. So watermark is the content you actually embed into the image, and blockchains are ways to actually authenticate if the watermark is consistent with the original contents we invited into the signal. So they can work together. You can imagine that we have watermark also being part of the blockchain, uploaded to the remote distributed server. So they can work hand in hand in this case. But watermarks can also work independently from a single capture control mechanism for authenticity of digital visual media.

Mr. BEYER. Thank you. And Ms.—

Dr. LYU. Thank you.

Mr. BEYER. Ms. Francois, again, you talked about how the big data players, the Facebooks and Twitters, obviously are a huge part of the potential problem—source material, and have to be part of the solution, and you mentioned regulation as one of the pieces of the NSF/NIST piece. Not that you can do it in 45 seconds, but anything that you guys can prepare to help our Energy and Commerce Committee, the committees in both houses, looking at how we manage the social media giants would be very, very appreciated. Because understanding how they've gone from basically unregulated unicorn game changers in our society, to how they can properly play within the rules, is going to be a really, really big challenge for us.

Ms. FRANCOIS. I think it's going to be a lot of moving pieces. It's a complex problem, as I said, and I do believe that there's a lot of different bodies of regulation that can be applied and brought to bear to tackle it. One that is often left out of the conversation that I just want to highlight here is consumer protection. Dr. Farid talked about how the advertisers are getting the fake clicks. This can be a consumer protection issue. So different bodies of regulation, from cyber security to consumer protection, to address a whole of the disinformation problem, plus serious pressure to ensure that the data that the field needs is being shared in a way that makes it—for people.

Mr. BEYER. Yes. Thank you very much, and I yield back.

Chairwoman SHERRILL. Thank you. Next I'd recognize Mr. Waltz for 5 minutes.

Mr. WALTZ. Thank you, Madam Chairwoman. Ms. Francois, going back to the disinformation campaigns that the Russians, the Iranians, and others have ongoing, the FBI and Department of Homeland Security have briefed us that they're confident, at least at this point in time, that active hacking into our election infrastructure has diminished, at least for now.

Although I, and other colleagues, have worked to ensure that critical infrastructure is secured going forward, and this Committee has done work on that as well, but I'm interested in the disinformation piece of it, are you seeing increasing evidence of our adversaries conducting disinformation against individuals, whether they're thought leaders, journalists, politicians? For example, I could foresee hawks on Iran policy, or Russia, or others being specifically targeted during an election in order to change that election outcome, and therefore change our policy and voices. Are you seeing an increase there? What types of techniques are you seeing, and where are you seeing it, aside from the United States?

One of the things that I've pushed is for us to share what we're gathering. For example, the Taiwanese elections, or other elections, for us to create a collaborative approach with our allies as well. This is a problem with the West, and I think with free speech and free thought, as much as it is with, you know, 2020 elections. And I'd welcome your thought.

And then second, sorry, what would you think the response would be if we took more of a deterrence measure? For example, sending the signal that the Iranians, the Russians, and other bad actors, they have their own processes, and they have their own concerns, and often these regimes are more concerned with their own survival than they are with anything else, and at least demonstrating that we have that capability to interfere as well. I know that may present a lot of moral and ethical questions of whether we should have that capability, and whether we should demonstrate we should use it, but we've certainly taken that approach with nuclear weapons. And so I'd welcome your thoughts there.

Ms. FRANCOIS. Thank you. I want to start by saying that part of it—yes, I am seeing an increase. Part of it is an increase, the other part is simply just a reckoning, as I said. Iran is a good example. We see a lot of disinformation campaigns originating from the Iranian state, who's a very prolific actor in that space.

Now, people often ask me, is Iran following the Russian model. In reality the first Iranian campaign to use social media to target U.S. audiences date back from 2013, where we were asleep at the wheel, and not looking for them. So, despite our reckoning with sort of the diversity of actors who have been engaged with these techniques to target us, there is also an increase in both their scale and their sophistication. This is a cat-and-mouse game, and so what we also see is, as we detect actors and their techniques, they increase the sophistication. They make it harder for us to do the forensics that we need in order to catch those campaigns as they unfold.

Thank you for raising the question of deterrence. I do think that this ultimately is a cyber policy issue too, and therefore the govern-

ment has a role to play. In the case of the U.S. midterms in 2018, we saw U.S. Cyber Command target the Internet Research Agency in St. Petersburg in an act of this attempted cyber deterrence. So I do think that there is a governmental response too by putting this problem in the broader context of cyber issue and cyber conflict.

Mr. WALTZ. Thank you for raising that. I think it's important for my colleagues to note that was a policy change under this Administration that then allowed Cyber Command to take those kind of, what they call active defensive measures, and taking election security very seriously. I want to distinguish, though, between active defense and the potential, at least, and sending the signal that we have the potential for offense. And your thoughts there on the United States also participating in disinformation, or at least a deterrent capability?

At the end of the day I think we can only do so much in playing defense here. We can only counter so much of this cat-and-mouse game. We have to fundamentally change our adversaries' behavior, and put them at risk, and their regimes at risk, in my own view. But I'd welcome your thoughts in my remaining time.

Ms. FRANCOIS. Yes, I think the—8 minutes to answer this complex question on the dynamics of deterrence and resilience in cyberspace. I will say what immediately comes to mind is, of course, a question of escalation. How much of these dynamics contribute to escalation is something that is an unknown in this space.

So far I think that the approach of being much more aggressive in both catching these campaigns, deactivating them, and publicly claiming that we have found them, and this is what they look like, seems to be a welcome move in this area. I think by exposing what actors are doing, we are also contributing to raising the cost for them to engaging in these techniques.

Chairwoman SHERRILL. Well, that was well done——

Mr. WALTZ. Thank you.

Chairwoman SHERRILL [continuing]. Ms. Francois. Thank you. Next I recognize Mr. Gonzalez for 5 minutes.

Mr. GONZALEZ. Thank you, Madam Chair, and thank you for being here, to our witnesses, and your work on this topic. A very important topic, and one that's a little bit new to Congress, but one that, alongside of Madam Chair, and others on this Committee, we've been excited to lead on, and I think we're making progress, unlike some other areas of Congress that I'm a part of.

So, that being said, Dr. Lyu, I want to start with you, and I really just want to understand kind of where we are in the technology, from the standpoint of cost. So if, call it 2 decades ago, I used the Forrest Gump example, yesterday. You know, Forrest Gump, if you've seen the movie, makes it look like he's shaking hands with Presidents, and all kinds of things, and you can't tell the difference, except for you just know that there's no way that happened. Hollywood studio could've produced that, but it was costly back then, right, however much it costs. Today I think some numbers came out that you were citing that as, you know, roughly a couple thousand dollars. How quickly is the cost going down, to the point that this will be a weapon, if you will, that, you know, a 16-year-old sitting behind his computer could pull off?

