Written Testimony of David Evan Harris

U.S. Senate Committee on the Judiciary

Subcommittee on Privacy, Technology, & the Law

September 17th, 2024

Written Testimony of David Evan Harris

U.S. Senate Committee on the Judiciary
Subcommittee on Privacy, Technology, & the Law
September 17th, 2024

Chairman Blumenthal, Ranking Member Hawley, and Members of the Subcommittee:

It is an honor to appear before you today to discuss the harms and risks of artificial intelligence ("AI"). It is particularly heartening to see that this Committee has developed a promising framework and members have written bipartisan legislative proposals[1] that I earnestly hope will become law, as they are urgently needed to provide effective guardrails for AI systems.

My name is David Evan Harris. From 2018 to 2023, I worked at Facebook and Meta on the Civic Integrity and Responsible AI teams. In my role, I helped lead efforts to combat online election interference, protect public figures, and drive research to develop ethical AI systems and AI governance. Today, those teams do not exist.

In the past two years, there have been striking changes across the industry. Trust and safety teams have shrunk dramatically. Secrecy and lack of transparency have increased.

Since working at Meta, my priority has been the passage of binding and enforceable AI regulation.

In addition to teaching AI ethics and civic technology at UC Berkeley's Haas School of Business and serving as an advisor to the Brennan Center for Justice at NYU Law, I serve as Senior Policy Advisor to the California Initiative for Technology and Democracy.[2] In that capacity, I helped craft two bills about deepfakes and

[1] Sen. Blumenthal, Richard and Sen. Josh Hawley. "Bipartisan Framework for U.S. AI Act." September, 2023. https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf; and Sen. Klobuchar, Amy, Sen. Josh Hawley, Sen. Chris Coons, Sen. Susan Collins. "S.2770 - Protect Elections from Deceptive AI Act." September, 2023 https://www.congress.gov/bill/118th-congress/senate-bill/2770/text; and Sen. Durbin, Dick, Sen. Lindsey Graham, Sen. Amy Klobuchar, Sen. Josh Hawley. "S. 3696 - Disrupt Explicit Forged Images and Non-Consensual Edits Act of 2024 (DEFIANCE Act)." January, 2024. https://www.congress.gov/bill/118th-congress/senate-bill/3696/text; and Sen. Coons, Chris, Sen. Marsha Blackburn, Sen. Amy Klobuchar, Sen. Thom Tillis. "S. 4875 - Nurture Originals, Foster Art, and Keep Entertainment Safe (No Fakes) Act." https://www.congress.gov/bill/118th-congress/senate-bill/4875/text; and Sen. Thune, John, Sen. Amy Klobuchar. "S.3312 - Artificial Intelligence Research, Innovation, and Accountability Act of 2023." November, 2023. https://www.congress.gov/bill/118th-congress/senate-bill/3312/text; and Sen. Hawley, Josh, Sen. Richard Blumenthal. "S.1993 - A bill to waive immunity under section 230 of the Communications Act of 1934 for claims and charges related to generative artificial intelligence." June, 2023. https://www.congress.gov/bill/118th-congress/senate-bill/1993/text.

[2] California Initiative for Technology & Democracy, accessed September 16, 2024, https://cited.tech/.

elections in California that are currently awaiting the governor's signature. The Defending Democracy from Deepfake Deception Act and the Elections: Deceptive Media in Advertisements Act offer a framework for sensible regulation for AI in elections at the state and federal levels.

In addition to California, I have advised policymakers and leaders on AI policy across state, federal and international jurisdictions, including at the United Nations, NATO, the European Union, UK, Singapore, Arizona, California, the White House, and, of course, here in Congress.

My work inside the tech industry, and since then, working closely with policymakers in California, Arizona, and internationally, has strengthened my belief that not only do we urgently need effective oversight of AI, but that it is possible to achieve.

Today, I will focus on three areas that I believe you should consider:

> First, voluntary self-regulation does not work;
> Second, the solutions for AI safety and fairness exist in the framework and bills proposed by the members of the committee; and,
> Third, not all the horses have left the barn. There is still time.

First, voluntary self-regulation is a myth. We have seen repeatedly in tech that it does not work. Take just one example from my time at Facebook. In 2018, the company set out to make time on their platforms into "time well spent," reducing the number of viral videos and increasing content from friends and family. The voluntary policy opened up a vacuum that TikTok happily stepped into. Today, Facebook and Instagram are fighting to claw back market share from TikTok with reels—essentially those very viral videos that they sought to diminish.

When one tech company tries to be responsible, another, less responsible one steps in to fill the void.

If you need further evidence that voluntary self-regulation won't work, just look at the multitude of voluntary agreements that have been signed in the past year, including the White House Voluntary AI Commitments signed by the biggest AI developers last July.[3] These are positive steps. But not enough.

The best reason not to trust AI companies to self-regulate, is that they don't even trust themselves. Most of the CEOs from the largest AI companies are publicly calling for government regulation.

[3] White House, "Voluntary AI Commitments," September 2023, https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf; and White House, "Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," July 21, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/#:~:text=As%20part%20of%20this%20commitment,help%20move%20toward%20safe%2C%20secure%2C.

But beware. While these leaders may want regulation, they often do not want *that regulation.* I have seen time and again Big Tech's lobbyists come in at the last minute and change all the "shalls" to "mays". Without the "shalls," the legislation becomes voluntary. And that's only one of the many tricks up their sleeves. Unbiased technical and legal advisors are essential to combat their dilution and evisceration of your legislation. We are out there. Use us.

Second, we don't need silver bullets. This committee's framework already has many of the answers. Two recommendations in particular are essential components for regulation: AI companies should be held liable for their products, and embedding hard-to-remove provenance data in AI-generated content should be required. It is encouraging to see Senate Bill 3875 on transparency in elections - a bill that has passed out of committee - includes a disclaimer requirement for some federal election advertising that has been substantially AI-generated. More steps like this are needed.

Which brings me to my final point: the horses have not left the barn. The misconception is that it is too late to do anything. For all of the dizzyingly fast releases of AI voice and image deepfakes, there are many more uses of the technology that have not yet been released.

Realistic audio and video deepfakes in real time. Hundreds of thousands of independently-operating AI chatbots that can propagate malicious viewpoints. Automated sextortion schemes. Unchecked AI decision-making that determines job acceptances, loan approvals, insurance prices, and prison sentences. These are just the ones we know about.

We need to move quickly with meaningful regulations on AI. Though certainly a challenge, it is possible. But the time is now. Do not be fooled by the ponies who have exited the barn. Take action now with the promising framework and bills that you have already proposed. You still have a chance to rein in the Clydesdales and centaurs waiting behind the barn door.

Thank you.

APPENDICES

a. Quotes from AI CEOs calling for regulation
b. New York Times: Inside the A.I. Arms Race That Changed Silicon Valley Forever
c. Centre for International Governance Innovation: Not Open and Shut: How to Regulate Unsecured AI
d. IEEE Spectrum: AI Image Generators Make Child Sexual Abuse Material (CSAM) - IEEE Spectrum
e. AI companies promised to self-regulate one year ago. What's changed? | MIT Technology Review
f. Chairman Warner Shares Responses from AI Companies on Efforts to Crack Down on Malicious Use - Press Releases
g. Graphika: A Revealing Picture AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business
h. Free Press: Big Tech Backslide: How Social-Media Rollbacks Endanger Democracy Ahead of the 2024 Elections
i. CAIDP: Still no Guardrails
j. Brennan Center for Justice: Safeguards for Using Artificial Intelligence in Election Administration

Selected Quotations from CEOs Calling For AI Regulation

Sam Altman, CEO of OpenAI

- “I think if this technology goes wrong, it can go quite wrong. And we want to be vocal about that,” he said. “We want to work with the government to prevent that from happening.”[4]

Satya Nadella, CEO of Microsoft

- “I think [a global regulatory approach to AI is] very desirable, because I think we’re now at this point where these are global challenges that necessitate global norms and standards.”[5]

Sundar Pichai, CEO of Alphabet

- “AI is too important not to regulate, and too important not to regulate well.”[6]

Dario Amodei, CEO of Anthropic

- “In our view these concerns merit an urgent policy response. The ideal policy response would address not just the specific risks we’ve identified above, but would at the same time provide a framework for addressing as many other risks as possible – without, of course, hampering innovation more than is necessary...It is worth stating clearly that given the current difficulty of controlling AI systems even where safety is prioritized, there is a real possibility that these rigorous standards would lead to a substantial slowdown in AI development, and that this may be a necessary outcome.”[7]

Mark Zuckerberg, CEO of Meta

- “So I agree that Congress should engage with AI to support innovation and safeguards. This is an emerging technology, there are important equities to balance here, and the government is ultimately responsible for that...We think policymakers, academics, civil

[4] Cecilia Kang, “OpenAI’s Sam Altman Urges A.I. Regulation in Senate Hearing,” *New York Times*, May 16, 2023, https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

[5] Linda Lacina, “Davos 2024: A conversation with Satya Nadella, World Economic Forum, January 16, 2024, https://www.weforum.org/podcasts/meet-the-leader/episodes/davos-2024-conversation-microsoft-satya-nadella/#:~:text=Satya%20Nadella%3A%20I%20think%20it%27s,core%20research%2C%20that%20is%20needed.

[6] Sundar Pichai, “Google CEO: Building AI responsibly is the only race that really matters,” *Financial Times*, May 23, 2023, https://www.ft.com/content/8be1a975-e5e0-417d-af51-78af17ef4b79.

[7] United States Senate Subcommittee on Privacy, Technology, and the Law, “Written Testimony of Dario Amodei, Ph.D.,” July 25, 2023, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

society and industry should all work together to minimize the potential risks of this new technology, but also to maximize the potential benefits."[8]

Jensen Huang, CEO of NVIDIA

- "We have to develop the technology safely, we have to apply the technology safely, and we have to help people use the technology safely…Whether it's the plane that I came in, cars or medicine, all of these industries are heavily regulated today."[9]

Arvind Krishna, CEO of IBM

- "As long as it is light touch and allows innovation to happen, I absolutely will be pro regulation."[10]

Aidan Gomez, CEO of Cohere

- "I would love to see clear guardrails and guidelines. Right now the state of regulation ... the picture is so fuzzy. I think it should be a human right to know whether the content you're consuming is machine-generated or human-generated."[11]

Kevin Baragona, CEO of DeepAI

- "If we can set clear guidelines against the use of deep fakes to steal face and voice of people without permission, and similar regulation against other types of misinformation would be a great start."[12]

Marc Benioff, CEO of Salesforce

- "I think the government is going to have to really step it up and really offer another level of regulation…They get an F for social media and how they've handled them. We've seen that play out. I hope that in AI they really step up and do the right thing."[13]

[8] Meta, "Mark Zuckerberg's Remarks at AI Forum," September 13, 2023, https://about.fb.com/news/2023/09/mark-zuckerbergs-remarks-at-ai-forum/.

[9] Shreyas Sinha, "How OpenAI's Sam Altman And Nvidia's Jensen Huang Think A.I. Should Be Regulated," *Observer*, February 16, 2024, https://observer.com/2024/02/nvidia-jensen-huang-openai-sam-altman-ai-regulation/.

[10] *Bloomberg*, "IBM CEO on Future of AI, Regulation," May 6, 2024, https://www.bloomberg.com/news/videos/2024-05-06/ibm-ceo-on-future-of-ai-regulation#:~:text=As%20long%20as%20it%20is,I%20think%20that%27s%20a%20problem.

[11] Kia Kokalitcheva, "What They're Saying: Cohere CEO Aidan Gomez," *Axios*, July 2, 2023, https://www.axios.com/2023/07/01/cohere-ceo-aiden-gomez-interview.

[12] Ian Krietzberg, "DeepAI Founder and CEO Says a Sci-Fi Future Is On Its Way," *TheStreet*, June 25, 2023, https://www.thestreet.com/technology/deepai-founder-and-ceo-says-a-sci-fi-future-is-on-its-way.

[13] Brian Sozzi, "Govenrment Must 'Step it up' on Regulating AI: Salesforce CEO Marc Benioff," *Yahoo Finance*, September 19, 2023, https://finance.yahoo.com/news/government-must-step-it-up-on-regulating-ai-salesforce-ceo-marc-benioff-

The New York Times:

Inside the A.I. Arms Race That Changed Silicon Valley Forever

ChatGPT's release a year ago triggered a desperate scramble among tech companies and alarm from some of the people who helped invent it.

By Karen Weise, Cade Metz, Nico Grant and Mike Isaac

At 1 p.m. on a Friday shortly before Christmas last year, Kent Walker, Google's top lawyer, summoned four of his employees and ruined their weekend.

The group worked in SL1001, a bland building with a blue glass facade betraying no sign that dozens of lawyers inside were toiling to protect the interests of one of the world's most influential companies. For weeks they had been prepping for a meeting of powerful executives to discuss the safety of Google's products. The deck was done. But that afternoon Mr. Walker told his team the agenda had changed, and they would have to spend the next few days preparing new slides and graphs.

