AI For The People

Internet users, not the government or Big Tech, should control the AI to filter online content

Testimony by Michael Shellenberger

Before the Senate Committee on Homeland Security and Governmental Affairs

On the topic of:
"Governing AI Through Acquisition and Procurement"

September 14, 2023

Chairman Peters, ranking member Paul, and Committee members thank you for your stated concern with the implications of AI for our civil liberties and constitutional rights, and for requesting my testimony. I am honored to provide it.

The ability to create deep fakes and fake news through the use of AI is a major threat to democracy, say many experts. "AI-generated images and videos have triggered a panic among researchers, politicians and even some tech workers who warn that fabricated photos and videos could mislead voters, in what a U.N. AI adviser called in one interview the 'deepfake election,'" reported the Washington Post late last month. "The concerns have pushed regulators into action. Leading tech companies recently promised the White House they would develop tools to allow users to detect whether media is made by AI."[1]

But the threat of AI to elections today is as overblown as the threat of Russian disinformation to elections in 2020. Never before has the U.S. been better prepared to detect deep fakes and fake news than we are today. In truth, the U.S. Department of Defense has been developing such tools for decades. In 1999, Defense Advanced Research Applications (DARPA) described its funding for R&D as having the goal of "total situational awareness" through "data mining," "face recognition," and computer networks to evaluate "semantic content." The proposal anticipates the direction of the technology over the following 25 years.[2]

Before elaborating on this point, I want to emphasize that I view AI as a human, not a machine, problem, as well as dual-use technology with the potential for good and bad. My attitude toward AI is the same, fundamentally, as it is toward other powerful tools we have developed, from nuclear energy to biomedical research. With such powerful tools, democratic civilian control and transparent use of these technologies allow for their safe use, while secret, undemocratic, and military control increases the danger. The problem, in a nutshell, is not with the technology of computers attempting to emulate human thinking through algorithms, but rather who will control it and how.

---

[1] Cat Zakrzewski, "ChatGPT breaks its own rules on political messages," Washington Post, August 28, 2023.

[2] J. Brian Sharkey, "Charging Into the Next Millenium: Total Information Awareness," *Accessed via Internet Archive*, June 7-10, 1999.

There is a widespread belief that users already choose their own content on social media platforms. We choose who to follow, and see their posts on the Facebook, X, Instagram, Facebook, and YouTube feeds. In truth, social media platforms decide a significant portion of what users see. YouTube's recommendation algorithm, for example, determines 70% of what people watch on the platform, a share that did not change between 2018[3] and 2022.[4]

The amount of recommended content is lower on other platforms. Meta said last year that just 15% of total Facebook feed content is recommended content from non-followed accounts,[5] while 40 percent of Instagram's feed content is.[6]

But Meta CEO Mark Zuckerberg said last year that he expects Facebook will double the percentage of recommended content by the end of 2023. And users have little to no control over what is recommended to them. In fact, research published in late 2022 found that users have little control over the videos that YouTube feeds them.[7] On every other platform, the algorithms are hidden from users.

The heavy lifting of censorship or "content moderation" was by 2021 done overwhelmingly by AI. Zuckerberg said, "more than 95% of the hate speech that [Facebook] take[s] down is done by an AI [artificial intelligence] and not by a person. . . . And I think it's 98 or 99% of the terrorist content that we take down is identified by an AI and not a person."[8] Similarly, 99% of Twitter's content takedowns started with machine learning.[9]

The problem with AI technology today funded by the US government, whether DARPA or National Science Foundation (NSF), is fundamentally around the control of these technologies by small groups of individuals and institutions remarkably unaccountable to the citizens of the United States. While there is always a diversity of

---

[3] Ashley Rodriguez, "YouTube's algorithms drive 70% of what we watch," *QZ*, July 13, 2018.

[4] Hana Kiros, "Hated that video? YouTube's algorithm might push you another just like it," *MIT Tech Review*, September 20, 2022.

[5] Meta, Q2 2022 Earnings, July 27, 2022.

[6] Rachael Davies, "Nearly half of the posts you see on Instagram are from accounts you don't follow," *Evening Standard,* April 28, 2023.

