TESTIMONY OF SPENCER OVERTON
Patricia Roberts Harris Research Professor of Law
George Washington University

Before the
Subcommittee on Cybersecurity, Information Technology, and Government Innovation
U.S. House Committee on Oversight and Accountability

Hearing on
"Advances in Deepfake Technology"

November 8, 2023*

Overview

While deepfake audio and video technologies can open opportunities for creativity and education, they also expand opportunities for harms disproportionately borne by women, communities of color, and religious minorities.[1] These challenges only grow as deepfake technology becomes more advanced and accessible.[2]

The vast majority of deepfake videos are deepfake pornography featuring women rather than men, and these videos are often weaponized to intimidate and silence women. Deepfake technology is fueling online gendered and racial harassment, which studies show causes real harms—including anxiety, depression, isolation, and reduced academic performance. Deepfake technology is driving racial distrust—such as falsely attributing racially-insensitive comments to political candidates, and has the potential to diminish the credibility of authentic video showing misconduct by officials. Deepfake technology can also enhance attempts by Russian officials to impersonate Americans and interfere in U.S. elections. Unfortunately, even technologies developed to detect deepfakes that could prevent some of these harms are less effective in identifying synthetic videos featuring people of color.

---

* Exchanges with K.J. Bagchi, Danielle Keats Citron, and Maya Wiley helped develop the ideas in this written testimony. Eleonora Viotto provided invaluable research assistance.

[1] Mary Anne Franks & Ari Ezra Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD. L. REV. 892, 896 (2019) ("[T]he costs of deep fakes are not equally borne across society. Research demonstrates that women and racial and sexual minorities are more likely to be victimized by these abuses and to suffer more severe consequences because of them."); Raquel Vazquez Llorente & Sam Gregory, *Regulating Transparency in Audiovisual Generative AI: How Legislators Can Center Human Rights*, TECH POLICY PRESS, Oct. 18, 2023 ("While there are creative and commercial benefits to generative AI and synthetic media, these tools are connected to a range of harms that are impacting disproportionately those communities that were already vulnerable to mis- and disinformation, or targeted and discriminated against because of their gender, race, ethnicity, or religion.").

[2] HOME SECURITY HEROES, 2023 STATE OF DEEPFAKES: REALITIES, THREATS, AND IMPACT (2023) ("Another significant factor driving advancements in deepfake technology is the increasing accessibility of user-friendly tools, software, and online communities."); Danielle Citron, *Combatting Online Harassment*, DEMOCRACY: A JOURNAL OF IDEAS, Spring 2023 ("Technological advances have made it cheaper and easier for perpetrators to wreak havoc. With machine-learning algorithms, perpetrators can now create frighteningly realistic fake sex videos.").

I. Abusive Deepfakes Silence Political Participation by Women

Nina Jankowicz is a 34-year old researcher who specializes in state-sponsored disinformation and online gendered abuse. After being appointed as executive director of the U.S. Department of Homeland Security's Disinformation Governance Board in April 2021, online activists erroneously attacked Jankowicz as a free speech censor, published her contact information, and threatened her with rape and death.[3] Security experts advised Jankowicz and her husband to relocate, which was not feasible because she was nine-months pregnant at the time.[4]

About a year later the Biden Administration dissolved the board and Jankowicz resigned, and thereafter she discovered she was the subject of at least three artificially-generated videos that appear to show Jankowicz engaging in sex acts. As Jankowicz wrote about the deepfake videos:

> Although they may provide cheap thrills for the viewer, their deeper purpose is to humiliate, shame, and objectify women, especially women who have the temerity to speak out. I am somewhat inured to this abuse, after researching and writing about it for years. But for other women, especially those in more conservative or patriarchal environments, appearing in a deepfake-porn video could be profoundly stigmatizing, even career- or life-threatening. . . . Users can also easily find deepfake-porn videos of the singer Taylor Swift, the actress Emma Watson, and the former Fox News host Megyn Kelly; Democratic officials such as Kamala Harris, Nancy Pelosi, and Alexandria Ocasio-Cortez; the Republicans Nikki Haley and Elise Stefanik; and countless other prominent women. By simply existing as women in public life, we have all become targets, stripped of our accomplishments, our intellect, and our activism and reduced to sex objects for the pleasure of millions of anonymous eyes.[5]

The use of deepfakes to intimidate women is not limited to the United States. After investigative reporter Rana Ayyub criticized Prime Minister Narendra Modi in a BBC broadcast, a source in the Modi government instructed Ayyub to check her text messages. She saw a deepfake sex video with her face, which would eventually be shared over 40,000 times in group text messages and on social media sites.[6] Her home address and cell phone were posted online, and she was swamped with rape and death threats and inquiries about her rates for sex. Ayyub stopped writing and suffered heart palpitations and anxiety, and has since been unable to publish in India's news outlets.[7]

---

[3] *See* Danielle Citron, *Combatting Online Harassment*, DEMOCRACY: A JOURNAL OF IDEAS, Spring 2023; Nina Jankowicz, *I Shouldn't Have to Accept Being in Deepfake Porn*, THE ATLANTIC, June 25, 2023.