In fact, the entire agenda of the company had changed — all in the course of nine days. Sundar Pichai, Google's chief executive, had decided to ready a slate of products based on artificial intelligence — immediately. He turned to Mr. Walker, the same lawyer he was trusting to defend the company in a profit-threatening antitrust case in Washington, D.C. Mr. Walker knew he would need to persuade the Advanced Technology Review Council, as Google called the group of executives, to throw off their customary caution and do as they were told.

It was an edict, and edicts didn't happen very often at Google. But Google was staring at a real crisis. Its business model was potentially at risk.

153810082.html?guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABt3MeXk2ZPB8xli6lcnDkkFM_T5Bkx-XePxDbOe9W8coOoQWd6JuCkIfxmsCPcFt3O3UitdQbwphwbVrJyFNAmGstzMRGr_9y8xQRNZzOsPnKmZil6cH-TjWNkU6xZaIuu0JOODNnXosiOgAcyiNFreAGfZvAiwWI5Bb1WzuTjZ&guccounter=2.

What had set off Mr. Pichai and the rest of Silicon Valley was ChatGPT, the artificial intelligence program that had been released on Nov. 30, 2022, by an upstart called OpenAI. It had captured the imagination of millions of people who had thought A.I. was science fiction until they started playing with the thing. It was a sensation. It was also a problem.

At the Googleplex, famed for its free food, massages, fitness classes and laundry services, Mr. Pichai was also playing with ChatGPT. Its wonders did not wow him. Google had been developing its own A.I. technology that did many of the same things. Mr. Pichai was focused on ChatGPT's flaws — that it got stuff wrong, that sometimes it turned into a biased pig. What amazed him was that OpenAI had gone ahead and released it anyway, and that consumers loved it. If OpenAI could do that, why couldn't Google?

Why not plow ahead? That's the question that loomed over A.I.'s adolescence — the year or so after the technology made the leap from lab to living room. There was hand-wringing over chatbots writing seductive phishing emails and spewing disinformation, or high schoolers using them to cheat their way to an A. Doomsayers insisted that unfettered A.I. could lead to the end of humankind.

For tech company bosses, the decision of when and how to turn A.I. into a (hopefully) profitable business was a more simple risk-reward calculus. But to win, you had to have a product.

By Monday morning, Dec. 12, the team at SL1001 had a new agenda with a deck labeled "Privileged and Confidential/Need to Know." Most attendees tuned in over videoconference. Mr. Walker started the meeting by announcing that Google was moving ahead with a chatbot and A.I. capabilities that would be added to cloud, search and other products.

"What are your concerns? Let's get in line," Mr. Walker said, according to Jen Gennai, the director of responsible innovation.

There would be guardrails, but approvals would be fast-tracked. Mr. Walker called it the "green lane" approach. It was all laid out in the deck. Opportunities for "Green Lane streamlining" were identified. Dangers were color-coded. Blue indicated risks where "mitigations" were "required." Risks that were "controllable with minimum thresholds/mitigations" were rendered in orange.

In one chart, under "Hate & Toxicity," the plan was to "curb stereotypes, toxicity and hate speech in outputs." One topic was: "What are we missing in order to fast-track approvals?"

Not everyone was on board. "My standards are as high if not higher than they usually are, and we will be going through a review process with all of this," Ms. Gennai remembered a cloud executive saying.

Eventually a compromise was reached. They would limit the rollout, Ms. Gennai said. And they would avoid calling anything a product. For Google, it would be an experiment. That way it didn't have to be perfect. (A Google spokeswoman said the A.T.R.C. did not have the power to decide how the products would be released.)

What played out at Google was repeated at other tech giants after OpenAI released ChatGPT in late 2022. They all had technology in various stages of development that relied on neural networks — A.I. systems that recognized sounds, generated images and chatted like a human. That technology had been pioneered by Geoffrey Hinton, an academic who had worked briefly with Microsoft and was now at Google. But the tech companies had been slowed by fears of rogue chatbots, and economic and legal mayhem.

Once ChatGPT was unleashed, none of that mattered as much, according to interviews with more than 80 executives and researchers, as well as corporate documents and audio recordings. The instinct to be first or biggest or richest — or all three — took over. The leaders of Silicon Valley's biggest companies set a new course and pulled their employees along with them.

Over 12 months, Silicon Valley was transformed. Turning artificial intelligence into actual products that individuals and companies could use became the priority. Worries about safety and whether machines would turn on their creators were not ignored, but they were shunted aside — at least for the moment.

At Meta, Mark Zuckerberg, who had once proclaimed the metaverse to be the future, reorganized parts of the company formerly known as Facebook around A.I.

Elon Musk, the billionaire who co-founded OpenAI but had left the lab in a huff, vowed to create his own A.I. company. He called it X.AI and added it to his already full plate.

Satya Nadella, Microsoft's chief executive, had invested in OpenAI three years before and was letting the start-up's cowboys tap into its computing power. He sped up his plans to incorporate A.I. into Microsoft's products — and give Google a poke in its searching eye.

"Speed is even more important than ever," Sam Schillace, a top executive, wrote Microsoft employees. It would be, he added, an "absolutely fatal error in this moment to worry about things that can be fixed later."

Image

A 'Low Key Research Preview'

The strange thing was that the leaders of OpenAI never thought ChatGPT would shake up Silicon Valley. In early November 2022, a few weeks before it was released to the world, it didn't really exist as a product. Most of the 375 employees working in their new offices, a former mayonnaise factory, were focused on a more powerful version of technology, called GPT-4, that could answer almost any question using information gleaned from an enormous collection of data scraped from seemingly everywhere.

It was revolutionary, but there were problems. Sometimes the tech spewed hate speech and misinformation. The engineers at OpenAI kept postponing the launch and talking about what to do.

One option was to release an older, less powerful version of the technology — and just see what happened. The idea, according to four people familiar with OpenAI's work, was to watch the public's reaction and use it to work out the kinks.

And though some executives have downplayed it, they wanted to beat the competition. Lots of tech companies were working on their own A.I. chatbots. But the people to beat were at Anthropic, started the year before by researchers and engineers who left OpenAI because they thought that Sam Altman, its chief executive, had not made safety a priority as A.I. grew more powerful. The defectors had helped build the technology that OpenAI was so excited about before they trooped out the door.

In mid-November 2022, Mr. Altman; Greg Brockman, OpenAI's president; and others met in a top-floor conference room to discuss the problems with their breakthrough tech yet again. Suddenly Mr. Altman made the decision — they would release the old, less-powerful technology.

The plan was to call it Chat with GPT 3.5 and put it out by the end of the month. They referred to it as a "low key research preview." It didn't feel like a big-deal decision to anyone in the room.

"We plan to frame it as a research release," Mira Murati, OpenAI's chief technology officer, told staff over Slack. "This reduces risk in all dimensions while allowing us to learn a lot," she wrote. "We are aiming to move quickly over the next few days to make it happen."

The underlying code was a bit of a blob. It needed to be converted into something regular people without Ph.D.s could interact with. Mr. Altman and other executives asked a group of engineers to graft a graphical user interface — a GUI, pronounced gooey — onto the blob. A GUI is the face of an application, where you type and press buttons.

A GUI had been created earlier that year to show the technology to Bill Gates, Microsoft's founder, at his home outside Seattle. They stuck the same GUI on and changed the name to ChatGPT. About two weeks after Mr. Altman made his decision, they were good to go.

On Nov. 29, the night before the launch, Mr. Brockman hosted drinks for the team. He didn't think ChatGPT would attract a lot of attention, he said. His prediction: "no more than one tweet thread with 5k likes."

Mr. Brockman was wrong. On the morning of Nov. 30, Mr. Altman tweeted about OpenAI's new product, and the company posted a jargon-heavy blog item. And then, ChatGPT took off. Almost immediately, sign-ups overwhelmed the company's servers. Engineers rushed in and out of a messy space near the office kitchen, huddling over laptops to pull computing power from other projects. In five days, more than a million people had used ChatGPT. Within a few weeks, that number would top 100 million. Though nobody was quite sure why, it was a hit. Network news programs tried to explain how it worked. A late-night comedy show even used it to write (sort of funny) jokes.

After things settled down, OpenAI employees used DALL-E, the company's A.I. image generator, to make a laptop sticker labeled "Low key research preview." It showed a computer about to be consumed by flames.

Zuckerberg Gets Warned

Actually, months earlier Meta had released its own chatbot — to very little notice.

BlenderBot was a flop. The A.I.-powered bot, released in August 2022, was built to carry on conversations — and that it did. It said that Donald J. Trump was still president and that President Biden had lost in 2020. Mark Zuckerberg, it told a user, was "creepy." Then two weeks before ChatGPT was released, Meta introduced Galactica. Designed for scientific research, it could instantly write academic articles and solve math problems. Someone asked it to write a research paper about the history of bears in space. It did. After three days, Galactica was shut down.

Mr. Zuckerberg's head was elsewhere. He had spent the entire year reorienting the company around the metaverse and was focused on virtual and augmented reality.

But ChatGPT would demand his attention. His top A.I. scientist, Yann LeCun, arrived in the Bay Area from New York about six weeks later for a routine management meeting at Meta, according to a person familiar with the meeting. Dr. LeCun led a double life — as Meta's chief

A.I. scientist and a professor at New York University. The Frenchman had won the Turing Award, computer science's most prestigious honor, alongside Dr. Hinton, for work on neural networks.

As they waited in line for lunch at a cafe in Meta's Frank Gehry-designed headquarters, Dr. LeCun delivered a warning to Mr. Zuckerberg. He said Meta should match OpenAI's technology and also push forward with work on an A.I. assistant that could do stuff on the internet on your behalf. Websites like Facebook and Instagram could become extinct, he warned. A.I. was the future.

Mr. Zuckerberg didn't say much, but he was listening. There was plenty of A.I. at work across Meta's apps — Facebook, Instagram, WhatsApp — but it was under the hood. Mr. Zuckerberg was frustrated. He wanted the world to recognize the power of Meta's A.I. Dr. LeCun had always argued that going open-source, making the code public, would attract countless researchers and developers to Meta's technology, and help improve it at a far faster pace. That would allow Meta to catch up — and put Mr. Zuckerberg back in league with his fellow moguls. But it would also allow anyone to manipulate the technology to do bad things.
At dinner that evening, Mr. Zuckerberg approached Dr. LeCun. "I have been thinking about what you said," Mr. Zuckerberg told his chief A.I. scientist, according to a person familiar with the conversation. "And I think you're right."

In Paris, Dr. LeCun's scientists had developed an A.I.-powered bot that they wanted to release as open-source technology. Open source meant that anyone could tinker with its code. They called it Genesis, and it was pretty much ready to go. But when they sought permission to release it, Meta's legal and policy teams pushed back, according to five people familiar with the discussion.

Caution versus speed was furiously debated among the executive team in early 2023 as Mr. Zuckerberg considered Meta's course in the wake of ChatGPT.

Had everyone forgotten about the last seven years of Facebook's history? That was the question asked by the legal and policy teams. They reminded Mr. Zuckerberg about the uproar over hate speech and misinformation on Meta's platforms and the scrutiny the company had endured by the news media and Congress after the 2016 election.

Open sourcing the code might put powerful tech into the hands of those with bad intentions and Meta would take the blame. Jennifer Newstead, Meta's chief legal officer, told Mr. Zuckerberg that an open-source approach to A.I. could attract the attention of regulators who already had the company in their cross hairs, according to two people familiar with her concerns.

At a meeting in late January in his office, called the aquarium because it looked like one, Mr. Zuckerberg told executives that he had made his decision. Parts of Meta would be reorganized and its priorities changed. There would be weekly meetings to update executives on A.I. progress. Hundreds of employees would be moved around. Mr. Zuckerberg declared in a Facebook post that Meta would "turbocharge" its work on A.I.

Mr. Zuckerberg wanted to push out a project fast. The researchers in Paris were ready with Genesis. The name was changed to LLaMA, short for "Large Language Model Meta AI," and released to 4,000 researchers outside the company. Soon Meta received over 100,000 requests for access to the code.

But within days of LLaMA's release, someone put the code on 4chan, the fringe online message board. Meta had lost control of its chatbot, raising the possibility that the worst fears of its legal and policy teams would come true. Researchers at Stanford University showed that the Meta system could easily do things like generate racist material.

On June 6, Mr. Zuckerberg received a letter about LLaMA from Senators Josh Hawley of Missouri and Richard Blumental of Connecticut. "Hawley and Blumental demand answers from Meta," said a news release.

The letter called Meta's approach risky and vulnerable to abuse and compared it unfavorably with ChatGPT. Why, the senators seemed to want to know, couldn't Meta be more like OpenAI?

Under the Tent at Microsoft

For Mr. Nadella, the realization that OpenAI's tech could change everything did not come as an "Aha!" moment. After investing $1 billion in 2019, Microsoft slowly started playing with the start-up's code. First up was GitHub, the company's code storage service. A few teams of engineers started experimenting with OpenAI's tech to help them write code.