[7] Hana Kiros, "Hated that video? YouTube's algorithm might push you another just like it," *MIT Tech Review*, September 20, 2022

[8] Feerst, Alex. "The Use of AI in Online Content Moderation" *Digital Governance Working Group*, Sept. 2022. (p. 2.)

[9]Kristen Ruby, "Twitter Artificial Intelligence," *Ruby Media Group,* December 26, 2022.

agendas and motivations behind what decision-makers in the AI space are doing, many U.S. government-funded individuals and institutions behind deep fake alarmism are, not coincidentally, demanding greater governmental or nongovernmental control over social media platforms and Internet companies.

Why is that? Why have elements within the US government promoted AI for online censorship? And can AI be used to advance free speech and free expression instead?

AI and the Censorship Industrial Complex

This Censorship Industrial Complex of government agencies and government contractors has its roots in the war on terrorism and the expansion of surveillance after 9/11. President George W. Bush that year authorized the National Security Agency to monitor Americans who were suspected of having a 'nexus to terrorism,' resulting in the Agency's now-infamous and illegal interception of information."[10] In 2003 DARPA told Congress that NSA was its "experimental partner" using [Total Information Awareness (TIA)] and AI to detect false information.[11] Ten years later, in 2013, a US military contractor named Edward Snowden revealed to reporters that the NSA was collecting telephone records of millions of Verizon customers,[12] and accessing Google and Facebook to secretly collect data.[13]

During the same period, the U.S. intelligence community (IC) and DOD alike recognized how essential AI would become to their operations overall. In 2013, a New York Times report on the NSA's use of AI foreshadowed how "counter-disinformation" experts would, nearly a decade later, describe fighting misinformation online.[14] "Computers could instantly sift through the mass of Internet

---

[10] Scott Shane, "Giving In to the Surveillance State," *New York Times*, August 22, 2012.

[11] DARPA, "Report to Congress Regarding the Terrorism Information Awareness Program," *DARPA Information Awareness Office*, May 20, 2003.

[12] Glenn Greenwald, "NSA collecting phone records of millions of Verizon customers daily," *The Guardian*, June 6, 2013.

[13] Barton Gellman &Laura Poitras, "U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program," *The Washington Post*, June 7, 2013.

[14] James Risen & Eric Lichtblau, "How the U.S. Uses Technology to Mine More Data More Quickly," *The New York Times*, June 8, 2013.

communications data," reported the Times, "see patterns of suspicious online behavior and thus narrow the hunt for terrorists." In 2014, the DOD unveiled its "Third Offset Strategy," which emphasized that AI would change how the US prepared for cyberwar with China and Russia.[15]

In 2015, DARPA launched the funding track that directly resulted in the AI tools that leading Internet and social media companies use today. That fall, DARPA invited proposals for its MediFor program.[16] The goal? Develop a science and practice for "determining the authenticity and establishing the integrity of visual media."[17] DARPA funded universities to create the MediFor platform to automatically detect manipulations.[18]

DARPA's warning eight years ago is identical to the Washington Post's warning about deep fake last month. "Mirroring this rise in digital imagery is the associated ability for even relatively unskilled users to manipulate and distort the message of the visual media," warned DARPA. "While many manipulations are benign, performed for fun or for artistic value, others are for adversarial purposes, such as propaganda or misinformation campaigns."

The adoption of AI grew alongside alarmism about deep fakes and "misinformation," and "disinformation" more broadly. In 2016, Facebook reported it had developed AI to automatically censor offensive live videos.[19] In early January 6, 2017, outgoing Obama Administration DHS Secretary Jeh Johnson designated "election infrastructure" as "critical infrastructure," which would become the mandate of the Cybersecurity and Infrastructure Security Agency (CISA), which Congress created the following year to protect. In 2018, journalists revealed that Facebook was using AI to predict users' future actions for advertisers.[20]

---

[15] Gentile et al., A History of the Third Offset, 2014–2018, *Rand Corporation*, 2021.

[16] Dr. William Corvey, Media Forensics (MediFor) (Archived), *darpa.mil*, nd.

[17] Media Forensics (MediFor) Grant DARPA-BAA-15-58, *grants.gov*, September 29, 2015.