[4] Danielle Citron, *Combatting Online Harassment*, DEMOCRACY: A JOURNAL OF IDEAS, Spring 2023.

[5] Nina Jankowicz, *I Shouldn't Have to Accept Being in Deepfake Porn*, THE ATLANTIC, June 25, 2023.

[6] DANIELLE KEATS CITRON, THE FIGHT FOR PRIVACY: PROTECTING DIGNITY, IDENTITY, AND LOVE IN THE DIGITAL AGE 56, 107-08 (2022).

[7] *See also* Eliza Mackintosh & Swati Gupta, *Troll armies, 'deepfake' porn videos and violent threats. How Twitter became so toxic for India's women politicians*, CNN, Jan. 22, 2020 ("…India's patriarchal social structure has taken on a new dimension online, where men vandalize women's internet profiles, use filthy language to describe their sex appeal, publish intimate images without their consent or share doctored imagery -- known as "deepfakes" -- depicting

In the United States, a study of online harassment of candidates during the 2020 congressional election (including threats of violence, non-consensual image/video sharing, and doxxing) found that women of color were more likely to be the target of "sexist abuse (as compared to white women), racist abuse (as compared to men of color), and violent abuse (four times more than white candidates and two times more than men of color)."[8]

Deepfake porn is not simply obscenity that "hurt[s] the women's feelings," but is an anti-democratic form of harassment designed to undermine public confidence in women as leaders and legitimate participants in public policy debates.[9] Online abuse silences victims and chills participation in democracy by women around the globe.[10]

II. Most Deepfakes are Pornography Featuring Women

Broadly, deepfake pornography featuring women accounts for the vast majority of deepfake videos, and the volume is quickly rising. Deepfake pornography accounts for 98% of deepfake videos online, and 99% of all deepfake porn features women while only 1% features men.[11] The total number of deepfake porn videos produced in 2023 increased 464% from 2022 (to 21,019 from 3,725), and in 2023 the monthly traffic of the leading ten dedicated deepfake porn websites reached over 34 million.[12] The U.S. accounts for 20% of the world's deepfake pornography, trailing only South Korea (which accounts for 53%).[13]

When asked about their reaction if someone close to them became a victim of deepfake porn, 73% of American males surveyed expressed a desire to report the incident to authorities and 68% indicated they would feel shocked and outraged by the violation of privacy.[14]

---

them in pornography. India's youngest parliamentarian, Chandrani Murmu, was subjected to such a "deepfake," with her face superimposed onto an obscene video, before she was elected last year.").

[8] Dhanaraj Thankur & DeVan Hankerson Madrigal, *An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States*, CENTER FOR DEMOCRACY & TECHNOLOGY, Oct. 27, 2022.

[9] Women's Media Center, *Katie Hill, Deepfakes, and How "Political Risk" is Defined*, Nov. 7, 2019 ("Depicting women politicians as sexually sullied, violable, voracious, or promiscuous is a gendered tactic of harassment that undermines women as politicians, has spillover effects on the public's confidence in women as leaders and citizens, and is profoundly anti-democratic.").

[10] Danielle Citron, *Combatting Online Harassment*, DEMOCRACY: A JOURNAL OF IDEAS, Spring 2023 ("Women in politics around the globe avoid expressing views that might provoke online attacks. Finland's first Black female member of Parliament, Bella Forsgrén, has explained that she thinks twice not only about how she talks about issues, but also about the decisions that she makes, lest she face online backlash. The threat of online abuse has also deterred women and minorities from considering political careers. Eighty percent of Australian women surveyed said that the mistreatment of female politicians online made it less likely they would go into politics."); Jonathon W. Penney, *Internet Surveillance, Regulation, and Chilling Effects Online: A Comparative Case Study*, 6 INTERNET POL'Y REV., no. 2, 2017, at 1, 19; Danielle Keats Citron, *From Bad to Worse: Cyber Stalking and Free Speech* (draft manuscript).

[11] HOME SECURITY HEROES, 2023 STATE OF DEEPFAKES: REALITIES, THREATS, AND IMPACT (2023) (also indicating that 94% of those featured in deepfake pornography videos are actresses, singers, or social media influencers).