Over dinner in Microsoft's boardroom with a friend in the summer of 2021, Mr. Nadella said he was beginning to see the technology as a game changer. It would touch every part of Microsoft's business and every human being, he predicted. (The GitHub experiment eventually became a product: GitHub Copilot.)

A year later, Mr. Nadella got a peek at what would become GPT-4. Mr. Nadella asked it to translate a poem written in Persian by Rumi, who died in 1273, into Urdu. It did. He asked it to transliterate the Urdu into English characters. It did that, too. “Then I said, ‘God, this thing,’” Mr. Nadella recalled in an interview. From that moment, he was all in.

Microsoft’s $1 billion investment in OpenAI had already grown to $3 billion. Now Microsoft was planning to increase that to $10 billion.

Even for Microsoft, which was sitting on $105 billion in cash, that was real money. OpenAI was structured as a nonprofit. Microsoft would not get a board seat. But it had the right to use OpenAI’s code. That meant Microsoft and OpenAI were partners and competitors.

At the end of the summer of 2022, Microsoft’s offices weren’t yet back to their prepandemic bustle. But on Sept. 13, Mr. Nadella summoned his top executives to a meeting at Building 34, Microsoft’s executive nerve center. It was two months before Mr. Altman made the decision to release ChatGPT.

He and Mr. Brockman demonstrated GPT-4 for the group. First they asked it biology questions. Then Mr. Brockman let the executives try to stump the chatbot. At one point the chatbot was asked a question about photosynthesis. Not only did it answer, but it ruled out other possibilities. Peter Lee, the head of Microsoft Research, was surprised it seemed to know how to reason. He turned to Microsoft’s chief scientist, who was sitting next to him, and asked, “What is going on there?!”

Then Mr. Nadella took the lectern to tell his lieutenants that everything was about to change. This was an executive order from a leader who typically favored consensus. “We are pivoting the whole company on this technology,” Eric Horvitz, the chief scientist, later remembered him saying. “This is a central advancement in the history of computing, and we are going to be on that wave at the front of it.”

It all had to stay secret for the time being. Not everyone would be brought into the tent, and at Microsoft, tents were where the important stuff happened. Three “tented projects” were set up in early October to get the big pivot started. They were devoted to cybersecurity, the Bing search engine, Microsoft Word and related software.

About two months later, Yusuf Mehdi, a marketing executive, demonstrated the Bing chatbot for some members of the board. They weren’t sold on it. They found the product overly complicated and without a clear vision to communicate to consumers. Mr. Nadella’s team hadn’t nailed it.

Two weeks later, Mr. Mehdi met with the full board. This time the version he demonstrated was more simple and consumer-friendly. It was a go.

Microsoft invited journalists to its Redmond, Wash., campus on Feb. 7 to introduce a chatbot in Bing to the world. They were instructed not to tell anybody they were going to a Microsoft event, and the topic wasn't disclosed.

But somehow, Google found out. On Feb. 6, to get out ahead of Microsoft, it put up a blog post by Mr. Pichai announcing that Google would be introducing its own chatbot, Bard. It didn't say exactly when.
Image

Mr. Altman had just arrived at Microsoft's conference center for a dry run of the show when Mr. Mehdi grabbed him and showed him Mr. Pichai's post.

"'Oh my gosh, this is hysterical,'" Mr. Mehdi recalled Mr. Altman saying. Just then Mr. Nadella walked out of the room where he had been rehearsing. Mr. Altman suggested that he and Mr. Nadella take a selfie. He posted it on Twitter to tweak Google.

"Hello from redmond! excited for the event tomorrow," tweeted Mr. Altman, who had more than 1.3 million Twitter followers.

By the morning of Feb. 8, the day after Microsoft announced the chatbot, its shares were up 5 percent. But for Google, the rushed announcement became an embarrassment. Researchers spotted errors in Google's blog post. An accompanying GIF simulated Bard saying that the Webb telescope had captured the first pictures of an exoplanet, a planet outside the solar system. In fact, a telescope at the European Southern Observatory in northern Chile got the first image of an exoplanet in 2004. Bard had gotten it wrong, and Google was ribbed in the news media and on social media.

It was, as Mr. Pichai later said in an interview, "unfortunate." Google's stock dropped almost 8 percent, wiping out more than $100 billion in value.

A Google Goodbye

There was no question the Bing chatbot put Microsoft ahead of Google, and in spring 2023 Mr. Nadella bought more than $2 billion in computer chips to keep it that way, according to two people familiar with the budget. "We have a big order coming to you, a really big order coming to you," Mr. Nadella gleefully told Jensen Huang, Nvidia's chief executive, Mr. Huang said.

Mr. Pichai, at Google, felt like a scuba diver. The fallout from Google's announcement about Bard was tumultuous, and that was like navigating the rough top foot of an ocean. But underneath the surface, the water was calm, and he was focused on the coming release of Google's A.I. products.

Mr. Pichai oversaw more than 2,000 researchers divided between two labs, Google Brain and DeepMind. In April, he merged them. Google DeepMind would develop an A.I. system called Gemini. To run it, Mr. Pichai chose Demis Hassabis, a founder of DeepMind. Mr. Hassabis had long and loudly warned that A.I. could destroy humanity. Now he would be in charge of leading Google to artificial intelligence supremacy.

Geoffrey Hinton, Google's best-known scientist, had always poked fun at people like Dr. Hassabis — the doomers, rationalists and effective altruists who worried that A.I would end mankind in the near future. He had developed much of the science behind artificial intelligence as a professor at the University of Toronto and became a wealthy man after joining Google in 2013. He is often called the godfather of A.I.

But the new chatbots changed everything for him. The science had moved more quickly than he had expected. Microsoft's introduction of its chatbot convinced him that Google would have no choice but to try to catch up. And the corporate race shaping up between tech giants seemed dangerous.

"If you think of Google as a company whose aim is to make profits," Dr. Hinton said in April, "they can't just let Bing take over from Google search. They've got to compete with that. When Microsoft decided to release a chatbot as the interface for Bing, that was the end of the holiday period."

Dr. Hinton spent a lot of time mulling his own role in the development of A.I. Sometimes he felt regretful. Other times he jokingly sent friends a video of Edith Piaf singing "Non, Je Ne Regrette Rien." But finally, he decided to quit.

For the first time in more than 50 years, he stepped away from research. And then in April, he called Mr. Pichai and said goodbye.

MIT Technology Review

Featured Topics Newsletters Events Podcasts SIGN IN SUBSC

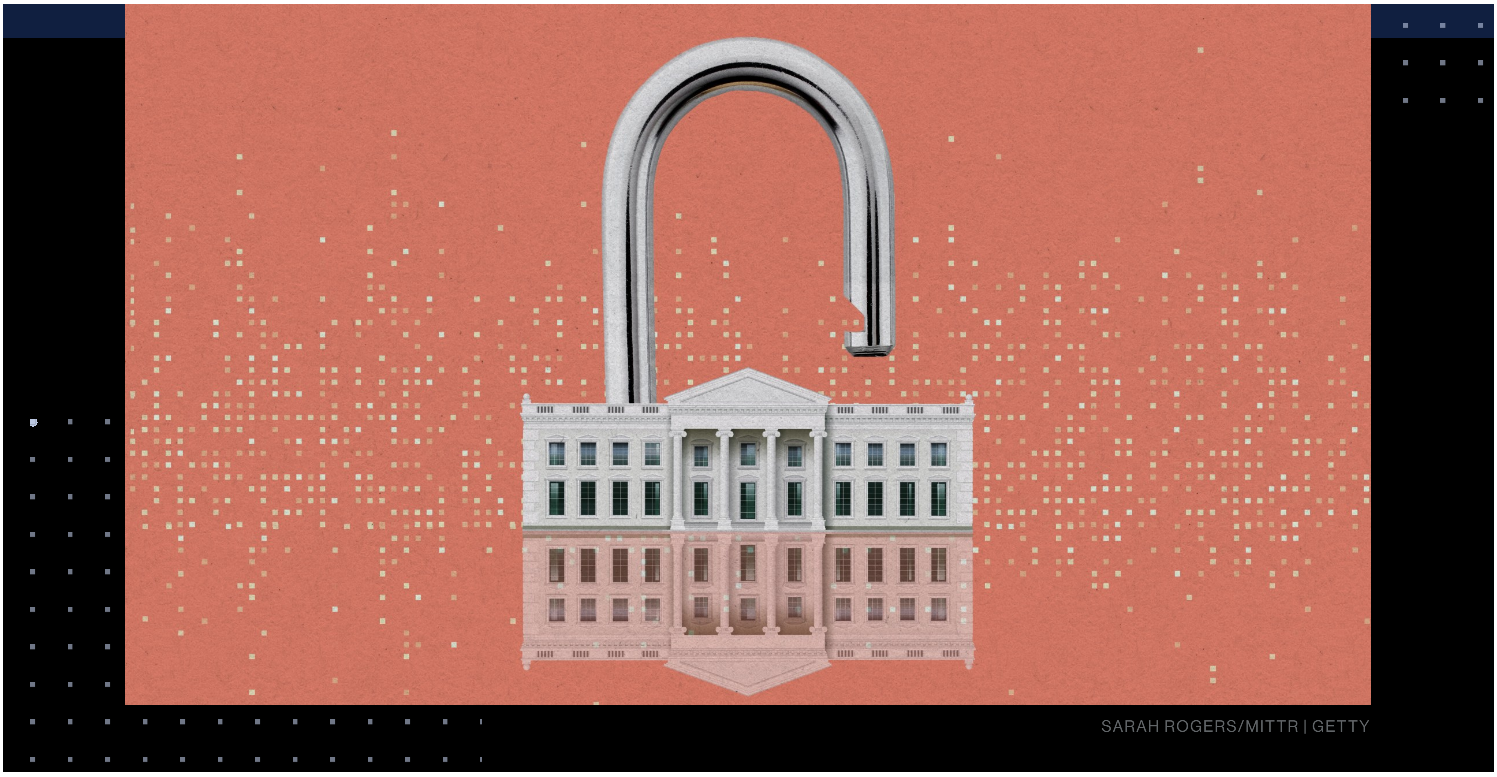
ARTIFICIAL INTELLIGENCE

AI companies promised to self-regulate o year ago. What's changed?

The White House's voluntary AI commitments have brought better red-teaming practices and watermarks, but no meaningful transparency or accountability.

By Melissa Heikkilä

July 22, 2024



SARAH ROGERS/MITTR | GETTY

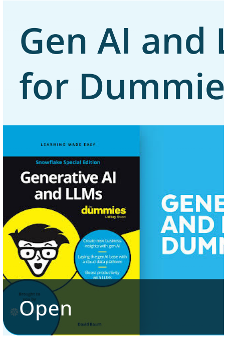
One year ago, on July 21, 2023, seven leading AI companies—Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI—committed with the White House to a set of eight voluntary commitments on how to develop AI in a safe and trustworthy way.

These included promises to do things like improve the testing and transparency around AI systems, and share information on potential harms and risks.

On the first anniversary of the voluntary commitments, *MIT Technology Review* asked the AI companies that signed the commitments for details on their work so far. Their replies show that the tech sector has made some welcome progress, with big caveats.

The voluntary commitments came at a time when generative AI mania was perhaps at its frothiest, with companies racing to launch their own models and make them bigger and better than their competitors'. At the same time, we started to see developments such as fights over copyright and deepfakes. A vocal lobby of influential tech players, such as Geoffrey Hinton, had also raised concerns that AI could pose an existential risk to humanity. Suddenly, everyone was talking about the urgent need to make AI *safe*, and regulators everywhere were under pressure to do something about it.

Until very recently, AI development has been a Wild West. Traditionally, the US has been loath to regulate its tech giants, instead relying on them to regulate themselves. The voluntary commitments are a good example of that:



POPULAR

How to fix a Windows PC
global outage

Rhiannon Williams

A controversial Chinese
still hopeful about embry
Here's why.

Zeyi Yang

Google DeepMind's new
now solve complex math

Rhiannon Williams

OpenAI has released a n
that you can talk to

Melissa Heikkilä

they were some of the first prescriptive rules for the AI sector in the US, but they remain voluntary and unenforceable. The White House has since issued an executive order, which expands on the commitments and also applies to other tech companies and government departments.

"One year on, we see some good practices towards their own products, but [they're] nowhere near where we need them to be in terms of good governance or protection of rights at large," says Merve Hickok, the president and research director of the Center for AI and Digital Policy, who reviewed the companies' replies as requested by *MIT Technology Review*. Many of these companies continue to push unsubstantiated claims about their products, such as saying that they can supersede human intelligence and capabilities, adds Hickok.

One trend that emerged from the tech companies' answers is that companies are doing more to pursue technical fixes such as red-teaming (in which humans probe AI models for flaws) and watermarks for AI-generated content.

But it's not clear what the commitments have changed and whether the companies would have implemented these measures anyway, says Rishi Bommasani, the society lead at the Stanford Center for Research on Foundation Models, who also reviewed the responses for *MIT Technology Review*.