Dr. LYU. I think this is basically, you know, we used to call this Moore's Law, where the computational power just got doubled every 18 months, and I think Moore's Law has already been broken with the coming of GPUs. The computational power that are at our hand is extremely higher than we have imagined before, and this trend is growing. So I will predict in the coming years it will become cheaper and easier, and also better to produce these kind of videos, and the computer hardware and algorithms will all get rapid improvements.

Mr. GONZALEZ. Yes.

Dr. LYU. So that's coming. I think it's a coming event. Thank you.

Mr. GONZALEZ. Thanks. And I actually think, you know, we talk a lot about great power competition in Iran, and China, and Russia, and I think that makes sense. I'm also maybe equally concerned about just a random person somewhere in society who has access to this, and can produce these videos without any problem, and the damage that that can cause. And I don't know that we've talked enough about that, frankly.

But switching to Ms. Francois, you talked about how you found 70 countries use computational propaganda techniques in your analysis. And obviously a lot of this is spread through the platforms, and I think you talked really well about just how you can go down rabbit holes in the engagement metrics, and things like that. What do you think, and Dr. Farid, I'd welcome your comments as well, what do the platforms themselves need to be doing differently? Because it strikes me that they're being somewhat, or I would say, I would argue grossly irresponsible with how they manage some of the content on their systems today.

Ms. FRANCOIS. That's a great question. I just want it precise that the 70 countries method comes from the Oxford Internet Institute report that was published today.

Mr. GONZALEZ. OK. Thank you.

Ms. FRANCOIS. For me, the platform's play here is actually quite simple, and I would say clearer roles, more aggressive action, more transparency.

Mr. GONZALEZ. Yes.

Ms. FRANCOIS. Let's start with clearer roles. Some platforms still don't have a rule that governments are not allowed to leverage their services in order to manipulate and deceive. And they will say they have rules that kind of go to this point, you know, tangentially, but there's still a lot of more clear rules that need to be established. To the second point, aggressive enforcement. There's still a lot of these campaigns that go under the radar, and that go undetected. They need to put the means to the table to make sure that they actually are able to catch, and detect, and take down as much of this activity as possible. My team, this week, published a large report on a spam campaign that was targeting Hong Kong protestors from Chinese accounts, and then they—

Mr. GONZALEZ. Yes.

Ms. FRANCOIS [continuing]. Had to take it down. There's more that they can do. Finally, transparency. It's very important that the platform continue, and increase, their degree of transparency

in saying what they're seeing on their services, what they're taking down, and share the data back to the field.

Mr. GONZALEZ. Yes. I think that makes a lot of sense. My fear is, you know, we're going to do the best we can. I don't know that, one, this is intellectually difficult to figure out, as Congress, and it's also politically difficult, which, to me, puts it in that, like, Never Never Land, if it's going to take a while. So my hope is that the social medial platforms understand their responsibility, and come to the forefront with exactly what you said, because if not, I don't know that we're going to get it right, frankly.

But with my final question, I'll throw just the word mental health, and the platforms themselves, and misinformation. Any studies that you're aware of that are showing the impacts on mental health, in particular teenagers, with respect to what's going on on the platforms today? Anybody can answer that.

Ms. FRANCOIS. Again, I want to say that in this field we direly lack the data, infrastructure, and access to be able to do robust at-scale studies. So there is a variety of wonderful studies that are doing their best with small and more qualitative approaches, which are really, really important, but we're still direly lacking an important piece of doing rigorous research in this area.

Mr. GONZALEZ. Thank you. And I'll follow up with additional questions on how we can get that data, and be smarter about that in Congress. So, thank you, I yield back.

Mr. BEYER [presiding]. Thank you very much, sir. Dr. Farid, I understand you developed a seminal tool for Microsoft called PhotoDNA that detects and weeds out child pornography as it's posted online. Can you talk about how this tool works? Could this be used to address harmful memes and doctored images? And how do the social media companies respond to this?

Dr. FARID. So PhotoDNA was a technology that I developed in 2008–2009 in collaboration with Microsoft and the National Center for Missing and Exploited Children (NCMEC). Its goal was to find and remove the most horrific child sexual abuse material (CSAM) online. The basic idea is that the technology reaches into an image, extracts a robust digital signature that will allow us to identify that same piece of material when it is reuploaded. NCMEC is currently home to 80 million known child sexual abuse material, and so we can stop the proliferation and redistribution of that content.

Last year alone, in one year, the National Center for Missing and Exploited Children's CyberTipline received 18 million reports of CSAM being distributed online. That's 2,000 an hour. 97, 98 percent of that material was found with PhotoDNA. It has been used for over a decade, and has been highly effective. Two more things. That same core technology can be used, for example, to find the Christchurch video, the Speaker Pelosi video, the memes that are known to be viral and dangerous. Once content is detected, the signature can be extracted, and we can stop the redistribution.

And to your question of how the technology companies respond, I think the answer is not well. They were asked in 2003 to do something about the global distribution of child sexual abuse material, and for 5 years they stalled, they did absolutely nothing. We're not talking about complicated issues here, gray areas. We are talking about 4-year-olds, 6-year-olds, 8-year-olds being violently raped,

and the images and the videos of them, through these horrific acts, being distributed online. And the moral compass of Silicon Valley for the last decade has been so fundamentally broken they couldn't wrap their heads around their responsibility to do something about that.

That doesn't bode well, by the way, for going forward, so I think that history is really important, and we have to remember that they come begrudgingly to these issues, and so we have to coax them along the way.

Mr. BEYER. Thank you very much. So there—these images have digital signatures, even before we talk about the capture control technology—

Dr. FARID. Yes.

Mr. BEYER [continuing]. Or the watermark—

Dr. FARID. That's exactly right. These don't have to be captured with specific hardware. So what we do is, after the point of recording, we reach in and we find a distinct signature that will allow us to identify, with extremely high reliability, that same piece of content. And that can be child abuse material, it can be a bomb-making video, it can be a conspiracy video, it can be copyright infringement material. It can be anything.

Mr. BEYER. But it has to show up first—

Dr. FARID. That's right.

Mr. BEYER [continuing]. In the public space—

Dr. FARID. Yes.