[18] Contractors included Notre Dame, Purdue University, Duke University, Ideal Innovations Inc, Schaefer Corporation, University of Siena, New York University, University of Southern California, Politecnico di Milano, Unicamp, NVIDIA, Columbia University, Dartmouth, University of Albany, UC Berkeley, and Kitware.

[19] Kristina Cooke, "Facebook developing artificial intelligence to flag offensive live videos," *Reuters*, December 1, 2016.

[20] Sam Biddle, "Facebook uses artificial intelligence to predict your future actions for advertisers, says confidential document," *The Intercept*, April 13, 2018.

In 2019, DARPA launched "Semantic Forensics," the successor to Medifor. SemaFor funded think-tanks, academic institutions, software companies, social media, and search engine organizations as part of a four-year project to develop AI meant to detect deep fakes, or synthetic or manipulated media.[21] It gave contracts to five primary organizations: Kitware, PAR Government, STR, Lockheed Martin, and SRI International, with this financing further divided amongst other universities and research institutes.

Commercial interests in both policing deep fake and advocating policies to censor synthetic media popped up during this period. Also in 2019, a new nongovernmental organization called The "DeepTrust Alliance" launched a series of events called the "Fix Fake Symposia."[22] The DeepTrust Alliance described itself as "the ecosystem to tackle disinformation," and its website invited audiences to "Join the global network actively driving policy and technology to confront the threat of malicious deep fakes and disinformation."[23]

The goal of Deep Trust appeared to be to advocate for policies aimed at criminalizing "digital harms," including forms of speech that hurt people. "If the behavior is malicious," said the group's CEO, Kathryn Harrison, in 2020, "that's a problem. Laws need to be extended to digital harms… There needs to be a standard set of practices" across social media platforms.[24] "I want to see society put more safeguards in place," she said. "This is like cars, right? When you first had cars, you didn't have seat belts…. We're in a very similar situation in the media ecosystem and can save information at light speed but no safety net. That's what we need to build."

It was also in 2020 that DHS' CISA created an "Election Integrity Partnership" to censor election skepticism. It partnered with four groups: Graphika, the University of Washington, the Atlantic Council's DFR Lab, and the Stanford Internet Observatory. Graphikia and UW are DARPA's Semafor grantees. In Deep Trust's report, it names those four groups and progressive philanthropic donors, and other NGOs and government. EIP claims it classified 21,897,364 individual posts

---

[21] Semantic Forensics (SemaFor) Grant HR001119S0085, *sam.gov*, November 19, 2019.

[22] Aros Harrinson, "Deepfake, Cheapfake: The Internet's Next Earthquake?"*Fix Fake Symposium Proceedings Part 1*, 2020.

[23] DeepTrust Alliance, Homepage, *deeptrustalliance.org*, nd.

[24]Jon Prial & Kathryn Harrison, "Episode 133: Tackling Digital Disinformation with Kathryn Harrison," *Georgian Impact Podcast*, December 11, 2020.

comprising unique "misinformation incidents" from August 15, 2020, to December 12, 2020, from a larger 859 million set of tweets connected to"misinformation narratives."[25]

By January of 2021, CISA unilaterally broadened its scope "to promote more flexibility to focus on general" misinformation, disinformation, and malinformation. Where misinformation can be unintentional, disinformation is defined as deliberate, while malinformation can include accurate information that is "misleading." Two months later, DARPA announced that it had funded Accenture Federal Services (AFS), Google/Carahsoft, New York University (NYU), NVIDIA, and Systems & Technology Research (STR) to "develop automated tools that aid analysts as they tackle the looming rise of automated multimodal media manipulation," otherwise known as deep fakes or fake news.[26]

While social media platforms use AI to identify and censor content, the decisions of what to censor, and how remain in the hands of humans, specifically executives at social media platforms. And so those individuals and groups that wished to see greater censorship by social media platforms rolled out a major initiative in the spring of 2022 to establish a US government agency to do precisely that. In April, DHS announced that it had created a "Disinformation Governance Board," ostensibly to protect national security by fighting disinformation, misinformation, and malinformation on social media.[27] One week earlier, former U.S. President Barack Obama gave a speech at Stanford calling for government regulation of online speech with the same justification as Deep Trust's Kathryn Harrison: preventing harm and protecting democracy.