One year is a long time in AI. Since the voluntary commitments were signed, Inflection AI founder Mustafa Suleyman has left the company and joined Microsoft to lead the company's AI efforts. Inflection declined to comment.

"We're grateful for the progress leading companies have made toward fulfilling their voluntary commitments in addition to what is required by the executive order," says Robyn Patterson, a spokesperson for the White House. But, Patterson adds, the president continues to call on Congress to pass bipartisan legislation on AI.

Without comprehensive federal legislation, the best the US can do right now is to demand that companies follow through on these voluntary commitments, says Brandie Nonnecke, the director of the CITRIS Policy Lab at UC Berkeley.

But it's worth bearing in mind that "these are still companies that are essentially writing the exam by which they are evaluated," says Nonnecke. "So we have to think carefully about whether or not they're … verifying themselves in a way that is truly rigorous."

Here's our assessment of the progress AI companies have made in the past year.

Commitment 1

The companies commit to internal and external security testing of their AI systems before their release. This testing, which will be carried out in part by independent experts, guards against some of the most significant sources of AI risks, such as biosecurity and cybersecurity, as well as its broader societal *effects.*

All the companies (excluding Inflection, which chose not to comment) say they conduct red-teaming exercises that get both internal and external testers to probe their models for flaws and risks. OpenAI says it has a separate preparedness team that tests models for cybersecurity, chemical, biological, radiological, and nuclear threats and for situations where a sophisticated AI model can do or persuade a person to do things that might lead to harm. Anthropic and OpenAI also say they conduct these tests with external experts before launching their new models. For example, for the launch of Anthropic's latest model, Claude 3.5, the company conducted predeployment testing with experts at the UK's AI Safety Institute. Anthropic has also allowed METR, a research nonprofit, to do an "initial exploration" of Claude 3.5's capabilities for autonomy. Google says it also conducts internal red-teaming to test the boundaries of its model, Gemini, around election-related content, societal risks, and national security concerns. Microsoft says it has worked with third-party evaluators at NewsGuard, an organization advancing journalistic integrity, to evaluate risks and mitigate the risk of abusive deepfakes in Microsoft's text-to-image tool. In addition to red-teaming, Meta says, it evaluated its latest model, Llama 3, to understand its performance in a series of risk areas like weapons, cyberattacks, and child exploitation.

Related Story

Three things to know about the White House's executive order on AI

Experts say its emphasis on content labeling, watermarking, and transparency represents important steps forward.

But when it comes to testing, it's not enough to just report that a company is taking actions, says Bommasani. For example, Meta, Amazon and Anthropic said they had worked with the nonprofit Thorn to combat risks to child safety posed by AI. Bommasani would have wanted to see more specifics about how the interventions that companies are implementing actually reduce those risks.

"It should become clear to us that it's not just that companies are doing things but those things are having the desired effect," Bommasani says.

RESULT: Good. The push for red-teaming and testing for a wide range of risks is a good and important one. However, Hickok would have liked to see independent researchers get broader access to companies' models.

Commitment 2

The companies commit to sharing information across the industry and with governments, civil society, and academia on managing AI risks. This includes best practices for safety, information on attempts to circumvent safeguards, and technical collaboration.

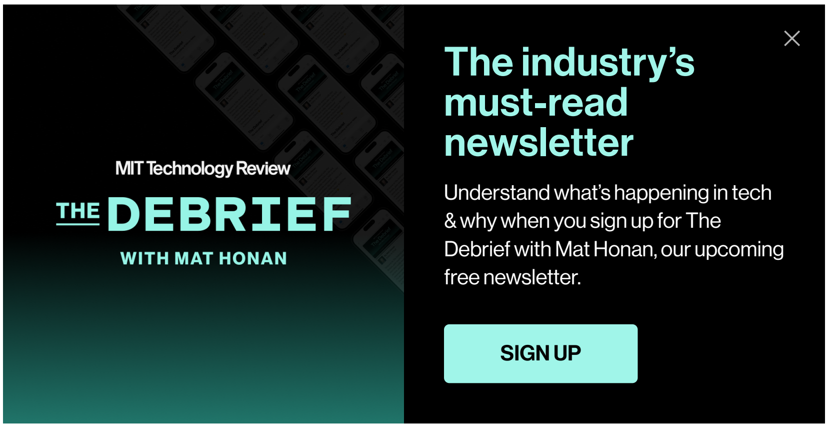
After they signed the commitments, Anthropic, Google, Microsoft, and OpenAI founded the Frontier Model Forum, a nonprofit that aims to facilitate discussions and actions on AI safety and responsibility. Amazon and Meta have also joined.

Engaging with nonprofits that the AI companies funded themselves may not be in the spirit of the voluntary commitments, says Bommasani. But the Frontier Model Forum could be a way for these companies to cooperate with each other and pass on information about safety, which they normally could not do as competitors, he adds.

"Even if they're not going to be transparent to the public, one thing you might want is for them to at least collectively figure out mitigations to actually reduce risk," says Bommasani.

All of the seven signatories are also part of the Artificial Intelligence Safety Institute Consortium (AISIC), established by the National Institute of Standards and Technology (NIST), which develops guidelines and standards for AI policy and evaluation of AI performance. It is a large consortium consisting of a mix of public- and private-sector players. Google, Microsoft, and OpenAI also have representatives at the UN's High-Level Advisory Body on Artificial Intelligence.

Many of the labs also highlighted their research collaborations with academics. For example, Google is part of MLCommons, where it worked with academics on a cross-industry AI Safety Benchmark. Google also says it actively contributes tools and resources, such as computing credit, to projects like the National Science Foundation's National AI Research Resource pilot, which aims to democratize AI research in the US. Meta says it is also part of the AI Alliance, a network of companies, researchers and nonprofits, and specifically engages in open source AI and the developer community.



Many of the companies also contributed to guidance by the Partnership on AI, another nonprofit founded by Amazon, Facebook, Google, DeepMind, Microsoft, and IBM, on the deployment of foundation models.

RESULT: More work is needed. More information sharing is a welcome step as the industry tries to collectively make AI systems safe and trustworthy. However, it's unclear how much of the effort advertised will actually lead to meaningful changes and how much is window dressing.

Commitment 3

The companies commit to investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights. These model weights are the most essential part of an AI system, and the companies agree that it is vital that the model weights be released only when intended and when security risks are considered.

Many of the companies have implemented new cybersecurity measures in the past year. For example, Microsoft has launched the Secure Future Initiative to address the growing scale of cyberattacks. The company says its model weights are encrypted to mitigate the potential risk of model theft, and it applies strong identity and access controls when deploying highly capable proprietary models.

Google too has launched an AI Cyber Defense Initiative**.** In May OpenAI shared six new measures it is developing to complement its existing cybersecurity practices, such as extending cryptographic protection to AI hardware. It also has a Cybersecurity Grant Program, which gives researchers access to its models to build cyber defenses.

Related Story

Three ways AI chatbots are a security disaster

Large language models are full of security vulnerabilities, yet they're being embedded into tech products on a vast scale.

Amazon mentioned that it has also taken specific measures against attacks specific to generative AI, such as data poisoning and prompt injection, in which someone uses prompts that direct the language model to ignore its previous directions and safety guardrails.

Just a couple of days after signing the commitments, Anthropic published details about its protections, which include common cybersecurity practices such as controlling who has access to the models and sensitive assets such as model weights, and inspecting and controlling the third-party supply chain. The company also works with independent assessors to evaluate whether the controls it has designed meet its cybersecurity needs.

RESULT: Good. All of the companies did say they had taken extra measures to protect their models, although it doesn't seem there is much consensus on the best way to protect AI models.

Commitment 4

The companies commit to facilitating third-party discovery and reporting of vulnerabilities in their AI systems. Some issues may persist even after an AI system is released and a robust reporting mechanism enables them to be found and fixed quickly.

For this commitment, one of the most popular responses was to implement bug bounty programs, which reward people who find flaws in AI systems.

Anthropic, Google, Microsoft, Meta, and OpenAI all have one for AI systems. Anthropic and Amazon also said they have forms on their websites where security researchers can submit vulnerability reports.

It will likely take us years to figure out how to do third-party auditing well, says Brandie Nonnecke. "It's not just a technical challenge. It's a socio-technical challenge. And it just kind of takes years for us to figure out not only the technical standards of AI, but also socio-technical standards, and it's messy and hard," she says.

Nonnecke says she worries that the first companies to implement third-party audits might set poor precedents for how to think about and address the socio-technical risks of AI. For example, audits might define, evaluate, and address some risks but overlook others.

RESULT: More work is needed. Bug bounties are great, but they're nowhere near comprehensive enough. New laws, such as the EU's AI Act, will require tech companies to conduct audits, and it would have been great to see tech companies share successful examples of such audits.

Commitment 5

The companies commit to developing robust technical mechanisms to ensure that users know when content is AI generated, such as a watermarking system. This action enables creativity with AI to flourish but reduces the dangers of fraud and deception.

Many of the companies have built watermarks for AI-generated content. For example, Google launched SynthID, a watermarking tool for image, audio, text, and video generated by Gemini. Meta has a tool called Stable Signature for images, and AudioSeal for AI-generated speech. Amazon now adds an invisible watermark to all images generated by its Titan Image Generator. OpenAI also uses watermarks in Voice Engine, its custom voice model, and has built an image-detection classifier for images generated by DALL-E 3. Anthropic was the only company that hadn't built a watermarking tool, because watermarks are mainly used in images, which the company's Claude model doesn't support.

Related Story

All the companies excluding Inflection, Anthropic, and Meta are also part of the Coalition for Content Provenance and Authenticity (C2PA), an industry coalition that embeds information about when content was created, and whether it was created or edited by AI, into an image's metadata. Microsoft and OpenAI

Why Big Tech's watermarking plans are some welcome good news

Big Tech is throwing its weight behind a standard that could add a "nutrition label" to images, video, and audio.

automatically attach the C2PA's provenance metadata to images generated with DALL-E 3 and videos generated with Sora. While Meta is not a member, it announced it is using the C2PA standard to identify AI-generated images on its platforms.

The six companies that signed the commitments have a "natural preference to more technical approaches to addressing risk," says Bommasani, "and certainly watermarking in particular has this flavor."

"The natural question is: Does [the technical fix] meaningfully make progress and address the underlying social concerns that motivate why we want to know whether content is machine generated or not?" he adds.

RESULT: Good. This is an encouraging result overall. While watermarking remains experimental and is still unreliable, it's still good to see research around it and a commitment to the C2PA standard. It's better than nothing, especially during a busy election year.

Commitment 6

The companies commit to publicly reporting their AI systems' capabilities, limitations, and areas of appropriate and inappropriate use. This report will cover both security risks and societal risks, such as the effects on fairness and bias.

The White House's commitments leave a lot of room for interpretation. For example, companies can technically meet this public reporting commitment with widely varying levels of transparency, as long as they do *something* in that general direction.

The most common solutions tech companies offered here were so-called model cards. Each company calls them by a slightly different name, but in essence they act as a kind of product description for AI models. They can address anything from the model's capabilities and limitations (including how it measures up against benchmarks on fairness and explainability) to veracity, robustness, governance, privacy, and security. Anthropic said it also tests models for potential safety issues that may arise later.

Microsoft has published an annual Responsible AI Transparency Report, which provides insight into how the company builds applications that use generative AI, make decisions, and oversees the deployment of those applications. The company also says it gives clear notice on where and how AI is used within its products.

Meta also has released its new Llama 3 model with a detailed and extensive technical report. The company also updated its Responsible Use Guide which includes guidance on how to use and responsibly deploy advanced large language models.

RESULT: More work is needed. One area of improvement for AI companies would be to increase transparency on their governance structures and on the financial relationships between companies, Hickok says. She would also have liked to see companies be more public about data provenance, model training processes, safety incidents, and energy use.

Commitment 7

The companies commit to prioritizing research on the societal risks that AI systems can pose, including on avoiding harmful bias and discrimination, and protecting privacy. The track record of AI shows the insidiousness and prevalence of these dangers, and the companies commit to rolling out AI that mitigates them.

Tech companies have been busy on the safety research front, and they have embedded their findings into products. Amazon has built guardrails for Amazon Bedrock that can detect hallucinations and can apply safety, privacy, and truthfulness protections. Anthropic says it employs a team of researchers dedicated to researching societal risks and privacy. In the past year, the company has pushed out research on deception, jailbreaking, strategies to mitigate discrimination, and emergent capabilities such as models' ability to tamper with their own code or engage in persuasion. And OpenAI says it has trained its models to avoid producing hateful content and refuse to generate output on hateful or extremist content. It trained its GPT-4V to refuse many requests that require drawing from stereotypes to answer. Google DeepMind has also released research to evaluate dangerous capabilities, and the company has done a study on misuses of generative AI.

All of them have poured a lot of money into this area of research. For example, Google has invested millions into creating a new AI Safety Fund to promote research in the field through the Frontier Model Forum. Microsoft says it has committed $20 million in compute credits to researching societal risks through the National AI Research Resource and started its own AI model research accelerator program for academics, called the Accelerating Foundation Models Research program. The company has also hired 24 research fellows focusing on AI and society.