Mr. BEYER [continuing]. At least once, and we have to know that it's there in order to capture this—

Dr. FARID. That's the drawback. But the good news is that technology works at scale. It works at the scale of a billion uploads to Facebook a day, and 500 hours of YouTube videos a minute. And that's a really hard engineering problem to tackle, but this technology actually works, unlike many of the other algorithms that have extremely high error rates, and would simply have too many mistakes.

Mr. BEYER. Thank you very much. Dr. Lyu, you talked about using AI to find AI, and that more deep neural networks are used to detect the fakes, but there's the sense that the good guys are always trying to catch up with the bad guys, you know, the cat-and-mouse. Is there any way around the cat-and-mouse nature of the problem? Which, by the way, we just saw before, it's got to be out there before you can tag it and chase it down.

Dr. LYU. That's a very good question. Actually, I think on this point, I'm more pessimistic because I don't think there's a way we can escape that, because that's the very nature of this kind of problem. Unlike other research areas, where the problem's fixed, we're basically dealing with a moving target. Whenever we have new detection or deterrent algorithms, the adversaries will always try to improve their algorithm to beat us. So I think, in the long run, this will be the situation that will keep going.

But I—that also emphasize Dr. Farid's point that we need more investment onto the side of detection and protection for the sake that, you know, we have a lot more resources put into making deep fakes for, you know, all kinds of reasons, but the investment in detection has not been catching up with that level. So that's part of

my testimony, is encouraging the Federal Government to put more investment into this important area. Thank you.

Mr. BEYER. Ms. Francois?

Ms. FRANCOIS. Yes, if I may add a very simple metaphor here, I think we also have a leveling of the playing field issue. We're currently in a situation where there are a lot of cats, and very few mice. We need to bring the resources to the table that correspond to the actual scale and gravity of the problem.

Mr. BEYER. OK. Great. Thank you very much. I now recognize the gentleman from Ohio, Mr. Gonzalez.

Mr. GONZALEZ. Thanks. Didn't know I was going to get a few extra seconds. So I just want to drill down on that data-sharing component. So you mentioned that we just need a better data-sharing infrastructure. Can you just take me as deep as you can on that? What do we need specifically? Just help me understand that.

Ms. FRANCOIS. Yes. There are many different aspects to what we need, and I think that the—both the infrastructure, people involved, and type of data depend on the type of usage. So, for instance, facilitating academic access to at-scale data on the effects of technology on society is ultimately a different issue than ensuring that cybersecurity professionals have access to the types of forensics that correspond to a high-scale manipulation campaign that enables them to build better detection tools. And so I think the first step in tackling this problem is recognizing the different aspects of it.

Mr. GONZALEZ. Got it.

Ms. FRANCOIS. Of course, the key component here is security and privacy, which here go hand in hand. What you don't want is to enable scenarios like Cambridge Analytica, where data abuses lead to more manipulation. Similarly, when we see disinformation campaigns, we often see a lot of real citizens who are caught into these nets, and they deserve the protection of their privacy.

If you go down sort of the first rabbit hole of ensuring that cybersecurity professionals have access to the type of data and associated forensics that they need in order to do this type of detection at scale, and to build the forensics tool we need at scale, there's still, as I said, a lot we can do. The platforms right now are sharing some of the data that they have on these types of campaigns, but in a completely haphazard way. So they're free to decide when they want to share, what they want to share, and in which format. Often the format, they're sharing them in are very inaccessible, so my team has worked to create a database that makes that accessible to researchers. That's one step we can take.

And, again, and I'll wrap on that, because this can be a deep rabbit hole—

Mr. GONZALEZ. Yes.

Ms. FRANCOIS [continuing]. You pushed me down this way. Again, if we take the Russia example, for instance when we scope a collection around something that we consider to be of national security importance, we need to make sure we have the means to ensure that the picture we're looking at is comprehensive.

Mr. GONZALEZ. Right.

Ms. FRANCOIS. Our own false sense of security, in looking at the data, thinking that they represent the comprehensive picture of

what happened, and was directed at us, is a problem in our preparations for election security.

Mr. GONZALEZ. Thank you. Dr. Farid, any additional thoughts on that?

Dr. FARID. Yes. I just wanted to mention, and I think Ms. Francois mentioned this, there is this tension between privacy and security, and you're seeing this particularly with Cambridge Analytica. And I will mention too that this is not, again, just a U.S. issue, this is a global issue. And with things like GDPR (General Data Protection Regulation), it has made data sharing extremely more complex for the technology sector.

Mr. GONZALEZ. Yes.

Dr. FARID. So, for example, we've been trying to work with the sector to build tools to find child predators online, and the thing we keep running up against is we can't share this stuff because of GDPR, we can't share it because of privacy. I think that's a little bit of a false choice, but there is a sensitivity there that we should be aware of.

Mr. GONZALEZ. Yes. That's fair. I agree with you. Certainly, I think what you highlight, which I agree with, is there are gray areas—

Dr. FARID. Yes.

Mr. GONZALEZ [continuing]. OK, but there also, like, big bright lines. Child pornography, let's get that off our platforms.

Dr. FARID. Yes, I agree. And feels to me like, if you share child pornography, you have lost the right to privacy. I don't think you have a right to privacy anymore once you've done that, I should have access to your account. So I think there's a little bit of a false narrative coming out here, but I still want to recognize that there are some sensitivities, particularly with the international standards. The Germans have very specific rules—

Mr. GONZALEZ. Yes.

Dr. FARID [continuing]. The Brits, the EU, et cetera.

Mr. GONZALEZ. So the last question, and this is maybe a bit of an oddball, so with the HN site that was ultimately brought down, I believe Cloudflare was their host, is that—

Dr. FARID. Yes.

Mr. GONZALEZ. So we talk a lot about the platforms themselves, right, but we don't always talk about the underlying infrastructure—

Dr. FARID. Yes.

Mr. GONZALEZ [continuing]. And maybe what responsibilities they have.

Dr. FARID. Yes.

Mr. GONZALEZ. Any thoughts on that? Should we be looking there as well?

Dr. FARID. You should. And it is complicated, because—

Mr. GONZALEZ. Yes.

Dr. FARID [continuing]. When you go to a Cloudflare—as the CEO came out and said, I woke up 1 day, and I thought, I don't like these guys, and I'm going to kick them off my platform. That is dangerous.