[12] *Id.*

[13] *Id.*

[14] *Id.* (detailing the results of a survey of the attitudes of 1522 American males about deepfake pornography).

Victims of online harassment stop using their phones, social media accounts, and withdraw from digital engagement.[15] A 2019 study found that politically-active women sent fewer tweets after online attacks.[16] "[Y]ounger women are most likely to self-censor to avoid potential online harassment: 41% of women ages 15 to 29 self-censor, compared with 33% of men of the same age group and 24% of internet users ages 30 and older (men and women)."[17] A study of 14,000 girls from 31 countries found that 19 percent of girls harassed frequently online said they use social media platforms less and 12 percent said that they stopped using them altogether.[18] Cyber harassment causes various other harms, including anxiety and depression, challenges maintaining employment and finding new jobs, and challenges developing personal relationships.[19]

III. Racial Harassment Deepfakes

Easy access to deepfake technology is also fueling racial harassment. For example, in early February 2023 in Putnam County, New York – about 60 miles north of New York City – students made and circulated a deepfake video of a local middle school principal saying Black students should be sent back to Africa, calling them monkeys and the n-word, and ending with "I am bringing my machine gun to school."[20] Other similar videos used racist slurs against Black and Latino students, said the "KKK legacy will return," and showed an animated version of the middle school where a shooter fires at Black and Latino students. While school district officials disciplined the high school students who created the videos, the sheriff's office found they had not committed any crimes.

Putnam County Afro-Latina parent Abigail Santana said her 10-year-old always used to be excited about attending school. After seeing the threatening deepfakes on TikTok, however, she started

---

[15] Francesa Stevens et al., *Cyber Stalking, Cyber Harassment, and Adult Mental Health: A Systematic Review*, 24 CYBERPSY., BEHV & SOC. NET. REV. 367, 372 (2021); Delanie Woodlock, *The Abuse of Technology in Domestic Violence and Stalking*, 23 VIOLENCE AGAINST WOMEN 584 (2017); Soheila Pashang et al., *The Mental Health Impact of Cyber Sexual Violence on Youth Identity*, 17 INT'L J. MENTAL HEALTH & ADDICTION 1119 (2019).

[16] Kristen Zeiter et al, *Tweets that Chill: Analyzing Online Violence Against Women in Politics*, NAT'L DEM. INST. (June 14, 2019)

[17] AMANDA LENHART, MICHELE YBARRA, KATHRYN ZICKUHR & MYESHIA PRICE-FEENEY, DATA & SOC'Y RSCH. INST. & CTR. FOR INNOVATIVE PUB. HEALTH RSCH., ONLINE HARASSMENT, DIGITAL ABUSE, AND CYBER STALKING IN AMERICA 4 (2016); Danielle Citron, *Combatting Online Harassment*, DEMOCRACY: A JOURNAL OF IDEAS, Spring 2023 ("Victims' very ability to speak freely is in jeopardy. When under assault online, women are more likely to self-censor their speech. Younger women are the most likely to self-censor to avoid further abuse. They shut down their social media accounts because they worry that keeping them will invite more abuse and provide another vector for perpetrators to reach them.").

[18] Sharon Gould et al, *Free to Be Online? Girls' and Young Women's Experiences of Online Harassment*, PLAN INTERNATIONAL 31 (2020).

[19] Danielle Citron, *Combatting Online Harassment*, DEMOCRACY: A JOURNAL OF IDEAS, Spring 2023 ("Cyber harassment impacts every aspect of victims' lives. They lose their jobs and have difficulty finding new employment. They change their names because their online search results include abuse that makes it impossible for them to work or date. They suffer severe anxiety, depression, and PTSD. They spend hours asking sites to take down destructive posts. They check their search results and inboxes with dread. Victims describe the experience as a never-ending nightmare.").

[20] María Luisa Paúl, *Students Made a Racist Deepfake of a Principal. It Left Parents in Fear*, WASH. POST, Mar. 14, 2023.

receiving texts from her child like "I just want to go home. I feel nervous and anxious. Please, could someone pick me up?"[21]

While online harassment generally can cause "anxiety, depression, sadness, anger, fear, shame, embarrassment, isolation, low self-esteem, paranoia, stomach aches, panic attacks, post-traumatic stress disorder (PTSD), self-harming behavior, and heart palpitations,"[22] certain populations experience particular challenges. One study found that exposure to online racial discrimination increased depression and anxiety and reduced confidence in academic abilities among Black and Latino adolescents.[23] The 2022 ADL Online Hate and Harassment survey found that among Black Americans who had experienced online harassment or were worried about future harassment, 22% had trouble sleeping, concentrating, or felt anxious.[24] When someone is targeted for harassment online, they are more likely to self-censor and withdraw from freely expressing themselves on the platform.[25] Bystanders and onlookers are also more likely to self-censor to avoid being targeted themselves.