Related Story

How existential risk became the biggest meme in AI

"Ghost stories are contagious."

RESULT: Very good. This is an easy commitment to meet, as the signatories are some of the biggest and richest corporate AI research labs in the world. While more research into how to make AI systems safe is a welcome step, critics say that the focus on safety research takes attention and resources from AI research that focuses on more immediate harms, such as discrimination and bias.

Commitment 8

The companies commit to develop and deploy advanced AI systems to help address society's greatest challenges. From cancer prevention to mitigating climate change to so much in between, AI—if properly managed—can contribute enormously to the prosperity, equality, and security of all.

Since making this commitment, tech companies have tackled a diverse set of problems. For example, Pfizer used Claude to assess trends in cancer treatment research after gathering relevant data and scientific content, and Gilead, an American biopharmaceutical company, used generative AI from Amazon Web Services to do feasibility evaluations on clinical studies and analyze data sets.

Google DeepMind has a particularly strong track record in pushing out AI tools that can help scientists. For example, AlphaFold 3 can predict the structure and interactions of all life's molecules. AlphaGeometry can solve geometry problems at a level comparable with the world's brightest high school mathematicians. And GraphCast is an AI model that is able to make medium-range weather forecasts. Meanwhile, Microsoft has used satellite imagery and AI to improve responses to wildfires in Maui and map climate-vulnerable populations, which helps researchers expose risks such as food insecurity, forced migration, and disease.

Related Story

Google DeepMind's weather AI can forecast extreme weather faster and more accurately

It said Hurricane Lee would make landfall in Nova Scotia three days sooner than traditional methods predicted.

OpenAI, meanwhile, has announced partnerships and funding for various research projects, such as one looking at how multimodal AI models can be used safely by educators and by scientists in laboratory settings It has also offered credits to help researchers use its platforms during hackathons on clean energy development.

RESULT: Very good. Some of the work on using AI to boost scientific discovery or predict weather events is genuinely exciting. AI companies haven't used AI to prevent cancer yet, but that's a pretty high bar.

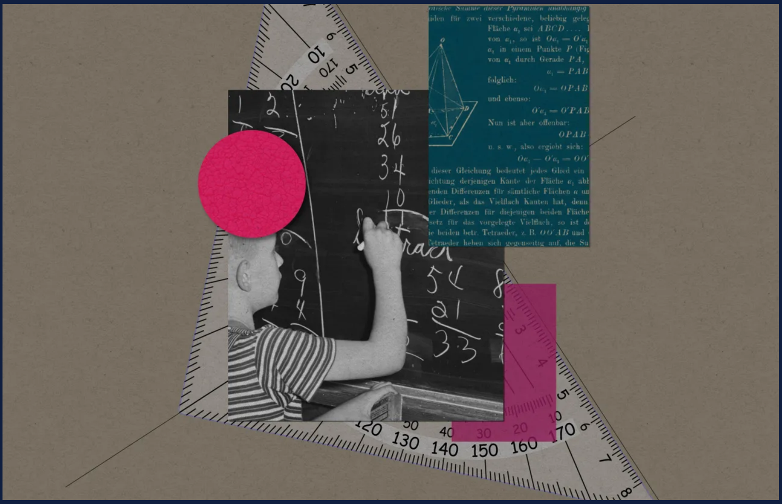
Overall, there have been some positive changes in the way AI has been built, such as red-teaming practices, watermarks and new ways for industry to share best practices. However, these are only a couple of neat technical solutions to the messy socio-technical problem that is AI harm, and a lot more work is needed. One year on, it is also odd to see the commitments talk about a very particular type of AI safety that focuses on hypothetical risks, such bioweapons, and completely fail to mention consumer protection, nonconsensual deepfakes, data and copyright, and the environmental footprint of AI models. These seem like weird omissions today.

UPDATE: This story has been updated to include additional information from Meta.

by Melissa Heikkilä

DEEP DIVE

ARTIFICIAL INTELLIGENCE



Google DeepMind's new AI systems can now solve complex math problems

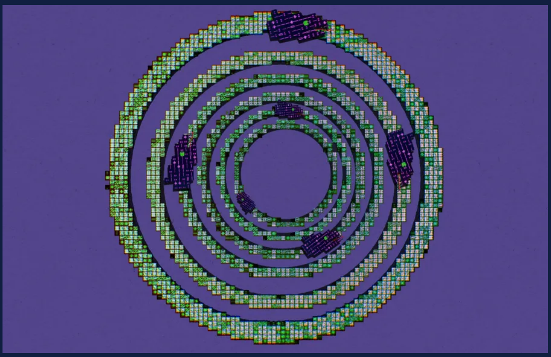
AlphaProof and AlphaGeometry 2 are steps toward building systems that can reason, which could unlock exciting new capabilities.

By Rhiannon Williams

OpenAI has released a new ChatGPT bot that you can talk to

The voice-enabled chatbot will be available to a small group of people today, and to all ChatGPT Plus users in the fall.

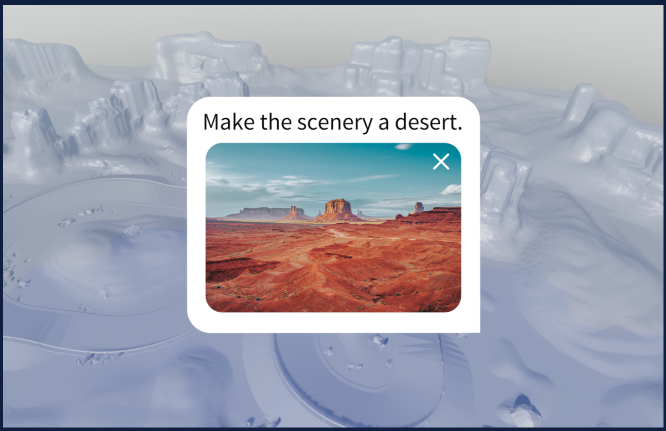
By Melissa Heikkilä



AI trained on AI garbage spits out AI garbage

As junk web pages written by AI proliferate, the models that rely on that data will suffer.

By Scott J Mulligan



Roblox is launching a generative AI that builds 3D environments in a snap

It will make it easy to build new game environments on the platform, even if you don't have any design skills.

By Scott J Mulligan

STAY CONNECTED

Illustration by Rose Wong

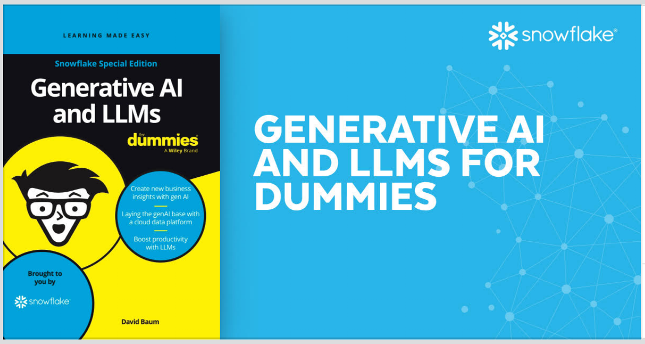
Get the latest updates from MIT Technology Review

Discover special offers, top stories, upcoming events, and more.

Enter your email

→

Privacy Policy


LEARNING MADE EASY
Snowflake Special Edition
Generative AI and LLMs
dümmies
A Wiley Brand
Create new business insights with gen AI
Laying the genAI base with a cloud data platform
Boost productivity with LLMs
Brought to you by
snowflake
David Baum
snowflake
GENERATIVE AI AND LLMS FOR DUMMIES

The latest iteration of a legacy

Founded at the Massachusetts Institute of Technology in 1899, MIT Technology Review is a world-renowned, independent media company whose insight, analysis, reviews, interviews and live events explain the newest technologies and their commercial, social and political impact.

Advertise with MIT Technology Review

Elevate your brand to the forefront of conversation around emerging technologies that are radically transforming business. From event sponsorships to custom content to visually arresting video storytelling, advertising with MIT Technology Review creates opportunities for your brand to resonate with an unmatched audience of technology and business elite.

READ ABOUT OUR HISTORY

ADVERTISE WITH US

About us

Careers

Custom content

Advertise with us

International Editions

Republishing

MIT Alumni News

Help & FAQ

My subscription

Editorial guidelines

Privacy policy

Terms of Service

Write for us

Contact us

IEEE Spectrum

GUEST ARTICLE ARTIFICIAL INTELLIGENCE

Was an AI Image Generator Taken Down for Making Child Porn? › An inquiry into models that can create child sexual abuse material may have yielded results

BY DAVID EVAN HARRIS DAVE WILLNER 30 AUG 2024

David Evan Harris is a chancellor's public scholar at UC Berkeley.

Dave Willner is a fellow at Stanford University's program on governance of emerging technologies.

MIKE KEMP/GETTY IMAGES

Why are AI companies valued in the millions and billions of dollars creating and distributing tools that

billions of dollars creating and distributing tools that can make AI-generated child sexual abuse material (CSAM)?

An image generator called Stable Diffusion version 1.5, which was created by the AI company Runway with funding from Stability AI, has been particularly implicated in the production of CSAM. And popular platforms such as Hugging Face and Civitai have been hosting that model and others that may have been trained on real images of child sexual abuse. In some cases, companies may even be breaking laws by hosting synthetic CSAM material on their servers. And why are mainstream companies and investors like Amazon, Google, Nvidia, Intel, Salesforce, and Andreessen Horowitz pumping hundreds of millions of dollars into these companies? Their support amounts to subsidizing content for pedophiles.

As AI safety experts, we've been asking these questions to call out these companies and pressure them to take the corrective actions we outline below. And we're happy today to report one major triumph: seemingly in response to our questions, Stable Diffusion version 1.5 has been removed from Hugging Face. But there's much still to do, and meaningful progress may require legislation.

The Scope of the CSAM Problem

Child safety advocates began ringing the alarm bell last year: Researchers at Stanford's Internet Observatory and the technology non-profit Thorn published a troubling report in June 2023. They found that broadly available and "open-source" AI image-generation tools were already being misused by malicious actors to make child sexual abuse material. In some cases, bad actors were making their own custom versions of these models (a process known as fine-tuning) with real child sexual abuse material to generate bespoke images of specific victims.

Last October, a report from the U.K. nonprofit Internet Watch Foundation (which collects reports of child sexual abuse material) detailed the ease with which malicious actors are now making photorealistic AI-generated child sexual abuse material, at scale. The researchers included a "snapshot" study of one dark web CSAM forum, analyzing more than 11,000 AI-generated images posted in a one-month period; of those, nearly 3,000 were judged severe enough to be classified as criminal. The report urged stronger regulatory oversight of generative AI models.

AI models can be used to create this material because they've seen examples before. Researchers at Stanford discovered last December that one of the most significant data sets used to train image-generation models included hundreds of pieces of CSAM. Many of the most popular downloadable open-source AI image

Many of the most popular downloadable open-source AI image generators, including the popular Stable Diffusion version 1.5 model, were trained using this data. While Runway created that version of Stable Diffusion, Stability AI paid for the computing power to produce the dataset and train the model, and Stability AI released the subsequent versions.

Runway did not respond to a request for comment. A Stability AI spokesperson emphasized that the company did not release or maintain Stable Diffusion version 1.5, and says the company has “implemented robust safeguards” against CSAM in subsequent models, including the use of filtered data sets for training.

Also last December, researchers at the social media analytics firm Graphika found a proliferation of dozens of “undressing” services, many based on open-source AI image generators, likely including Stable Diffusion. These services allow users to upload clothed pictures of people and produce what experts term nonconsensual intimate imagery (NCII) of both minors and adults, also sometimes referred to as deepfake pornography. Such websites can be easily found through Google searches, and users can pay for the services using credit cards online. Many of these services only work on women and girls, and these types of tools have been used to target female celebrities like Taylor Swift and politicians like U.S. representative Alexandria Ocasio-Cortez.

AI-generated CSAM has real effects. The child safety ecosystem is already overtaxed, with millions of files of suspected CSAM reported to hotlines annually. Anything that adds to that torrent of content—especially photorealistic abuse material—makes it more difficult to find children that are actively in harm's way. Making matters worse, some bad actors are using existing CSAM to generate synthetic images of these survivors—a horrific re-violation of their rights. Others are using the readily available "nudifying" apps to create sexual content from benign imagery of real children, and then using that newly generated content in sexual extortion schemes.

One Victory Against AI-Generated CSAM

Based on the Stanford investigation from last December, it's well-known in the AI community that Stable Diffusion 1.5 was trained on child sexual abuse material, as was every other model trained on the LAION-5B data set. These models are being actively misused by malicious actors to make AI-generated CSAM. And even when they're used to generate more benign material, their use inherently revictimizes the children whose abuse images went into their training data. So we asked the popular AI hosting platforms Hugging Face and Civitai why they hosted Stable Diffusion 1.5 and derivative models, making them available for free download?

It’s worth noting that Jeff Allen, a data scientist at the Integrity Institute, found that Stable Diffusion 1.5 was downloaded from Hugging Face over 6 million times in the past month, making it the most popular AI image-generator on the platform.