Mr. GONZALEZ. That's very—

Dr. FARID. Yes. But Ms. Francois said it very well. Clear rules, enforce the rules, transparency. We have to have due process. So define the rules, enforce them consistently, and tell me what you're doing. I can fix this problem for the CEO of Cloudflare. Just tell me what the rules are. So—but I don't think they get a bye just because they're the underlying hardware of the Internet. I think they should be held to exactly the same standards, and they should be held to exactly the same standards of defining, enforcing, and transparency.

And, by the way, I'll also add that cloud services are going to be extremely difficult. So, for example, we've made progress with YouTube on eliminating terror content, but now they're just moving to Google Drive, and Google is saying, well, Google Drive is a cloud service, so it's outside of this platform. So I do think we have to start looking at those core infrastructures.

Mr. GONZALEZ. OK. I appreciate your perspective. Frankly, I don't know what I net out on it, I just know it's something that I think we should be looking at—

Dr. FARID. I agree.

Mr. GONZALEZ [continuing]. And weighing, so thank you.

Mr. BEYER. Thank you. Dr. Lyu, you know, Ms. Francois just talked about a level playing field, you know, that, the bad guys have a lot more tools and resources than the good guys.

Dr. LYU. Right.

Mr. BEYER. We talked a lot about the perils of deep fakes, but are there any constructive applications?

Dr. LYU. Actually—

Mr. BEYER [continuing]. Where we want to use deep fakes in a good way?

Dr. LYU. Yes, indeed. Actually, the technology behind deep fake, as I mentioned in my opening remark, is of dual use. So there's a beneficial side of using this technology. For instance, the movie industry can use that, reduce their costs. There are also ways to actually make sure a message can be broadcast to multilingual groups without, you know, regenerating the media in different languages. It is also possible to use this technology to protect privacy. For instance, for people like whistleblowers, or, you know, victims in violent crime. If they don't want to expose their identity, it's possible to use this technology, replacing the face, but leaving the facial expression intact there.

The negative effect of deep fake, this kind of technology, you get a lot of spotlight, but there's also this dual use that we should also be aware of. Thank you very much.

Mr. BEYER. Thank you. Ms. Francois, are there any good bots?

Ms. FRANCOIS. Yes. They're really fun. One of them systematically tweets out every edit to Wikipedia that is made from the Congress Internet infrastructure. In general what I'm trying to say is there are good bots. Some of them are fun and creative, and I think they do serve the public interest. I do not think that there are good reasons to use an army of bots in order to do coordinated amplification of content. I think when you are trying to manipulate the behavior to make it look like a broader number of people are in support of your content than actually is the case, I do not see any particularly good use of that.

Mr. BEYER. I want to send you one of my daughter's bots. She has a perfectly normal Twitter account, and then she has the Twitter bot account, where she leverages off of her linguistics background, and I cannot make heads nor tails of what it does. But perhaps—

Ms. FRANCOIS [continuing]. Can look at it.

Mr. BEYER [continuing]. You can. Yes, it's—

Ms. FRANCOIS. OK.

Mr. BEYER. She says it's OK. Dr. Farid, you talked—it would be a mistake for the tech giants to transform their system from end-to-end encrypted systems, that would make the problem only worse. Can you walk us through that?

Dr. FARID. Sure, and I'm glad you asked the question. So let's talk about what end-to-end encryption is. So the idea is I type a message on my phone, it gets encrypted, and sent over the wire. Even if it's a Facebook service, Facebook cannot read the message. Under a lawful warrant, you cannot read the message. Nobody can read the message until the receiver receives it, and then they decrypt. So that's called an end-to-end encryption. Everything in the middle is completely invisible. WhatsApp, for example, owned by Facebook, is end-to-end encrypted, and it is why, by the way, WhatsApp has been implicated in horrific violence in Sri Lanka, in the Philippines, in Myanmar, in India. It has been linked with election tampering in Brazil, in India, and other parts of the world, because nobody knows what's going on on the platform.

So last year, you heard me say, 18 million reports to the National Center for Child Sexual Abuse Material, more than half of those came from Facebook Messenger, currently unencrypted. If they encrypt, guess what happens? Ten million images of child sexual abuse material, I can no longer see. This is a false pitting of privacy over security, and it's completely unnecessary. We can run PhotoDNA, the technology that I described earlier, on the client so that, when you type the message and attach an image, we can extract that signature. That signature is privacy preserving, so even if I hand it to you, you won't be able to reconstruct the image, and I can send that hash, that signature, along with the encrypted message, over wire, pull the hash off, compare it to a database, and then stop the transmission.

And I will mention, by the way, when Facebook tells you that this is all about privacy, is that on WhatsApp, their service, if somebody sends you a link, and that link is malware, it's dangerous to you, it will be highlighted in the message. How are they doing that? They are reading your message. Why? For security purposes. Can we please agree that protecting you from malware is at least as important as protecting 4-year-olds and 6-year-olds and 8-year-olds from physical sexual abuse?

We have the technology to do this, and the rush to end-to-end encryption, which, by the way, I think is a head fake. They're using Cambridge Analytica to give them plausible deniability on all the other issues that we have been trying to get them—progress on, from child sexual abuse, to terrorism, to conspiracies, to disinformation. If they end-to-end encrypt, we will lose the ability to know what's going on on their platforms, and you have heard very eloquently from my colleague that this will be a disaster. You

should not let them do this without putting the right safeguards in place.

Mr. BEYER. So you were just making a powerful argument now for national and international level banning end-to-end encryption?

Dr. FARID. I wouldn't go that far. We want end-to-end encryption for banking, for finance. There are places where it is the right thing to do, but there are other places where we have to simply think about the balance. So, for example, in my solution I didn't say don't do end-to-end encryption. I said put the safeguards in place so that if somebody's transmitting harmful content, I can know about it.

I have mixed feelings about the end-to-end encryption, but I think, if you want to do it, and we should think seriously about that, you can still put the safeguards in place.

Mr. BEYER. And blockchain is not end-to-end encryption?

Dr. FARID. No, it is not.

Mr. BEYER. But it gets close?

Dr. FARID. These are sort of somewhat orthogonal separate issues, right? What we are talking about is a controlled platform saying that—everything that comes through us, we will no longer be able to see. That is super convenient for the Facebooks of the world, who don't want to be held accountable for the horrible things happening on their platforms, and I think that's the core issue here.

Mr. BEYER. Great, thanks. Anything else? All right. I think Mr. Gonzalez and I are done and thank you very much. It's a very, very interesting mission, and don't be discouraged that there weren't more Members here, because everyone's in their office watching this, and have their own questions. So thank you very much for being here, and thanks for your witness stuff. And the record will remain open for 2 weeks for additional statements from the Members, and, additionally, we may have questions of you to answer in writing. So thank you very much.