Black and Latino people are more likely to be harassed online because of their race than white people. A 2021 Pew report found, for example, that 54% of Black respondents and 47% of Latino respondents stated they experienced online harassment because of their race or ethnicity, compared to 17% of white respondents.[26]

A 2023 ADL report found that an AI synthetic speech start up realized that individuals were using the tool to create deepfake audio of celebrities saying hateful rhetoric, such as a character from the video game Halo sharing "graphic instructions for murdering Jewish people and Black people, such as 'toss kikes into active volcanoes' and 'grind n***** fetuses in the garbage disposal.'"[27] The report also revealed a deepfake video of actress Emma Watson reading Adolf Hitler's book

---

[21] María Luisa Paúl, *Students Made a Racist Deepfake of a Principal. It Left Parents in Fear*, WASH. POST, Mar. 14, 2023.

[22] Francesca Stevens, et al., *Cyber Stalking, Cyber Harassment, and Adult Mental Health: A Systematic Review*, 24 CYBERPSYCHOLOGY, BEHAVIOR, AND SOCIAL NETWORKING, 367-376 (June 24, 2021) .

[23] Alvin Thomas, et al., *Taking the good with the bad?: Social Media and Online Racial Discrimination Influences on Psychological and Academic Functioning in Black and Hispanic Youth,* 52 Journal of Youth and Adolescence 245 (2023) (study finding that exposure of Black and Latino adolescents to online racial discrimination increased depression and anxiety, and reduced confidence in academic abilities). *See also* Xiangyu Tao and Celia B. Fisher, *Exposure to Social Media Racial Discrimination and Mental Health among Adolescents of Color*, 51 J. YOUTH ADOLESC. 30 (2022) (finding that "exposure to individual and vicarious social media racial discrimination increased depressive symptoms and drug use problems among" Black, Asian American, Indigenous, and Latinx youth).

[24] *See also* ADL, *Online Hate and Harassment: The American Experience 2022,* ADL 40 (June 20, 2022). https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2022

[25] DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 196-97 (2014) (detailing situations where online hate speech threatens victims into silence); Mary Anne Franks, *Fearless Speech*, 17 FIRST AM. L. REV. 294, 307 (2018) (detailing the chilling effects of online harassment, particularly among marginalized communities); Jon Penney, *Online Abuse, Chilling Effects, and Human Rights*, *in* CITIZENSHIP IN A CONNECTED CANADA: A RESEARCH AND POLICY AGENDA 211-15 (2020) (reviewing studies showing the chilling effects of online harassment and abuse))

[26] *See* Emily A. Vogels, *THE STATE OF ONLINE HARASSMENT,* PEW RESEARCH CENTER (Jan. 13, 2021); *see also* ADL, *Online Hate and Harassment: The American Experience 2022,* ANTI-DEFAMATION LEAGUE (June 20, 2022) (finding that in 2021, 59% of African-Americans surveyed reported they were harassed online because of their race, a sharp increase from 42% the previous year). *See also* ADL, *Online Hate and Harassment: The American Experience 2023,* ADL 19 (June 20, 2023) (finding that Black Americans experienced higher rates of both severe online harassment and any online harassment than other racial or ethnic groups).

[27] ADL, *The Dangers of Manipulated Media and Video: Deepfakes and More*, June 6, 2023.

Mein Kampf, and another deepfake video of former Disinformation Governance Board Executive Director Nina Jankowicz saying "the word disinformation was made up by Jews to define any information that Jews don't like" and describing Jews as a "scourge of evil."[28] Another deepfake video purporting to come from Planned Parenthood denounces "race mixing" and "urges white women to abort biracial fetuses."[29]

IV. Deepfakes Fuel Racial Distrust

Moving forward, deepfakes have the potential to tap into our confirmation biases and fuel social unrest and even violence, particularly "where distrust of certain individuals or communities already exists."[30] For example, it is not difficult to imagine deepfake videos of Black activists committing crimes or calling for violence against the police, or white police officers shooting an unarmed Black person while uttering racial slurs.[31]

Such tactics are already occurring. On the eve of Chicago's mayoral election in February 2023, a tweet from an account called Chicago Lakefront News distributed a likely deepfake video showing moderate "tough-on-crime" mayoral candidate Paul Vallas purportedly saying "'In my day' a police officer could kill as many as 17-18 civilians and 'no one would bat an eye.'"[32] Although the Chicago Lakefront News account was deleted the following day, the tweet was shared by thousands and Vallas lost the election to progressive candidate Brandon Johnson.