One month later, in May of 2022, DARPA launched its "Model Influence Pathways," or MIP, program to automate the process of discovering the origins and "pathways" of "misinformation, disinformation, and manipulated information."[28] The

---

[25] UW Center for an Informed Public, Digital Forensic Research Lab, Graphika, and Stanford Internet Observatory, "The Long Fuse: Misinformation and the 2020 Election," *Stanford Digital Repository: Election Integrity Partnership*, 2021.

[26]Matt Turek, "DARPA Announces Research Teams Selected to Semantic Forensics Program," *darpa.mil*, March 2, 2021.

[27] Amanda Seitz, "Disinformation board to tackle Russia, migrant smugglers," *AP*, April 28, 2022.

[28] Dr. Brian Kettler, Model Influence Pathways (MIP), *darpa.mil*, May 4, 2022.

goal of the program appears to be to develop tools so social media companies can reduce the virality or spread of disfavored social media posts. In that sense, it is within the vision of Stanford Internet Observatory's leader, Renee Diresta, who has long championed simply reducing the spread of disfavored views, rather than removing them from platforms outright. Preventing virality delivers most of the benefits of outright censorship with the benefit of not being noticed and thus not triggering the Streisand effect.[29]

The Federal Trade Commission in June of last year warned Congress about the dangers of using AI for censorship and urged "great caution." Good intentions weren't enough, said FTC, because "it turns out that even such well-intended AI uses can have some of the same problems — like bias, discrimination, and censorship — often discussed in connection with other uses of AI."[30] The FTC specifically pushed back against the idea, widely promoted by individuals and institutions within the Censorship Industrial Complex, that AI should be used to reduce harm. Noted the report authors, "while some harms refer to content that is plainly illegal, others involve speech protected by the First Amendment."

The FTC's warning was well-timed. Six months later, the Twitter Files would reveal Twitter executives over-ruling the determination by their own Trust and Safety team that President Donald J. Trump's tweets had not incited violence, but they deplatformed him anyway, under both external societal pressure and internal employee pressure. Shortly after, emails revealed White House staff demanding that Facebook executives censor "often-true" information about COVID-19 vaccine side effects under explicit or implicit financial threats, behaviors which the Fifth Circuit Court of Appeals last week ruled were unconstitutional.[31]

Both the Twitter and Facebook files exposed the large involvement, influence over, and infiltration by former government intelligence and security officials. "Facebook currently employs at least 115 people, in high-ranking positions, that formerly worked at FBI/CIA/NSA/DHS," noted an analyst. "17 CIA, 37 FBI, 23 NSA,

---

[29] Michael Shellenberger, "Why Renee Diresta Leads the Censorship Industry," *Public.Substack.com*, April 3, 2020.

[30] Federal Trade Commission, Combatting Online Harms Through Innovation, *Report to Congress*, June 16, 2022.

[31] Michael Shellenberger, "War on Free Speech War On Free Speech Means Social Media Users Must Be Free To Moderate Their Content," *Public*, September 9, 2023.

38 DHS."[32] This influence may carry over to today's people seeking to rescue, ostensibly independently, the legitimacy of the US government, which sits at the intersection of technology and foreign policy. Harrison, for example, worked in the French Ministry of Defense, received a graduate degree from Georgetown, and was a term member at the Council on Foreign Relations before working with IBM on AI and then founding Deep Trust.[33]

Why have elements within the US government promoted AI for online censorship? Part of the reason is a well-intentioned concern over real-world harm, and undermining of liberal democracy. But another part of it appears to stem from an inappropriate and exaggerated sense of entitlement by DARPA contractors to work with social media platforms to censor disfavored voices.

User-Based Content Moderation

The Fifth Circuit Court ruling showed the limits of the First Amendment to protect free speech online. The judges ruled that the Cybersecurity and Infrastructure Security Agency (CISA) of the Department of Homeland Security had likely not violated the First Amendment in creating an elaborate system for "flagging" content for Facebook, Twitter, and other social media platforms to censor. The court suggested that such mass flagging operations may be constitutionally protected free speech, at least if done right.