Dr. FARID. OK.

Mr. BEYER. You're excused, and the hearing is adjourned.

Dr. FARID. Thank you.

[Whereupon, at 3:26 p.m., the Subcommittee was adjourned.]

Appendix

ADDITIONAL MATERIAL FOR THE RECORD

REPORT SUBMITTED BY MS. CAMILLE FRANÇOIS



Actors, Behaviors, Content: A Disinformation ABC
Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses†

Camille François
 Graphika and Berkman Klein Center for Internet & Society at Harvard University¹
 September 20, 2019

Contents

Introduction	1
“A” is for Manipulative Actors	2
“B” is for Deceptive Behavior.....	4
“C” is for Harmful Content.....	6
Conclusion and recommendations	7
Appendix: Examples of Disinformation Campaigns Spanning the Three Vectors	7
Notes	8

Introduction

As the historic phenomenon of propaganda² unfolds today in a variety of social-media manifestations, a plethora of terms has emerged to describe its different forms and their implications for society: “fake news,” online disinformation, online misinformation, viral deception, etc.³ The speed and scale at which disinformation is now able to spread online has led to mounting pressure on regulators around the globe to address the phenomenon, yet its multifaceted nature makes it a difficult problem to regulate. Effective remedies must take into account the different vectors of contemporary disinformation and consider the multiplicity of stakeholders, tradeoffs in different approaches, disciplines, and regulatory bodies able to meaningfully contribute to responses.

† One in a series: A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: <https://www.ivir.nl/twg/>.

Major technology platforms have invested in better responses to disinformation, notably by adapting their community guidelines or terms of service. Observed through the lens of platform enforcement, “disinformation” breaks down into a number of different violations manifest on different products which are enforced by distinct teams. This points to a key concern with regard to the current industry responses to viral deception: while disinformation actors exploit the whole information ecosystem in campaigns that leverage different products and platforms, technology companies’ responses are mostly siloed within individual platforms (if not siloed by individual products!).

This concise “ABC” framework doesn’t aim to propose one definition or framework to rule them all, but rather seeks to lay out three key vectors characteristic of viral deception⁴ in order to guide regulatory and industry remedies. Manipulative **actors**, deceptive **behaviors**, harmful **content**: each vector presents different characteristics, difficulties, and implications. Unfortunately, they are also often intertwined in disinformation campaigns, suggesting that effective and long-term approaches will need to address these different vectors with appropriate remedies.

This “ABC” also seeks to reconcile approaches throughout applicable disciplines (e.g., cybersecurity, consumer protection, content moderation) and stakeholders. While the public debate in the U.S. has been largely concerned with **actors** (who is a Russian troll online?), the technology industry has invested in better regulating **behavior** (which accounts engage in coordinated and inauthentic behavior?) while governments have been most preoccupied with **content** (what is acceptable to post on social media?).

“A” is for Manipulative Actors

“On the Internet, nobody knows you’re a ~~dog~~ Russian military operative.”¹⁶

The Russian disinformation campaign targeting the U.S. 2016 presidential election⁶ has brought to the public’s attention how keen certain government actors were to leverage social media to manipulate and influence audiences at home and abroad by engaging in information operations. It has also painfully brought to light the lack of government and industry preparedness and proactivity in the face of these threats. The cybersecurity sector, which bears the brunt of detecting these threat actors and preventing their nefarious activities, had been most focused on protecting physical networks and not enough on detecting those actors on social media networks. Facebook’s April 2017 white paper on the issue of information operations (which also marks the first in-depth acknowledgement of this problem by a large technology platform) makes this point clearly and acknowledges that the Facebook cybersecurity team had to expand its scope to appropriately respond to this threat: “We have had to expand our security focus from traditional abusive behavior, such as account hacking, malware, spam and financial scams, to include more subtle and insidious forms of misuse, including attempts to manipulate civic discourse and deceive people.”¹⁷

Manipulative actors, by definition, engage *knowingly* and with clear intent in viral deception campaigns. Their campaigns are *covert*, designed to obfuscate the identity and intent of the actor orchestrating them. Throughout the technology industry, detection and enforcement of this vector of viral deception campaigns rely on the cat-and-mouse game of a) identifying threat actors willing and able

to covertly manipulate public discourse and b) keeping those actors from leveraging social media to do so,⁸ as they refine their strategies to evade detection.

Because this detection practice has its roots in the cybersecurity realm, terms of service and community guidelines do not always address these issues, or provide a clear basis to support detection and enforcement efforts against manipulative actors. Precedents in this area include platform rules laying out specific actors who are prevented from using the services (e.g., Foreign Terrorist Organizations⁹), but it is worth noting that no major platform to date has included language in its terms of service explicitly prohibiting governments from covertly using its services to conduct influence campaigns.¹⁰ Setting an industry precedent, in August 2019, following investigations disclosing that Chinese State-controlled media leveraged Twitter advertising to promote content critical of pro-democracy protests in Hong Kong, Twitter announced that it would no longer allow “State-controlled media” to use its advertising products.¹¹ The state-controlled media entities can continue to remain “organic users” (meaning normal and/or verified accounts on the Twitter platform), but their ability to use ads to reach users who are not already following them is now restricted. In doing so, Twitter will likely face difficulty determining which entities are “taxpayer funded entities” and “independent public broadcasters” allowed to use the advertising services vs. “state-controlled media (...) financially or editorially controlled by the state” prohibited from doing so. States have also used a variety of techniques to conceal their direct involvement in seemingly independent online media properties: the Kremlin-controlled Baltnews network¹² and the Iranian-controlled IUVM¹³ network are good illustrations.