Similarly, the pervasiveness of deepfakes could undermine truth and justice by casting doubt on actual video.[33] For example, recent bystander videos of police killings of people like George Floyd have resulted in convictions and calls for policy reform. If these authentic videos are suddenly denounced as deepfakes, the voices of many within marginalized communities are undermined. As Riana Pfefferkorn explains:

---

[28] *Id.*
[29] *Id.*
[30] Mary Anne Franks & Ari Ezra Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD. L. REV. 892, 896 (2019).
[31] *Id.*; *The National Security Challenge of Artificial Intelligence, Manipulated Media, and 'Deep Fakes': Hearing Before the H. Permanent Select Comm. On Intelligence*, 116th Cong. (June 13, 2019) (Testimony of Danielle Citron, Morton & Sophia Macht Professor of Law, University of Maryland Carey School of Law); Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1776, 1780-81 (2019) (detailing various scenarios in which deepfakes could undermine social cohesion).
[32] Joe Concha, *The Impending Nightmare that AI Poses for Media, Elections*, THE HILL, Apr. 23, 2023; Mekela Panditharatne and & Giansiracusa, *How AI Puts Elections at Risk - And the Safeguards Needed*, BRENNAN CENTER FOR JUSTICE, July 21, 2023 ("It wasn't a gaffe, a leak, or a hot-mic moment. It seemingly wasn't even the work of a sly impersonator who had perfected his Paul Vallas impression. The video was a digital fabrication, a likely creation of generative artificial intelligence that was viewed thousands of times.").
[33] Robert Chesney and Danielle Citron refer to this phenomenon as the "liar's dividend." Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1777-78 (2019) ("Imagine a situation in which an accusation is supported by genuine video or audio evidence. As the public becomes more aware of the idea that video and audio can be convincingly faked, some will try to escape accountability for their actions by denouncing authentic video and audio as deep fakes. Put simply: a skeptical public will be primed to doubt the authenticity of real audio and video evidence. . . . Hence what we call the liar's dividend: this dividend flows, perversely, in proportion to success in educating the public about the dangers of deep fakes.").

> If deepfakes cause society to move away from the current 'seeing is believing' paradigm for video footage, that shift may negatively impact individuals whose stories society is already less likely to believe. The proliferation of video recording technology has fueled a reckoning with police violence in the United States, recorded by bystanders and body-cameras. But in a world of pervasive, compelling deepfakes, the burden of proof to verify authenticity of videos may shift onto the videographer, a development that would further undermine attempts to seek justice for police violence.[34]

V. Deepfakes Facilitate Foreign Interference and Voter Suppression in U.S. Elections

In recent U.S. elections, foreign governments and domestic bad actors have targeted communities of color with lower-tech disinformation and voter suppression campaigns.[35] Deepfake video technology will only make these problems worse.

Digital blackface—online racial impersonation of Black people—is popular in part because of its effectiveness in spreading confusion and discrediting authentic movements. According to one study, presenting as a Black activist is the "most effective predictor of disinformation engagement by far."[36]

For example, on Election Day 2016, the operators of the Williams & Kalvin Facebook page — ostensibly two Black men from Atlanta who ran a popular Facebook page focused on Black media and culture — paid for and posted a Facebook ad targeted at Black users. The ad proclaimed: "We don't have any other choice this time but to boycott the election. This time we choose between two racists. No one represents Black people. Don't go to vote."[37]

After the November 2016 election, an investigation revealed that the Williams & Kalvin Facebook account was fake and was set up and operated by the Russian Internet Research Agency (the

---

[34] Riana Pfefferkorn, *The Threat Posed by Deepfakes to Marginalized Communities*, BROOKINGS, Apr. 21, 2021. Pferfferkorn also explains that if video verification technology is developed to authenticate that video is not synthetic but is not included in more affordable smartphones disproportionately used by many people of color, video taken with such smartphones may still be discounted as unreliable and prevent a judicial accountability for unwarranted police violence.

[35] *AI and the Future of Our Elections, Hearing Before S. Committee on Rules and Administration,* 118th Cong. (Sept. 27, 2023) (Testimony of Maya Wiley, President and CEO, The Leadership Conference on Civil and Human Rights) ("Disinformation, sometimes driven intentionally by foreign governments in our election cycles, often targets Black and Latino communities and poses significant risks to our society."); Christine Fernando, *Election Disinformation Targeted Voters of Color in 2020. Experts Expect 2024 to be Worse*, ASSOCIATED PRESS, July 29, 2023.