When we asked Hugging Face why it has continued to host the model, company spokesperson Brigitte Tousignant did not directly answer the question, but instead stated that the company doesn’t tolerate CSAM on its platform, that it incorporates a variety of safety tools, and that it encourages the community to use the Safe Stable Diffusion model that identifies and suppresses inappropriate images.

Then, yesterday, we checked Hugging Face and found that Stable Diffusion 1.5 is no longer available. Tousignant told us that Hugging Face didn't take it down, and suggested that we contact Runway—which we did, again, but we have not yet received a response.

It's undoubtedly a success that this model is no longer available for download from Hugging Face. Unfortunately, it's still available on Civitai, as are hundreds of derivative models. When we contacted Civitai, a spokesperson told us that they have no knowledge of what training data Stable Diffusion 1.5 used, and that they would only take it down if there was evidence of misuse.

Platforms should be getting nervous about their liability. This past week saw the arrest of Pavel Durov, CEO of the messaging app Telegram, as part of an investigation related to CSAM and other crimes.

What's Being Done About AI-Generated CSAM

The steady drumbeat of disturbing reports and news about AI-generated CSAM and NCII hasn't let up. While some companies are trying to improve their products' safety with the help of the

Tech Coalition, what progress have we seen on the broader issue?

In April, Thorn and All Tech Is Human announced an initiative to bring together mainstream tech companies, generative AI developers, model hosting platforms, and more to define and commit to Safety by Design principles, which put preventing child sexual abuse at the center of the product development process. Ten companies (including Amazon, Civitai, Google, Meta, Microsoft, OpenAI, and Stability AI) committed to these principles, and some also co-authored a related paper with more detailed recommended mitigations. The principles call on companies to develop, deploy, and maintain AI models that proactively address child safety risks; to build systems to ensure that any abuse material that does get produced is reliably detected; and to limit the distribution of the underlying models and services that are used to make this abuse material.

These kinds of voluntary commitments are a start. Rebecca Portnoff, Thorn's head of data science, says the initiative seeks accountability by requiring companies to issue reports about their progress on the mitigation steps. It's also collaborating with standard-setting institutions such as IEEE and NIST to integrate their efforts into new and existing standards, opening the door to third party audits that would "move past the honor system," Portnoff says. Portnoff also notes that Thorn is engaging with policy makers to help them conceive legislation that would be both technically feasible and impactful. Indeed, many experts say it's time to move beyond voluntary commitments.

We believe that there is a reckless race to the bottom currently underway in the AI industry. Companies are so furiously fighting to be *technically* in the lead that many of them are ignoring the *ethical* and possibly even *legal* consequences of their products. While some governments—including the European Union—are making headway on regulating AI, they haven't gone far enough. If, for example, laws made it illegal to provide AI systems that can produce CSAM, tech companies might take notice.

The reality is that while some companies will abide by voluntary commitments, many will not. And of those that do, many will take action too slowly, either because they're not ready or because they're struggling to keep their competitive advantage. In the meantime, bad actors will gravitate to those services and wreak havoc. That outcome is unacceptable.

What Tech Companies Should Do About AI-Generated CSAM

Experts saw this problem coming from a mile away, and child safety advocates have recommended common-sense strategies to combat it. If we miss this opportunity to do something to fix the situation, we'll all bear the responsibility. At a minimum, all companies, including those releasing open source models, should be legally required to follow the commitments laid out in Thorn's Safety by Design principles:

- Detect, remove, and report CSAM from their training data sets before training their generative AI models.
- Incorporate robust watermarks and content provenance systems into their generative AI models so generated images can be linked to the models that created them, as would be required under a California bill that would create Digital Content Provenance Standards for companies that do business in the state. The bill will likely be up for signature by Governor Gavin Newson in the coming month.
- Remove from their platforms any generative AI models that are known to be trained on CSAM or that are capable of producing CSAM. Refuse to rehost these models unless they've been fully reconstituted with the CSAM removed.
- Identify models that have been intentionally fine-tuned on CSAM and permanently remove them from their platforms.
- Remove "nudifying" apps from app stores, block search results for these tools and services, and work with payment providers to block payments to their makers.

There is no reason why generative AI needs to aid and abet the horrific abuse of children. But we will need all tools at hand—

Center for AI and Digital Policy

ChatGPT and the Federal Trade Commission: Still No Guardrails

Center for AI and Digital Policy
Washington, DC

July 2024

TABLE OF CONTENTS

I. Executive Summary

More than a year ago, the Center for AI and Digital Policy (CAIDP) filed a detailed, formal complaint with the Federal Trade Commission (FTC) about OpenAI, alleging that OpenAI had violated U.S. consumer protection law by releasing a consumer product without sufficient safeguards. CAIDP urged the FTC to act to protect consumers and ensure independent oversight of OpenAI and other AI companies.[1] In July 2023, both the New York Times and the Wall Street Journal reported that the FTC had launched the investigation sought by CAIDP. However, a year later, there is still no legal outcome, no judgment, and no settlement. There are Still No Guardrails for AI products sold to consumers in the United States.

The CAIDP OpenAI case is likely the most consequential AI investigation currently pending before the FTC. It could establish safeguards for AI services and bring transparency and accountability to the AI industry. Regulators in several other jurisdictions recognize these concerns and have acted. The urgency of the OpenAI case is underscored also by the absence of new federal laws in the United States to address new challenges resulting from the deployment of AI services. Unlike many other countries in the world,[2] the United States has still not enacted legislation to address public concerns even though polling data shows widespread concern in the U.S.

The purpose of this report *Still No Guardrails* is to review developments since the filing of CAIDP's original OpenAI complaint. Relevant is the range of enforcement actions initiated in other jurisdictions for the same concerns highlighted in our complaint. Alarming is the repeated warnings from AI experts. Obvious is the growing concern about the lack of governance and oversight of AI companies, particularly OpenAI.

In this document we set out an overview of our efforts to get the FTC to establish guardrails for AI. We highlight the accelerated deployment of OpenAI's large language model (LLM) GPT-4, the growing consumer concerns over AI, the views of AI experts, the FTC's mandate and prior statements on AI, and the need for the Federal Trade Commission to act.

[1] Merve Hickok, Christabel Randolph, Marc Rotenberg, *It's time for the FTC to act on ChatGPT*, Opinion, (Jun. 14, 2024), https://thehill.com/opinion/technology/4722343-its-time-for-the-ftc-to-act-on-chatgpt/

[2] See, e.g., European Parliament, *The First Regulation on Artificial Intelligence,* (Jun. 18, 2024), https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

II. Overview of the CAIDP Complaint to FTC re OpenAI and ChatGPT

In March 2023, CAIDP filed a detailed consumer complaint with the Federal Trade Commission (FTC) regarding ChatGPT and OpenAI. In the Complaint, the Supplements that followed, and appearances before the Commissioners at open Commission meetings, CAIDP set out the many problems with OpenAI's business practices and pointed to the FTC's previously issued guidance on AI products. CAIDP asked the Commission to halt the commercial release of new GPT models until necessary safeguards were established. CAIDP also said that such safeguards should be based on the FTC's previously issued guidance on AI products.

Summary of the CAIDP Complaint[3]

In the original Complaint, CAIDP explained that OpenAI released a product GPT-4 for the consumer market that is biased, deceptive, and a risk to privacy and public safety. OpenAI released AI-based products, DALL-E, GPT-4, OpenAI Five, ChatGPT, and OpenAI Codex for commercial use. OpenAI described these AI models as "products." OpenAI provided "pricing information" corresponding to the subscription levels. There was also downstream integration offered by OpenAI – plugins for GPT-4 was made available for routine consumer services, including travel, finance, and shopping.

The CAIDP Complaint quoted extensively from the GPT-4 Technical Report in which OpenAI acknowledged the specific dangers of "Disinformation and influence operations," "Proliferation of conventional and unconventional weapons," and "Cybersecurity." The Complaint highlighted that the outputs of ChatGPT cannot be proven or replicated. No independent assessment was undertaken prior to deployment.

CAIDP also relied on scientific evidence to highlight specific risks of bias, deception, harms to children, privacy, cybersecurity, and consumer protection. The Complaint set out established frameworks for AI governance – the OECD AI Principles, the Universal Guidelines for AI, which recommend the guardrails sought as relief in our complaint.

CAIDP emphasized that there should be independent oversight and evaluation of commercial AI products offered in the United States prior to release in the market. The specific relief sought in our complaint was:

- Halt further commercial deployment of GPT by OpenAI;

[3] CAIDP, *Complaint to the FTC - In re OpenAI and ChatGPT*, (Mar. 30, 2024), https://www.caidp.org/app/download/8450269463/CAIDP-FTC-Complaint-OpenAI-GPT-033023.pdf

- Require the establishment of independent assessment of GPT products prior to future deployment;
- Require compliance with FTC AI Guidance prior to further deployment of GPT
- Require independent assessment throughout the GPT AI lifecycle
- Establish a publicly accessible incident reporting mechanism for GPT-4 similar to the FTC's mechanisms to report consumer fraud;
- Initiate a rulemaking to establish baseline standards for products in the Generative AI market sector; and
- Provide such other relief as the Commission finds necessary and appropriate.

Summary of the First Supplement[4]

In the first Supplement, CAIDP highlighted that subsequent to the filing of the Complaint, consumer agencies around the world launched investigations of ChatGPT and AI experts called for regulation of AI services. The supplement covered the enforcement actions initiated in Italy, Canada, France, Australia, Germany, Spain, Japan, and the UK.

The First Supplement provided additional evidence from AI experts. It highlighted calls for regulation by federal agencies like CISA, NSA, and Department of Defense's Chief AI and Digital Officer who warned about the potential for generative artificial intelligence systems like ChatGPT to deceive citizens and threaten national security. The CAIDP Supplement cited corporate policies of Amazon, Samsung, Bank of America, Goldman Sachs, Wells Fargo, Citigroup and Deutsche Bank restricting employees from using ChatGPT due to privacy and data security concerns.

The First Supplement restated the prayer for relief sought in the Complaint. CAIDP urged the FTC to act without delay.

Summary of the Second Supplement[5]

In the Second Supplement, CAIDP described additional enforcement actions initiated in Korea, Brazil, Netherlands, Poland. The Second Supplement highlighted the aggressive business practices of OpenAI, contrary to warnings and cautions regarding accelerated deployment by AI experts. For example, OpenAI released GPTbot to scrape

[4] CAIDP, *Supplement to the Original Complaint to the FTC - In re OpenAI and ChatGPT,* (Jul. 10, 2023), https://www.caidp.org/app/download/8466615863/CAIDP-FTC-Supplement-OpenAI-07102023.pdf

[5] CAIDP, *Second Supplement to the Original Complaint to the FTC - In re OpenAI and ChatGPT,* (Nov. 14, 2023), https://www.caidp.org/app/download/8485816363/CAIDP-Supplement-FTC-OpenAI-11142023.pdf

the entire internet and the absence of any provenance measures for DALL-E3 and the propensity for generating 'racy content'.

Given the text to image capabilities being commercialized by OpenAI, the Second Supplement expanded upon the risks to democracy and elections, public safety risks, consumer concerns over deepfakes, voice clones, biometric privacy, fraud, and copyright abuse.

In the Second Supplement, CAIDP summarized a spate of class action lawsuits against OpenAI concerning the lack of transparency and unfair data practices. CAIDP also highlighted OpenAI's expansion of GPT-4 integration and the launch of voice, image capabilities of ChatGPT. CAIDP explained that these business practices raise concerns under FTC's Policy Statement on Biometric Information.

CAIDP urged the FTC to act. CAIDP cited OpenAI's accelerated deployment of GPT-4 notwithstanding documented and admitted risks is contrary to OpenAI's commitments to the administration and FTC's own business guidance.

II. The FTC Investigation into ChatGPT

In July 2023, the New York Times reported that the FTC had launched the investigation into OpenAI and ChatGPT sought by the Center for AI and Digital Policy.[6] The detailed document request made public by the Washington Post also indicated that the FTC identified copyright concerns in addition to the privacy and security risks CAIDP highlighted in its complaint.[7] The Wall Street Journal (WSJ) reported "In a civil subpoena to the company made public Thursday, the FTC says its investigation of ChatGPT focuses on whether OpenAI has "engaged in unfair or deceptive practices relating to risks of harm to consumers, including reputational harm."[8]

[6] New York Times, *F.T.C. Opens Investigation Into ChatGPT Maker Over Technology's Potential Harms,* (Jul. 13, 2023), https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html

[7] The Washington Post, *FTC investigates OpenAI over data leak and ChatGPT's inaccuracy*, (Jul. 13, 2023), https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/

[8] The Wall Street Journal, *ChatGPT Comes Under Investigation by the Federal Trade Commission*, (Jul.13, 2023), https://www.wsj.com/articles/chatgpt-under-investigation-by-ftc-21e4b3ef

In the document[9] made public by Washington Post, the FTC asked OpenAI to provide information about the data practices underlying the training of its large language model (LLM), the pre-release safety and risk assessment measures, the company's consumer marketing and advertising practices, its handling of users' personal information, and how the company offers downstream integrations of its GPT-4 product.