Note that this problem has little to do with “banning” anonymity or pseudonymity online: both serve important purposes in protecting vulnerable voices and enabling them to participate in critical conversations.¹⁴ Banning anonymity/pseudonymity would prevent such participation while doing little to prevent sophisticated and well-funded actors from exploiting this vector. The deceptive actors we are concerned with here are well-funded military and intelligence apparatus or campaign apparatus, not “somebody sitting on their bed that weighs 400 pounds,” as President Trump famously characterized the anonymous troll. Clint Watts describes these figures as “Advanced Persistent Manipulators,”¹⁵ a moniker that stresses the parallels and overlaps between the actors engaged in information operations¹⁶ and hacking.¹⁷

Similar to the challenge APT¹⁸ actors have posed to information and cyber security professionals, social media companies now face malign actors that can be labeled as Advanced Persistent Manipulators (APMs) on their platforms. These APMs pursue their targets and seek their objectives persistently and will not be stopped by account shutdowns and platform timeouts.... They have sufficient resources and talent to sustain their campaigns, and the most sophisticated and troublesome ones can create or acquire the most sophisticated technology.¹⁹

Since 2017, we have seen multiple examples of viral deception campaigns whose *primary* vector is a deceptive actor. Notable examples include false persona “Guccifer 2.0”²⁰ used by the GRU, false identities tying back to the Islamic Republic of Iran Broadcasting and operating on multiple platforms,²¹ and Facebook’s December 2018 takedown of accounts in Bangladesh that were found to be misrepresenting their true identity and attempting to mislead voters ahead of the elections.²²

Governments also have a role to play in detecting and mitigating harms caused by manipulative actors online, although defining the contours of government action in this space remains a largely unexplored policy question. Around the U.S. 2018 midterms elections, for instance, the U.S. government led actions to detect and share relevant information on manipulative actors with the technology sector²³ and to disrupt and deter these actors from engaging in information operations.²⁴

“B” is for Deceptive Behavior

“On the Internet, nobody knows you’re a ~~dog~~ bot army.”

Deceptive behavior is a fundamental vector of disinformation campaigns: it encompasses the variety of techniques viral deception actors may use to enhance and exaggerate the reach, virality and impact of their campaigns. Those techniques run from automated tools (e.g., bot armies used to amplify the reach and effect of a message) to manual trickery (e.g., paid engagement, troll farms). At the end of the day, deceptive behaviors have a clear goal: to enable a small number of actors to have the *perceived impact* that a greater number of actors would have if the campaign were organic.²⁵

Interestingly, while there are significant differences in the various disinformation definitions and terms of service applicable to the issue among technology companies, the focus on *deceptive behavior* appears to be a clear convergence point throughout the technology industry.

Google’s definition of disinformation, as laid out in its February 2019 White Paper on “How Google Fights Disinformation,” points to those deceptive behaviors as a core vector of how disinformation affects Google’s platforms:

We refer to [...] deliberate efforts to deceive and mislead using the speed, scale, and technologies of the open web as “disinformation.”²⁶

In Facebook’s case, deceptive behavior is mostly defined through the “Coordinated Inauthentic Behavior”²⁷ policy, which has led to numerous takedowns since it was implemented in 2018.²⁸ While Facebook has shared records and data points regarding the content and accounts taken down for their participation in “coordinated and inauthentic behavior,” enforcement in this realm remains opaque throughout the major technology companies.

While the detection and mitigation techniques in this area can be similar to spam detection, an area generally opaque for the public and regulators and not subject to much public scrutiny, the free speech implications of taking down *content* and social media *accounts* (especially political content during election cycles) justify much higher scrutiny of these practices. Relevant questions to technology platforms in this area include:

- Applicable rules: Which are the applicable policies set forth by the platform to address deceptive behaviors on their products?
- Enforcement: What enforcement options are available to the platforms to take action against accounts and content that violate the rules on deceptive behavior? Platforms generally acknowledge a range of options from content demotion to account suspension, although those enforcement options are rarely spelled out for users or made clear for users affected.

- Detection and prioritization: Which teams are effectively in charge of detecting deceptive behaviors, how much of this detection relies on machine learning classifiers (and which ones?), and how does prioritization of potential issues and focus areas work at the platform level?
- Transparency: How will affected users (including good faith actors mistakenly engaging in deceptive behaviors, consumers of information spread by deceptive behavior, bad faith actors seeking to best understand what telltale signs trigger enforcement, etc.) be notified when action is taken against content or accounts? Can those decisions be appealed, and if so, how? Will the platform share transparency metrics regarding its enforcement of rules relative to distortive behavior, both at the annual and the aggregate level (through the existing mechanism of Transparency Reports) and through press releases published when enforcement happens?
- Product vulnerabilities and changes: When deceptive behaviors exploit vulnerabilities in platforms and products, what changes are made to address them?²⁹

The industry's lack of proactivity in tackling some of these campaigns and growing public anxiety about disinformation have led regulators to craft frameworks to specifically address deceptive behavior. California's "Bot Law," for instance, is a clear attempt to regulate deceptive behavior on social media:

It shall be unlawful for any person to use a bot to communicate or interact with another person in California online, with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to incentivize a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election. A person using a bot shall not be liable under this section if the person discloses that it is a bot. The disclosure required by this section shall be clear, conspicuous, and reasonably designed to inform persons with whom the bot communicates or interacts that it is a bot.³⁰

The "Manipulative Actor" and "Deceptive Behavior" vectors are particularly challenging to address through effective regulatory frameworks because of the dramatic asymmetry of information between the platforms targeted by these campaigns and the rest of the world. While open-source investigation techniques and a few available tools allow others to scrutinize online activity for campaigns run by a manipulative actor or using deceptive techniques, it is undeniable that platforms have much more visibility into those issues than external researchers and stakeholders. Some platforms' community standards or terms of service either indirectly prevent the type of external research that may lead to detecting and exposing distortive behaviors (e.g., when existing and important safeguards also prevent researchers from collecting the data they'd need to analyze distortive behaviors) or directly seek to prevent it (e.g., with rules explicitly preventing the use of data in order to perform detection of deceptive behavior).

Finally, some of the platforms' own systems may actually enhance those deceptive behaviors by disinformation actors: algorithmic reinforcement is a core concern in this area.³¹ While anecdotal evidence suggests machine learning based recommendations systems may easily be gamed into promoting campaigns "boosted" by adversarial distortive behavior, the difficulties discussed above

with regard to external research have prevented more systematic examinations of these issues throughout the various platforms.

“C” is for Harmful Content

“On the Internet, nobody knows you’re a ~~dog~~ deepfake.”

Finally, it is sometimes the case that the content of posts and messages justifies classifying a campaign as an instance of viral deception. Content is the most visible vector of the three: while it is difficult for an observer to attribute messages to a manipulative actor or to observe behavioral patterns across a campaign, every user can see and form an opinion on the content of social media posts. This is likely why regulators have focused on content aspects when regulating disinformation.

This vector calls for detection and enforcement strategies in the realm of content moderation³². Unfortunately, regulatory and legal frameworks often struggle to properly define categories of “harmful content” they seek to regulate (see ongoing debates about the definitions of “violent extremism,” “hate speech,” “terrorist content,” etc.) or to properly take into account that a lot of the speech they consider to be “harmful” is protected under human rights law. Governments’ appetite to regulate viral deception through the content lens risk further eroding protections to freedom of expression online.