[36] *See also* Deen Freelon, et al., *Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation*, 40 SOCIAL SCIENCE COMPUTER REV. 560 (2022) (using a computational analysis of 5.2 million tweets by the Russian government-funded "troll farm" known as the Internet Research Agency and separating Black-presenting accounts from non-Black liberal accounts to find that presenting as a Black activist to "be the most effective predictor of disinformation engagement by far.").

[37] YOUNG MIE KIM, PROJECT DATA, UNCOVER: STRATEGIES AND TACTICS OF RUSSIAN INTERFERENCE IN US ELECTIONS: RUSSIAN GROUPS INTERFERED IN ELECTIONS WITH SOPHISTICATED DIGITAL CAMPAIGN STRATEGIES 9 (2018).

"Russian Agency").[38] While African Americans make up just 12.7% of the U.S. population, 37.04% of the unique Facebook pages believed to be created by the Russian Agency were focused on Black audiences,[39] and Black audiences accounted for over 38% of the ads purchased by the Russian Agency, 46.96% of the user impressions, and 49.84% of the user clicks.[40] Although federal law prohibits foreign nationals from spending any money in connection with U.S. elections,[41] the Russian Agency paid Facebook 1,350,489 rubles (about $20,257) for 1,087 different ads for two Black audience segments. The ad campaign resulted in 15,815,597 user impressions (users seeing the ad) and 1,563,584 user clicks (users engaging with the ad).[42]

Russian spending on disinformation targeted at Black voters in the U.S. continued in the 2020 election cycle. Facebook and Twitter acknowledged that they removed a network of Russian-backed accounts that originated in Ghana and Nigeria that posed as being operated by people in the United States (e.g., California, Florida, Louisiana, New York, New Jersey, North Carolina). The accounts attempted to build an audience with Black Americans with posts focusing on Black history, Black excellence, and "content about oppression and injustice, including police

---

[38] *See* Benjamin Fearnow, *Williams & Kalvin: Pro-Trump Facebook Stars Reportedly Worked for Kremlin, Accounts Removed*, INT'L BUS. TIMES (Oct. 10, 2017, 1:51 PM), (noting the "personal" account for Kalvin Johnson last posted in 2015); Issie Lapowsky, *House Democrats Release 3,500 Russia-Linked Facebook Ads*, WIRED (May 10, 2018, 10:00 AM), See also Deen Freelon, Michael Bossetta, Chris Wells, Josephine Lukito, Yiping Xia, and Kirsten Adams, *Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation*, 40 (3) SOCIAL SCIENCE COMPUTER REVIEW 560 (2022) (using a computational analysis of 5.2 million tweets by the Russian government-funded "troll farm" known as the Internet Research Agency to find that presenting as a Black activist to "be the most effective predictor of disinformation engagement by far").

[39] *See* RENEE DIRESTA ET AL., THE TACTICS & TROPES OF THE INTERNET RESEARCH AGENCY 12, 21 (2019), (calculating a total percentage of Black pages at 37.037%, based on numbers indicating that the "Facebook data provided posts from 81 unique pages" (the denominator) and that "[o]verall, 30 targeted Black audiences" (the numerator)); *ACS 2013-2017 Five Year Estimates*, U.S. CENSUS BUREAU (2017), https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_DP05&src=pt [https://perma.cc/YZW7-ETB6] (indicating a Black population in the United States of 12.7%); *see also* PHILIP N. HOWARD ET AL., COMPUTATIONAL PROPAGANDA RESEARCH PROJECT, THE IRA, SOCIAL MEDIA AND POLITICAL POLARIZATION IN THE UNITED STATES, 2012-2018, at 6 (2018) (indicating that Facebook provided data on 3,393 individual ads published from 2015-2017 that it believed originated from the Russian Agency to the U.S. Senate Select Committee on Intelligence, and the U.S. House Permanent Select Committee on Intelligence released details on 3,517 of such ads).

[40] *See* PHILIP N. HOWARD ET AL., COMPUTATIONAL PROPAGANDA RESEARCH PROJECT, THE IRA, SOCIAL MEDIA AND POLITICAL POLARIZATION IN THE UNITED STATES, 2012-2018, at 23 tbl.4 (2018) (providing raw numbers of the twenty audience segments on Facebook targeted by the Russian Agency, including the two audience segments of "African American Politics and Culture" and "Black Identity and Nationalism").

[41] 52 U.S.C. § 30121; 11 CFR 110.20.