I believe that the way CISA used AI to mass-flag so-called "Covid misinformation" in 2021, through its partnership with "The Virality Project," created by Stanford Internet Observatory (SIO) and others, was a government infringement on freedom of speech. Through such mass flagging, CISA indirectly demanded that Twitter and Facebook censors "often true" information about vaccine side effects. We believe that, with Biden simultaneously threatening the Section 230 legal status of the social media platforms, having CISA's partners make their demands constituted coercion.

---

[32] @nameredacted, twitter.com/NameRedacted247/status/1604641866342756352?s=20, X, December 18, 2022, 4:56 PM.

[33] Kathryn Ann Harrison *Experiences*. LinkedIn. Retrieved September 11, 2023.

But I also recognize that the Fifth Circuit court is saying that such AI-supported mass flagging by "government partners" like SIO could be constitutionally protected if it did not involve coercion or, on the flip side, any incentive to cooperate. The First Amendment prevents the government from "abridging" or limiting speech. It doesn't prevent government officials from telling publishers, whether of books, news articles, or social media posts, that, in their opinion, they shouldn't be publishing those books, articles, or social media posts. The line the Circuit Court wants to draw is on relatively direct and obvious coercion, not jawboning.

Whether or not the Supreme Court decides to hear the case and draw the line somewhere else, the ruling points to the need for Congress to take action to protect freedom of speech by defunding government contractors that advocate widened censorship by social media platforms, and exercising greater oversight over contractors developing AI tools.

The threat to our civil liberties comes not from AI but from the people who want to control it and use it to censor, rather than let users control, information. The obvious solution is for Congress to require that social media companies allow users to moderate their own content in exchange for Section 230's sweeping liability protections, which allow them to exist. This specific suggestion is something another committee will need to consider.

What this committee can consider is a related FTC recommendation, which is using the power of procurement to put AI tools in the hands of users, not the hands of big tech companies. "Filters that enable people, at their discretion, to block certain kinds of sensitive or harmful content are one example of such user tools," FTC notes. The way these tools work should be transparent; users should have a right to know how these tools work. Giving users control over what content they see and don't see is the solution most consistent with the American tradition of free of speech.

Users should be able to decide for themselves whether or not to use these filters and other tools, not Internet companies, the government, a nongovernmental organization, or anyone else. Some tools are already becoming available. Microsoft a Video Authenticator in 2020, while Adobe's Content Credentials allow users to detect whether the content is likely to be authentic and unaltered. Requiring people to affirmatively choose their filters will require more reflective and slow thinking about their content choices.

FTC errs in suggesting that Congress give government-certified researchers, rather than users, access to the algorithms and content moderating filters. A longstanding goal of its leaders is to allow US government-certified researchers to gain access to the data of social media platforms so they can then demand censorship of disfavored views behind closed doors. This is what the "Platform Accountability and Transparency Act," which Obama endorsed, would do. It would allow "researchers" to act as de facto censors. Such activities may be constitutional, but they are antithetical to the values of transparency, privacy, and free speech.

Finally, this committee should seek to encourage or even mandate that DARPA contractors be required to share their research in a more visible way, and stand for questions from the general public. Of the roughly 60 organizations, many if not most of which have been funded by the US government to fight "mis- and dis-information," that my colleagues and I emailed in the spring, none agreed to stand for an interview.[34] The refusal to speak to the public is an odd behavior from those whose livelihoods depend on the goodwill of the public. Congress should consider some provision whereby contractor recipients of taxpayer money must expose themselves to scrutiny.

At the same time, deep fakes and other forms of synthetic media are new, deception, disinformation, and misinformation are not. One of the oft-repeated claims of those advocating expanded online censorship is that, by allowing falsehoods to go viral and undetected, the Internet poses a heretofore unanticipated threat. But the same thing was said about the Gutenberg printing press, the radio, and television. The solution today, as then, is for users to correct misinformation with good information, for themselves, not other people.

None of the above information is likely to put an end to the alarmism about the threat to democracy from deep fakes and AI. But it may help expose much of it as coming from individuals and institutions with an interest in exploiting the alarmism for personal or political gain.

---

[34] Matt Taibbi, "Report on the Censorship-Industrial Complex," *Racket News,* April 25, 2023.