The intersection of harmful content and disinformation campaigns can manifest in several ways:

- Entire categories of content can be deemed “harmful” because they belong to the realm of viral deception, e.g., health misinformation.³³

Technology platforms have so far mostly proposed to address the categories of content deemed most “harmful” for their disinformation nature by adding context for users alongside the content, such as “flags” or “fact-checking” content. Some platforms though have taken a more radical route by banning entire categories of disinformation content from their services.

Photo-sharing platform Pinterest, for instance, takes action against harmful medical information shared on its platform. Its “Health Misinformation” policy reads:

“Pinterest’s misinformation policy prohibits things like promotion of false cures for terminal or chronic illnesses and anti-vaccination advice. Because of this, you’re not allowed to save content that includes advice where there may be immediate and detrimental effects on a Pinner’s health or on public safety.”³⁴

- The content of a campaign itself (not its diffusion mechanism) can be manipulated to deceive users and therefore belong to the realm of “disinformation” (e.g., use of manipulated media on the range from “deepfakes” to “cheap fakes”³⁵).
- “Harmful content” can be promoted by deceptive actors or by campaigns leveraging distortive behaviors (e.g., “troll farms amplifying harassment campaigns”).

It should indeed be noted that viral deception campaigns whose primary vector is a deceptive actor or distortive behavior can participate in amplifying other types of harmful content categories, such as hate speech, harassment, and violent extremism.

Conclusion and recommendations

Viral deception campaigns spread across platforms and through three core vectors: manipulative actors (A), deceptive behavior (B) and harmful content (C). As such, they represent a complex and multifaceted problem for policy makers and regulators to address. This “ABC” framework therefore offers a few modest recommendations for policy makers and regulators navigating this maze:

- Each dimension matters. Regulatory efforts focused on viral deception tend to exaggerate the role of harmful content: balanced approaches will consider how manipulative actors (both foreign and domestic) and deceptive behaviors contribute to the problem.
- Each dimension comes with its own set of challenges, tradeoffs, and policy implications. Specific disciplines may be necessary and/or best suited to address each of them. For instance, cybersecurity (and threat intelligence in particular) is a core component of how manipulative actors get detected; how the resulting signals get shared across the industry and with the relevant parties (researchers, public institutions) is a key policy question. Consumer protection frameworks (and stakeholders) may be ideally situated to help regulate deceptive behavior issues. Policies and regulatory frameworks that center around one type of remedy only (such as content takedowns) are insufficient.
- On a final (and related) note, Manipulative Actors (A) and Deceptive Behaviors (B) are dimensions on which the information asymmetry between the technology platforms on which this activity unfolds and the rest of the stakeholders in the debate is immense. How to ensure that the public, media, and policy stakeholders are able to meaningfully analyze both the issues and potential impacts of remedies in place is a fundamental question in this space.

Appendix: Examples of Disinformation Campaigns Spanning the Three Vectors

- A Disinformation Campaign in the Philippines (Facebook)

On March 28, 2019, Facebook removed 200 pages, groups and accounts engaged in “coordinated inauthentic behavior” on Facebook and Instagram in the Philippines. Facebook’s press release³⁶ highlights the manipulative actor along with the deceptive behavior elements of the campaign:

We’re taking down these Pages and accounts based on their behavior, not the content they posted. In this case, the people behind this activity coordinated with one another and used fake accounts to misrepresent themselves, and that was the basis for our action.

Follow-up analysis highlights that the content taken down by Facebook in this campaign did contain “harmful content,” notably in the form of hate speech and manipulated media (Photoshopped images of politicians in wheelchairs enticing viewers to question the health of candidates).³⁷

- The Russian Internet Research Agency’s “Columbia Chemical” Campaign (Twitter)

On September 11, 2014, a set of seemingly uncoordinated Twitter accounts engaged in disseminating news of a chemical incident and toxic fumes in the city of St. Mary Parish in Louisiana. Along with the social media campaign, videos of the “incident” were uploaded and officials and media were contacted by available channels with an alarming messaging – “Take shelter!” – and links to a dedicated website (www.jcolumbiachemical.com).³⁸

It wasn’t long until officials realized that the campaign, with its false images of the incident and alarming messages, constituted harmful content – “a hoax,” as it was initially described. It was later made clear that the accounts used to spread the content were coordinated to give the impression of a mounting local panic, using distortive behavior to create the illusion of a spontaneous wave of local panic.

It took a few more years for the major technology platforms and the U.S. Government to provide a final attribution on those accounts, confirming that the Internet Research Agency troll farm in Saint Petersburg was indeed the actor operating the accounts.³⁹

Notes

¹ Camille François works on cyber conflict and digital rights online. She is Chief Innovation Officer of Graphika, where she leads the company’s work to detect and mitigate disinformation, media manipulation and harassment in partnership with major technology platforms, human rights groups and universities around the world. She also is an affiliate of Harvard University’s Berkman Klein Center for Internet & Society. An earlier version of this paper was presented at the second meeting of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression, in Santa Monica, Calif., on May 9-12, 2019. The author thanks her colleagues in the working group for their engagement with this work during the sessions. Their feedback and encouragement greatly benefited this final paper.

² See for instance: Tworek, Heidi JS. *News from Germany: The Competition to Control World Communications, 1900–1945*. Harvard University Press, 2019.

³ For a thoughtful typology of the different aspects of the phenomenon, see for instance Claire Wardle and Hossein Derakhshan’s “Information Disorder” framework: <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>.

⁴ Viral deception here is used as an umbrella term for the multiple facets of contemporary disinformation online, see Jamieson, Kathleen Hall. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don’t, Can’t, and Do Know*. Oxford University Press, 2018.

⁵ I hope readers will forgive this 2019 edit to [the famous cartoon](#) published by Peter Steiner in the New Yorker on July 1993.