[42] *See* PHILIP N. HOWARD ET AL., COMPUTATIONAL PROPAGANDA RESEARCH PROJECT, THE IRA, SOCIAL MEDIA AND POLITICAL POLARIZATION IN THE UNITED STATES, 2012-2018, at 23 tbl.4 (2018) (providing raw numbers of the twenty audience segments on Facebook targeted by the Russian Agency, including the two audience segments of "African American Politics and Culture" and "Black Identity and Nationalism"). Domestic political operatives engaged in a similar plot in the 2020 election that targeted Black voters with robocalls that told voters that if they voted by mail their "personal information will be part of a public database that will be used by police departments to track down old warrants and be used by credit card companies to collect outstanding debts." Charlene Richards, *Robocalls to Voters Before 2020 Election Result in $5 Million Fine*, NBC NEWS, June 8, 2023. *See also* National Coalition on Black Civic Participation v. Wohl, 2023 WL 2403012 (S.D.N.Y., Mar. 8, 2023); Stephanie Saul, *Deceptive Robocalls Try to Frighten Detroit Residents About Voting by Mail*, NY TIMES (Aug. 27, 2020); Ryan J. Foley, *Conservative Hoaxers Face Charges Over False Voter Robocalls*, WASH. POST (Oct. 1, 2020).

brutality."[43] On June 18, 2020, the head of security policy at Facebook testified before Congress that the company disabled 1.7 billion fake accounts between January and March 2020 and had taken down "18 coordinated networks seeking to manipulate public debate, including three networks originating from Russia, two from Iran and two based here in the United States."[44] The Department of Homeland Security emphasized that Russian proxy websites "highlighted reduced in-person polling places in large cities due to the pandemic and the long lines this caused, claiming this would disproportionately suppress voting among African-Americans and expose them to the spread of COVID-19."[45]

In an analysis of 31 posts linked to the Russian Internet Research Agency from late 2019, University of Wisconsin professor Young Mie Kim found that the Russians were impersonating Americans[46] and were targeting "both sides of the ideological spectrum to sow division."[47] The Russian Agency's social media campaigns "exploit sharp political divisions already existing in our society" and "often create an 'us vs. them' discourse, feeding fear to activate or demobilize those who consider an issue personally important."[48] Professor Kim found that the Russian Agency's posts focused on "racial identity/conflicts, anti-immigration (especially anti-Muslim), nationalism/patriotism, sectarianism, and gun rights."[49]

Racial impersonation goes beyond elections. From 2016 to about 2019, a group of domestic non-Black extremists infiltrated a debate within the Black community about #Blaxit (Black people's exit), and set up fake accounts pretending to be Black users.[50] The extremists distributed memes branded in yellow and black designed to mimic Black Lives Matter, created an official Blaxit logo, and took other steps "to create the impression of an emergent movement of African repatriation by a group of Black Americans."[51] As one participant indicated, "[t]his is like catfishing an entire race."[52]

Rather than simply rely on social media account still photos and posts written in ethnic vernacular to impersonate people of color, user-friendly and affordable deepfake technology will allow both foreign governments and domestic bad actors to create realistic-looking synthetic videos of people of color. These nefarious actors will be able to more effectively target communities of color with

---

[43] *See* Clarissa Ward, et. al, *Russian election meddling is back -- via Ghana and Nigeria -- and in your feeds*, CNN (Apr. 11, 2020); Tony Romm and Craig Timberg, *Facebook, Twitter Suspend Russian-linked Operation Targeting African Americans on Social Media*, Wash. Post (March 12, 2020); Taylor Hatmaker, *Russian Trolls Are Outsourcing to Africa to Stoke U.S. Racial Tensions*, TECH CRUNCH (Mar. 12, 2020).

[44] *Online Foreign Influence Operations*, Hearing Before the U.S. House Intelligence Committee (June 18, 2020) (Testimony of Nathaniel Gleicher, the head of security policy at Facebook).

[45] *Id.*

[46] YOUNG MIE KIM, NEW EVIDENCE SHOWS HOW RUSSIA'S ELECTION INTERFERENCE HAS GOTTEN MORE BRAZEN, BRENNAN CENTER (March 5, 2020) ("The IRA . . . mimicked existing names similar to domestic political, grassroots, and community groups, as well as the candidates themselves. . . For example, the IRA mimicked the official account of the Bernie Sanders campaign, "bernie2020," by using similar names like "bernie.2020__".").

[47] *Id.* ("The IRA targets both sides of the ideological spectrum to sow division. This strategy is unique to Russian election campaigns, making it different than conventional persuasion-oriented propaganda or other foreign countries' election interference strategies.")