However, since that initial report of the investigation there is no further information on the investigation. The FTC's silence and delay is all the more troublesome because OpenAI, like many big tech firms, is cutting safety and security teams at the same time competition is increasing. Remarkably, experts inside and outside the company warn that the problems are far greater than the public is aware.[10]

IV. Consumer Concern over ChatGPT and LLM commercialization

When OpenAI released ChatGPT into the market, there were only 11 plugins available.[11] In a little over the year, the commercialization of its LLM GPT-4 has increased exponentially.

Accelerated commercialization and deployment

When the Washington Post reported on FTC investigating OpenAI, it noted "Analysts have called OpenAI's ChatGPT the fastest-growing consumer app in history, and its early success set off an arms race among Silicon Valley companies to roll out competing chatbots."[12]

AI products are evolving rapidly and being deployed downstream in consumer facing services. For example, ChatGPT is integrated with Snapchat used by many children and OpenAI has released GPTs which allows customization for any direct consumer use.

[9] Federal Trade Commission, *Civil Investigative Demand Schedule*, (FTC File No. 232-3044), https://www.washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf?itid=lk_inline_manual_4

[10] Merve Hickok, Christabel Randolph, Marc Rotenberg, *It's time for the FTC to act on ChatGPT*, Opinion, (Jun. 14, 2024), https://thehill.com/opinion/technology/4722343-its-time-for-the-ftc-to-act-on-chatgpt/

[11] CAIDP, *Complaint to the FTC - In re OpenAI and ChatGPT*, (Mar. 30, 2024), para.9 https://www.caidp.org/app/download/8450269463/CAIDP-FTC-Complaint-OpenAI-GPT-033023.pdf

[12] The Washington Post, *FTC investigates OpenAI over data leak and ChatGPT's inaccuracy*, (Jul. 13, 2023), https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/

ChatGPT's now augmented, multi-modal capabilities pose a significant threat to consumer safety, public safety, and election integrity.[13]

Following a proposed ban on using news publications and books to train AI chatbots in the U.K., OpenAI submitted a plea to the House of Lords communications and digital committee stating that it would be "impossible" to train AI models without using copyrighted materials, and that they believe copyright law "does not forbid training."[14]

Currently, OpenAI APIs are integrated into platforms such as Quizlet with more than 60 million students using it to study,[15] OpenAI launched a GPT store,[16] announced a new model GPT-4o that is multi-modal in input and output.[17] There are reports of GPT store being filled with spam and "several GPTs ripped from popular movie, TV and video game franchises"[18] and most recently, the eerie similarity of GPT-4o "Sky" voice with that of Scarlett Johansson resurfaced existing concerns over the business practices of AI companies in training their AI models.[19]

The Atlantic, Vox Media, Slack, Reddit, GitHub deals also show the aggressive commercialization practices that will now by default opt-in user data to train AI models.[20] What is more insidious is the anthropomorphization[21] of these systems which further augments the "dark pattern" effect and deceptive potential of GPT-4 tools.

[13] Statement of CAIDP and Encode Justice Re the FTC OpenAI Investigation, (Jan. 18, 2024), https://www.linkedin.com/posts/center-for-ai-and-digital-policy_aigovernance-consumerprotection-delayiscostly-activity

[14] TechCrunch, *ChatGPT: Everything you need to know about the AI-powered chatbot,* (Jun.17, 2024), https://techcrunch.com/2024/06/17/chatgpt-everything-to-know-about-the-ai-chatbot/

[15] OpenAI, *Introducing APIs for GPT-3.5 Turbo and Whisper*, (Apr. 24, 2024), https://openai.com/index/introducing-chatgpt-and-whisper-apis/

[16] OpenAI, *Introducing the GPT Store,* (Jan. 10, 2024), https://openai.com/index/introducing-the-gpt-store/

[17] OpenAI, *Hello GPT-4o,* (May 13, 2024), https://openai.com/index/hello-gpt-4o/

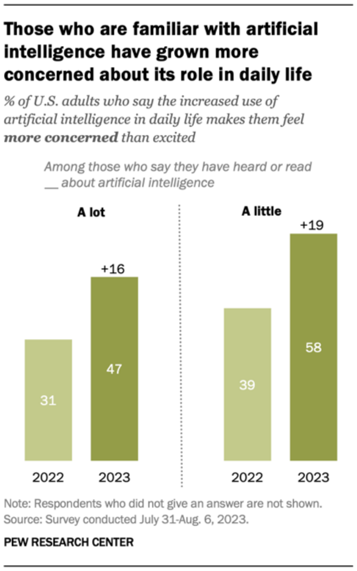
[18] TechCrunch, *OpenAI's chatbot store is filling up with spam,* (Mar. 20, 2024), https://techcrunch.com/2024/03/20/openais-chatbot-store-is-filling-up-with-spam/

[19] New York Times, *Scarlett Johansson's Statement About Her Interactions With Sam Altman,* (May 20, 2024), https://www.nytimes.com/2024/05/20/technology/scarlett-johansson-openai-statement.html

[20] TechCrunch, *ChatGPT: Everything you need to know about the AI-powered chatbot,* (Jun.17, 2024), https://techcrunch.com/2024/06/17/chatgpt-everything-to-know-about-the-ai-chatbot/

[21] Axios, *GPT-4o delivers human-like AI interaction with text, audio, and vision integration*, (May 14, 2023), https://www.artificialintelligence-news.com/2024/05/14/gpt-4o-human-like-ai-interaction-text-audio-vision-integration/

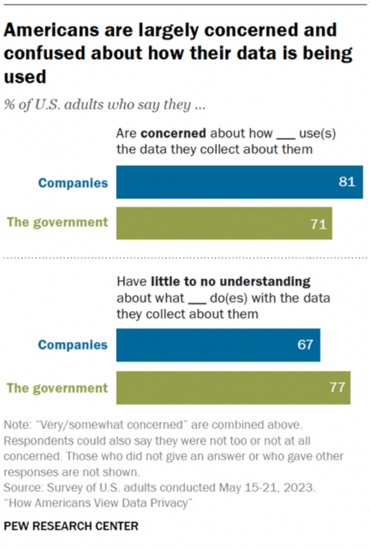
Public Opinion Surveys



Those who are familiar with artificial intelligence have grown more concerned about its role in daily life

% of U.S. adults who say the increased use of artificial intelligence in daily life makes them feel ***more concerned*** *than excited*

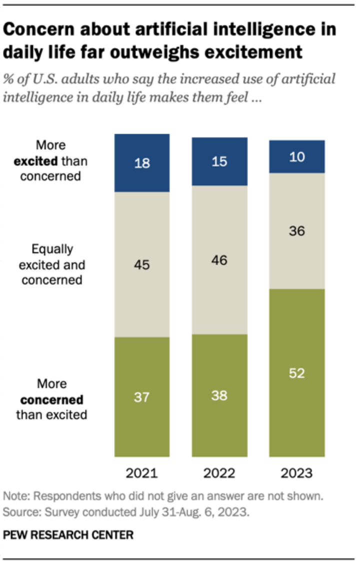
Note: Respondents who did not give an answer are not shown.
Source: Survey conducted July 31-Aug. 6, 2023.
PEW RESEARCH CENTER



Americans are largely concerned and confused about how their data is being used

% of U.S. adults who say they ...

Note: "Very/somewhat concerned" are combined above. Respondents could also say they were not too or not at all concerned. Those who did not give an answer or who gave other responses are not shown.
Source: Survey of U.S. adults conducted May 15-21, 2023. "How Americans View Data Privacy"
PEW RESEARCH CENTER



Concern about artificial intelligence in daily life far outweighs excitement

% of U.S. adults who say the increased use of artificial intelligence in daily life makes them feel ...

Note: Respondents who did not give an answer are not shown.
Source: Survey conducted July 31-Aug. 6, 2023.
PEW RESEARCH CENTER

The Pew Research Center in August 2023 reported that "Of those who have heard of ChatGPT, majorities of Democrats and Republicans say their greater concern is not enough regulation."[22] Just two months later in October 2023 Pew surveys showed that, "People's views on artificial intelligence (AI) are marked with distrust and worry about their data…As AI raises new frontiers in how people's data is being used, unease is high. Among those who've heard about AI, 70% have little to no trust in companies to make responsible decisions about how they use it in their products."[23]

The Pew polls all show growing public support for the regulation of AI products and services. "Democrats and Republicans alike are more concerned about insufficient government regulation of chatbots than excessive regulation."[24] Of those polled, 67% said the government would not go far enough to safeguard the public.

[22] Pew Research Center, *Most Americans haven't used ChatGPT; few think it will have a major impact on their job*, (Aug. 28, 2023), https://www.pewresearch.org/short-reads/2023/08/28/most-americans-havent-used-chatgpt-few-think-it-will-have-a-major-impact-on-their-job/

[23] Pew Research Center, *How Americans View Data Privacy,* (Oct. 18, 2023), https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/

[24] Pew Research Center, *Democrats and Republicans alike are more concerned about insufficient government regulation of chatbots than excessive regulation* (Aug. 28, 2023), https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/sr_23-11-21_ai-roundup_5-png/

Advocacy of EU Consumer Groups

On the other side of the Atlantic, following the CAIDP complaint to the FTC, the European Consumer Organisation (BEUC) called for EU and national authorities to launch an investigation into ChatGPT and similar chatbots.[25] The Norwegian Consumer Council (NCC) released an extensive report on the consumer harms of generative AI products which CAIDP cited in its supplementary complaint.[26] In the press release accompanying the report, the NCC stated "As long as the EU's AI Act is not applicable, authorities need to investigate where new generative AI-driven products and services may be harming consumers and enforce existing data protection, safety and consumer protection legislation."[27]

In March 2024, BEUC – the European consumer organization released the report on "Digital Fairness for Consumers"[28] which carefully lays out the agility required from consumer protection law to address economic and non-economic harm to consumers in the digital environment and considering the "engineering of consumer behavior" through AI systems.

V. Growing Concerns relating to OpenAI

As more individuals use AI to seek relationship advice, medical information or psychological counseling, experts say the risks to individuals are growing. In addition to potentially sharing specific pieces of data, generative AI tools can draw connections, or inferences providing a chillingly detailed understanding of our personhood.[29]

Amidst these growing concerns over accelerated commercialization of generative AI systems, the events surrounding OpenAI's leadership in November last year signals

[25] BEUC, *Investigation by EU authorities needed into ChatGPT technology,* (Mar. 30, 2023), https://www.beuc.eu/press-releases/investigation-eu-authorities-needed-chatgpt-technology
[26] CAIDP, *Supplement to the Original Complaint to the FTC - In re OpenAI and ChatGPT,* (Jul. 10, 2023), para. 110, https://www.caidp.org/app/download/8466615863/CAIDP-FTC-Supplement-OpenAI-07102023.pdf
[27] FORBRUKERRADET, *New report: Generative AI threatens consumer rights,* (Jun. 20, 2023), https://www.forbrukerradet.no/side/new-report-generative-ai-threatens-consumer-rights/
[28] BEUC, *Digital Fairness for Consumers*, (Mar. 2024), https://www.beuc.eu/sites/default/files/publications/BEUC-X-2024-032_Digital_fairness_for_consumers_Report.pdf
[29] Axios, *Generative AI's privacy problem,* (Mar. 14, 2024), https://www.axios.com/2024/03/14/generative-ai-privacy-problem-chatgpt-openai

the growing governance concerns relating to concentration of power both within and outside the company.[30]

<u>Concerns on internal governance</u>

Those most closely associated with OpenAI are now warning about a culture of recklessness and secrecy at the company at the same time it is racing to build the most powerful A.I. systems ever created.[31] Just this month, the New York Times reported,

> The members say OpenAI, which started as a nonprofit research lab and burst into public view with the 2022 release of ChatGPT, is putting a priority on profits and growth as it tries to build artificial general intelligence, or A.G.I., the industry term for a computer program capable of doing anything a human can.
>
> They also claim that OpenAI has used hardball tactics to prevent workers from voicing their concerns about the technology, including restrictive nondisparagement agreements that departing employees were asked to sign.[32]

Several current and former OpenAI and Google DeepMind employees warned about the lack of oversight for the artificial intelligence industry in a recent public letter.[33] The letter states "AI companies possess substantial non-public information about the capabilities and limitations of their systems, the adequacy of their protective measures, and the risk levels of different kinds of harm." This letter has been endorsed by leading AI experts Yoshua Bengio, Geoffrey Hinton, and Stuart Russell.