⁶ See the Mueller Report: <https://www.justice.gov/storage/report.pdf>

⁷ Jen Weedon, William Nuland and Alex Stamos, “Information Operations and Facebook”, v.1: <https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf>

⁸ For an example of a takedown solely motivated by the actor behind the content, see Facebook’s “The IRA Has No Place On Facebook” post on April 3, 2018: “We removed this latest set of Pages and accounts solely because they were controlled by the IRA — not based on the content.” <https://newsroom.fb.com/news/2018/04/authenticity-matters/>

⁹ See for instance Microsoft’s “*Approach to Terrorist Content*” statement (published May 20, 2016), which notes that “there is no universally accepted definition of terrorist content” and that Microsoft relies on organizations listed in the Consolidated United Nations Security Council Sanctions List to define and take action against terrorist content posted on its platforms: <https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/>

¹⁰ There are, however, multiple policies that indirectly cover aspects of these campaigns. An example with Google Ads’ “Misrepresentation Policy”: <https://support.google.com/adspolicy/answer/6020955?hl=en>

¹¹ In 2017, Twitter had similarly banned Russian State-controlled media Russia Today and Sputnik from using their advertising products (https://blog.twitter.com/en_us/topics/company/2017/Announcement-RT-and-Sputnik-Advertising.html). The August 2019 policy extends this ad-hoc remediation done in the wake of the investigation

- regarding the Kremlin's election interference efforts on social media to all of "state-controlled" media: https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html
- ¹² See the Aug. 29, 2019 BuzzFeed investigation, "This Is How Russian Propaganda Actually Works in the 21st Century": <https://www.buzzfeednews.com/article/holgerroonemaa/russia-propaganda-baltics-balmews>
- ¹³ See for instance DFRLab's "In Depth: Iranian Propaganda Network Goes Down," March 26, 2019, <https://medium.com/dfrlab/takedown-details-of-the-iranian-propaganda-network-d1fad32fd30>
- ¹⁴ For an examination of how manipulative actors use "pseudoanonymity" to "impersonate marginalized, underrepresented, and vulnerable groups to either malign, disrupt or exaggerate their cause," see Friedberg and Donovan's piece in the MIT JODS: <https://jods.mitpress.mit.edu/pub/2gnso48a>
- ¹⁵ Clint Watts, "Advanced Persistent Manipulators", Feb. 12, 2019: <https://securingdemocracy.gmfus.org/advanced-persistent-manipulators-part-one-the-threat-to-the-social-media-industry/>
- ¹⁶ For a global inventory of actors organized for social media manipulation, see: Bradshaw, Samantha, and Philip Howard. "Troops, trolls and troublemakers: A global inventory of organized social media manipulation." (2017).
- ¹⁷ See also "False Leaks: A Look at Recent Information Operations Designed To Disseminate Hacked Material," Camille Francois, CYBERWARCON 2018. Video: <https://www.youtube.com/watch?v=P8iXN8j4gMk>
- ¹⁸ APT here refers to Advanced Persistent Threat, a term commonly used in the threat intelligence industry to describe State-sponsored and state-affiliated groups engaged in hacking operations. See: https://en.wikipedia.org/wiki/Advanced_persistent_threat
- ¹⁹ Clint Watts, "Advanced Persistent Manipulators," Feb. 12, 2019: <https://securingdemocracy.gmfus.org/advanced-persistent-manipulators-part-one-the-threat-to-the-social-media-industry/>
- ²⁰ Guccifer is a social media persona who claimed to be the hacker who hacked the Democratic National Committee in 2016, and who used this deceptive identity to engage WikiLeaks and the media. The account was in reality operated by Russian military intelligence: https://en.wikipedia.org/wiki/Guccifer_2.0
- ²¹ See for instance Google's Kent Walker update on action taken against IRIB and broader State-Sponsored activity on Google's products: <https://blog.google/technology/safety-security/update-state-sponsored-activity/>
- ²² <https://newsroom.fb.com/news/2018/12/take-down-in-bangladesh/>
- ²³ See for instance reporting by the Associated Press, "Facebook blocks 115 accounts ahead of US midterm elections", Nov. 6, 2018, <https://www.apnews.com/19aabf8ba7b6466b859f4d0afd9e59be>. The AP reports: "Facebook acted after being tipped off Sunday by U.S. law enforcement officials. Authorities notified the company about recently discovered online activity 'they believe may be linked to foreign entities.'"
- ²⁴ See Ellen Nakashima's reporting in the Washington Post, "U.S. Cyber Command operation disrupted Internet access of Russian troll factory on day of 2018 midterms", Feb. 26, 2019, https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9c-36d6-11e9-af5b-b51b7ff322e9_story.html
- ²⁵ I am borrowing here from a definition my colleagues and I have used to frame detection techniques. See Francois, Barash, Kelly: <https://osf.io/ap2yz/>
- ²⁶ "How Google Fights Disinformation," Feb. 2019, available at: https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf
- ²⁷ See "Coordinated Inauthentic Behavior Explained," <https://newsroom.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/>
- ²⁸ A blog post entitled "Removing Bad Actors On Facebook", from July 2018, seems to be the first public reference to "coordinated and inauthentic behavior": <https://newsroom.fb.com/news/2018/07/removing-bad-actors-on-facebook/>
- ²⁹ An example of a product change directly motivated by a platform's need to tackle distortive behaviors on its products can be found in the January 2019 YouTube announcement: "'To that end, we'll begin reducing recommendations of borderline content and content that could misinform users in harmful ways':" <https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html>
- ³⁰ https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001
- ³¹ See for instance former YouTube engineer Guillaume Chaslot's project regarding algorithmic reinforcement of fringe and harmful views on YouTube: <https://algotransparency.org/methodology.html>
- ³² For an in-depth discussion of the various issues plaguing the content moderation industry, see Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019. or Gillespie, Tarleton. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.
- ³³ This is a good place for a quick reminder of the differences between misinformation and disinformation. [Dictionary.com](https://www.dictionary.com), which made "misinformation" the word of the year in 2018, defines it as "false information that is spread, regardless of whether there is intent to mislead." It describes disinformation as "deliberately misleading or biased information; manipulated narrative or facts; propaganda".
- ³⁴ <https://help.pinterest.com/en/article/health-misinformation>

³⁵ See Britt Paris and Joan Donovan, “Deep Fakes and Cheap Fakes”, published Sept. 18th 2019 by the Data & Society Research Institute, <https://datasociety.net/output/deepfakes-and-cheap-fakes/>

³⁶ <https://newsroom.fb.com/news/2019/03/cib-from-the-philippines/>

³⁷ <https://medium.com/graphika-team/archives-facebook-finds-coordinated-and-inauthentic-behavior-in-the-philippines-suspends-a-set-d02f41f527df>

³⁸ Adrian Chen's 2015 account in The New York Times Magazine is the first public account of this campaign: <https://www.nytimes.com/2015/06/07/magazine/the-agency.html>

³⁹ See for instance the reports commissioned by the Senate Select Intelligence Committee regarding the IRA's online activity targeting the USA: <https://comprop.oi.ox.ac.uk/research/ira-political-polarization/>