[48] *Id.*

[49] *Id.*

[50] Brandi Collins-Dexter, *Butterfly Attack: Operation Blaxit*, THE MEDIA MANIPULATION CASEBOOK (Oct. 16, 2020).

[51] *Id.*

[52] *Id.*

racialized disinformation messages to deter voting and fuel racial divisions.[53] Deepfake videos that target non-English speakers with disinformation likely will be particularly difficult for platforms to detect and remove.[54]

VI. Bias in Deepfake Detection Systems

While technologies are being developed to detect deepfakes, initial studies have demonstrated that many of these systems are more accurate in detecting deepfakes featuring whites than people of color.

One study of three popular deepfake detectors by researchers at the University of Southern California found up to a 10.7% difference in error rate depending on gender and race.[55] "In a real-world scenario, facial profiles of female Asian or female African are 1.5 to 3 times more likely to be mistakenly labeled as fake than profiles of the male Caucasian….For large scale commercial applications, this would indicate bias against millions of people."[56] The detectors had the highest error rates on videos with darker Black faces, particularly Black males.[57] Many of the detection tools have higher error rates with people of color because they are not trained on datasets that include a sufficiently robust number of images of people of color.[58] The authors emphasize the importance of representative datasets and auditing for increased transparency and accountability.[59]

---

[53] *Race, Media and Technology*, Shorenstein Ctr. on Media, Pols. and Pub. Pol'y, HARVARD KENNEDY SCHOOL (last visited Jul. 17, 2023) ("A term coined by Joan Donovan and Brandi Collins-Dexter, racialized disinformation refers to media manipulation campaigns that employ the strategic use of falsified racial or ethnic identities, and focus on race as a political wedge issue.").

[54] Aliya Bhatia, *Election Disinformation in Different Languages is a Big Problem in the US*, CTR. FOR DEMOCRACY & TECH., Oct. 18, 2022 (finding "Facebook failed to issue warning labels on 70% of misinformation in Spanish compared to only 29% in English"); The Leadership Conference on Civil and Human Rights and Common Cause, Comment Letter on Public Citizen Petition to Federal Election Commission for Rulemaking on Artificial Intelligence in Campaign Ads (Oct. 19, 2023).

[55] Loc Trinh & Yan Liu, *An Examination of Fairness of AI Models for Deepfake Detection*, ARXIV, May 2, 2021, at 2 (" Using facial datasets balanced by gender and race, we find that classifiers designed to detect deepfakes have large predictive disparities across racial groups, with up to 10.7% difference in error rate."). See Kyle Wiggers, *Deepfake detectors and datasets exhibit racial and gender bias, USC study shows,* VENTUREBEAT, May 6, 2021. *See also* Patrick Hall & Andrew Burt, *Do Deepfakes Discriminate? Auditing a Deepfake Detection System for Systemic Bias*, PHILADELPHIA FEDERAL RESERVE, Oct. 27, 2022 (showing that for every 1000 deepfakes detected with White faces, one expects 694 deepfakes with South Asian faces and 821 deepfakes with Black faces to be detected).

[56] Loc Trinh & Yan Liu, *An Examination of Fairness of AI Models for Deepfake Detection*, ARXIV, May 2, 2021, at 4-5.

[57] *Id.*, at 3 ("All detectors trained with [Blended Image datasets] perform worst on darker faces from the African subgroup, especially male African faces (3.5 - 6.7% difference in error rate)").

[58] *Bias in Facial Recognition is Handicapping Deepfake Detection*, BIOMETRIC UPDATE.COM, May 17, 2021 ("Harmful bias has been found in deepfake datasets and detection models by researchers from the University of Southern California. A commonly used dataset is "overwhelmingly" dominated by white subjects — particularly white females. The result of this skew is that deepfake detectors are less able to spot fraudulent images and video of people of color.").

[59] Loc Trinh & Yan Liu, *An Examination of Fairness of AI Models for Deepfake Detection*, ARXIV, May 2, 2021, at 6-7 (" We found large disparities in predictive performances across races, as well as large representation bias in widely used FaceForensics++. . . . Our work echoes the importance of benchmark representation and intersectional auditing for increased demographic transparency and accountability in AI systems.").

Conclusion

As deepfake technology becomes more ubiquitous in our nation the challenges that accompany deepfakes become more common, women, people of color, and religious minorities bear particular burdens. Not only do these populations disproportionately bear the costs of deepfake technology, but the tools emerging to detect and mitigate the harms of deepfake technology are less effective in protecting people of color. Members of Congress should come together on a bipartisan basis to understand the most significant emerging threats, determine the most constructive role of the federal government in preventing those threats, and take action to protect all Americans from the harms of deepfake technology.