The letter from OpenAI employees echoed concerns raised by Helen Toner, a former OpenAI board member. Toner stated in an interview that OpenAI CEO Sam Altman was fired by the former board of directors because he provided inaccurate information about safety mechanisms, did not clear major product releases with the board and kept related investments confidential.[34]

[30] Dave Lee, *Sam Altman Exposes the Charade of AI Accountability,* Opinion, Bloomberg, (Nov. 20, 2023), https://www.bloomberg.com/opinion/articles/2023-11-20/openai-sam-altman-exposes-the-charade-of-ai-accountability

[31] The New York Times, *Insiders Warn of a 'Reckless' Race for Dominance*, The Shift, (Jun. 5, 2024), https://www.nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html

[32] *Id.*

[33] A Right to Warn about Advanced Artificial Intelligence, https://righttowarn.ai

[34] Merve Hickok, Christabel Randolph, Marc Rotenberg, It's time for the FTC to act on ChatGPT, Opinion, (Jun. 14, 2024), https://thehill.com/opinion/technology/4722343-its-time-for-the-ftc-to-act-on-chatgpt/

Concerns on safety policies and practices

The widely reported concerns about a culture of gagging and chilling speech affects not only employment practices but safety assessments and evaluations -critical for advancing safe, secure, and trustworthy AI. The concerns of employees and insiders are mirrored by concerns on assurances of red-teaming and safety testing at AI companies. OpenAI's current usage policy prohibits outside researchers from intentionally circumventing safeguards and mitigations "unless supported by OpenAI," and yet advocates for AI companies like OpenAI to create more opportunities for researchers to scrutinize their models. OpenAI does deploy a network of third-party red teamers to conduct adversarial research of their models, but researchers must apply to be part of the program, and OpenAI ultimately sets the rules of engagement. [35]

MIT led an open call by 350+ AI, legal, and policy experts for "A Safe Harbor for Independent AI Evaluation" citing concerns over current practices and policies of AI companies that can chill independent evaluation.[36]

VI. Enforcement in other jurisdictions

Overall, there have been **over a dozen** investigations of OpenAI in different countries, targeting various aspects of its services.[37] The range of enforcement actions can be categorized into three broad clusters:

- **Consumer Data Privacy**: Privacy violation due to the collection and processing of data to train AI models.
- **Inaccurate Content**: Harm arising from inaccurate content generated by OpenAI services.
- **Competition**: Consolidation of market share through Microsoft's investment in OpenAI.

Even where enforcement actions have not yet concluded, investigations act as important information collection mechanisms for regulators to clarify the practices of OpenAI and other AI services and to prepare for more targeted legal standards. Significantly, these actions demonstrate that consumer protection and competition

[35] Cyberscoop, *AI companies promise to protect our elections. Will they live up to their pledges?,* (May 15, 2024), https://cyberscoop.com/ai-companies-election-transparency/
[36] A Safe Harbor for Independent AI Evaluation, https://sites.mit.edu/ai-safe-harbor/
[37] Stephanie Psaila, *Governments vs ChatGPT: Investigations around the world*, DIPLO (Jun. 16, 2023), https://www.diplomacy.edu/blog/governments-chatgpt-investigations/.

enforcement are not alternative choices but rather complementary routes for ensuring consumer safeguards and preventing market concentration.

Since the launch of ChatGPT in 2022, Data Protection Authorities (DPAs) in France, Germany, Italy, Ireland, Netherlands, Poland, and Spain have all initiated their respective investigation against ChatGPT.[38] Other countries, including Canada and Brazil, have also initiated their investigations based on their own data security laws in response to public complaints.[39] The Canadian complaint focused on the use and collection of data without consent, while the Brazilian complaint focused on accessing personal data retained by ChatGPT and information about how they are utilized.[40] There have been no further updates since these countries launched their investigations in 2023.

Inaccurate content generated by AI models also comes under the purview of EU data privacy laws. For instance, the GDPR[41] requires data processor to accurately process personal data and grants data subjects the right to rectify incorrect personal data.[42] Inaccurate content generated by AI models could violate Art. 5, while the inability to rectify personal data may go against data subjects' right to rectify.

OpenAI has an "extended partnership" with Microsoft. Since 2019, OpenAI has received over $13 billion in investment from Microsoft.[43] The two companies also cooperate in supercomputing, AI service, and cloud service.[44] Such deals have raised concerns about whether Microsoft has effectively acquired OpenAI or achieved

[38] See Psaila, supra note 37; Reuters, *Dutch privacy watchdog seeks information from OpenAI, flags concerns*, (Jun. 7, 2023), https://www.reuters.com/technology/dutch-privacy-watchdog-seeks-information-openai-flags-concerns-2023-06-07/; TechCrunch, *Poland opens privacy probe of ChatGPT following GDPR complaint*, (Sept. 21, 2023), https://techcrunch.com/2023/09/21/poland-chatgpt-gdpr-complaint-probe/.

[39] Office of the Privacy Commissioner of Canada [OPC], *OPC launches investigation into ChatGPT*, (Apr. 4, 2023), https://www.priv.gc.ca/en/opc-news/news-and-announcements/2023/an_230404/; Pedrp Spadoni, *Após denúncia, agência brasileira vai fiscalizar ChatGPT*, Olhar Digital (Jul. 26, 2023), https://olhardigital.com.br/2023/07/26/seguranca/agencia-brasileira-vai-fiscalizar-chatgpt-apos-denuncia/.

[40] OPC, supra note 16; Luca Belli, *Why ChatGPT does not comply with the Brazilian Data Protection Law and why I petitioned the Regulator*, MEDIANAMA (May 25, 2023), https://www.medianama.com/2023/05/223-chatgpt-brazilian-data-protection-law-ai-regulation/.

[41] 2016 O.J. (L 119) 1, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL%3A2016%3A119%3ATOC [hereinafter GDPR]

[42] GDPR arts. 5(1)(d), 16.

[43] See Jason Karaian, *Microsoft's Stock Hits Record High After Hiring OpenAI Outcasts*, The New York Times (Nov. 20, 2023), https://www.nytimes.com/2023/11/20/business/microsoft-stock-openai.html.

[44] Microsoft, *Microsoft and OpenAI extend partnership*, Microsoft Corporate Blogs (Jan. 23, 2023), https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/.

dominance within the AI market. In the United States, the Federal Trade Commission has launched a general inquiry into generative AI investments and is reported to specifically launch an antitrust investigation of the OpenAI-Microsoft partnership.[45]

Both the United Kingdom and the European Union have initiated early-stage reviews of the partnership, based on their respective antitrust laws. In December 2023, the Competition and Markets Authority of the UK, its principal antitrust regulator, announced that it was inviting public comments on the OpenAI-Microsoft partnership.[46] The investigation focused on whether Microsoft's investment in OpenAI constitutes a merger under the UK Enterprise Act of 2002, and whether the partnership substantially hindered market competition.[47] The UK CMA has since not issued any updates.

Similarly, in January 2024, the European Commission officials signaled that it was considering whether Microsoft's investment in OpenAI would be subject to review under the EU merger rule, as well as the market impact of the OpenAI-Microsoft partnership.[48] In April 2024, Reuters reported that the investment would avoid a formal EU merger review, but Microsoft could still face an antitrust investigation.[49]

Example #1: Italy's Enforcement against OpenAI

Italy was the first to take enforcement action against OpenAI. The Garante, Italy's DPA, launched its investigation of OpenAI in March 2023 and *temporarily banned ChatGPT* in Italy.[50] The enforcement action was initiated after a data breach of ChatGPT user information on March 20, 2023.

[45] FTC, *FTC Launches Inquiry into Generative AI Investments and Partnerships*, Press Release, (Jan. 25, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships; Matt O'Brien, *US antitrust enforcers will investigate leading AI companies Microsoft, Nvidia and OpenAI*, AP (Jun. 6, 2024), https://apnews.com/article/nvidia-openai-microsoft-ai-antitrust-investigation-ftc-doj-0adc9a4a30d4b581a4f07894473ba548.

[46] Competition and Markets Authority, *Microsoft / OpenAI partnership merger inquiry*, GOV.UK (Dec. 8, 2023), https://www.gov.uk/cma-cases/microsoft-slash-openai-partnership-merger-inquiry.

[47] Id.

[48] European Commission, *Commission launches calls for contributions on competition in virtual worlds and generative AI*, (Jan. 9, 2024), https://ec.europa.eu/commission/presscorner/detail/en/ip_24_85.

[49] Foo Yun Chee, *Exclusive: Microsoft's OpenAI partnership could face EU antitrust probe, sources say*, Reuters, https://www.reuters.com/technology/microsofts-openai-partnership-could-face-eu-antitrust-probe-sources-say-2024-04-18/

[50] Garante per la protezione dei dati personali [GPDP], *Artificial intelligence: stop to ChatGPT by the Italian SA*, (Mar. 31, 2023), https://gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9872832.

The announcement alleged several data privacy violations of ChatGPT, including:

- the lack of user access to information about what and how personal data are collected and used (GDPR Art. 12);
- the collection of personal information without legal basis (GDPR Art. 5);
- the processing of inaccurate personal information (GDPR Art. 6), and;
- the lack of age verification system (GDPR Art. 8).[51]

Along with the ban, the Garante has required OpenAI to clarify its GDPR compliance measures and implement a series of compliance measures, such as:

- Implement an age verification system by September 2023 that filters users under 13 and users between 13 to 18 without parental consent.
- Provide accessible notice on the OpenAI website on how data are processed for the operation of ChatGPT, as well as the data subjects' rights (users and non-users).
- Remove all reference to "contractual performance," which is a legal basis for processing data under GDPR Art. 6, and rely instead on "consent" or "legitimate interests" as the legal basis for processing data.
- Create a mechanism for data subjects, including users and non-users, to submit objection to the processing of personal data, and request rectification or erasure of personal data.[52]

As a result, OpenAI announced its improvement measures, including displaying required information on its website, and adding options for EU ChatGPT users to remove personal data or opt out of using their own data to train AI models through ChatGPT's privacy portal.[53] OpenAI has also implemented an age verification feature for Italian users later in 2023.[54]

The Garante lifted the temporary ban on April 28, 2023, after OpenAI "addressed or clarified" concerns of the Garante.[55] In its press release after lifting the ban, the Garante

[51] *Id.*

[52] GPDP, *ChatGPT: Italian SA to lift temporary limitation if OpenAI implements measures*, (Apr. 12, 2023), https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9874751.

[53] GPDP*, ChatGPT: OpenAI reinstates service in Italy with enhanced transparency and rights for European users and non-users*, (Apr. 28, 2024), https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9881490.

[54] Frank Hersey, *ChatGPT adds age verification in Italy to satisfy privacy enforcement*, MLEX (Oct. 13, 2023), https://mlexmarketinsight.com/news/insight/chatgpt-adds-age-verification-in-italy-to-satisfy-privacy-enforcement.

[55] GPDP, *supra* note 27.

welcomed the changes made by OpenAI, but vowed to continue its factfinding on OpenAI's compliance.[56]

In January 2024, the Garante concluded that available evidence showed that OpenAI *has violated* several GDPR provisions but did not specify the provisions violated in the announcement.[57]

In addition, after OpenAI launched Sora, its new text-to-video AI model, in March 2024, Garante launched another investigation to request clarification from OpenAI.[58] The Italian DPA required OpenAI to provide information on whether the new AI model will be available to EU users, how data is collected, processed, and used to train Sora algorithms, and whether sensitive personal data are collected.[59] If OpenAI intends to provide the service to EU Users, Garante also required information on Sora's legal basis for processing data and how it would inform users about their data rights.[60]

Example #2: Korea's Enforcement Action concerning ChatGPT data breach

On July 27, 2023, PIPC, the South Korean national data protection authority, announced an enforcement action against OpenAI concerning a breach of ChatGPT Plus subscriber information.[61] On March 20, 2023, subscriber information including user names, email addresses, payment addresses, and credit card information was made available to other ChatGPT subscribers, due to a bug in an open-source library used by OpenAI.[62] 687 South Korean users were impacted by the data breach.

The legal basis of the enforcement is the Korean Personal Information Protection Act (PIPA). Article 29 of PIPA specifies a duty for "personal information controller[s]" to

[56] *Id.*
[57] Garante also stated that it will also consider the EDPB ChatGPT Task Force Determination. *See,* GPDP, *ChatGPT: Italian DPA notifies breaches of privacy law to OpenAI*, (Jan. 29, 2024), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9991867.
[58] GPDP*, Artificial intelligence: the Italian Data Protection Authority opens an investigation into OpenAI's 'Sora'*, (Mar. 8, 2024), https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9991867.
[59] *Id.*
[60] *Id.*
[61]Personal Information Protection Commission [PIPC], *PIPC Imposes Administrative Sanctions on OpenAI, Issuing Recommendations to Improve Data Privacy Practices*, (Jul. 27, 2024), https://www.pipc.go.kr/eng/user/ltn/new/noticeDetail.do?bbsId=BBSMSTR_000000000001&nttId=2271.
[62] The information was leaked through incorrectly addressed subscription confirmation emails, or through displaying. OpenAI*, March 20 ChatGPT outage: Here's what happened*, (Mar. 24, 2023), https://openai.com/index/march-20-chatgpt-outage/.

take "technical managerial, and physical measures" to safeguard personal information.[63] Meanwhile, Article 39-4 of PIPA requires notification by "a provider of information and communications services" to "relevant users" and PIPC no later than 24 hours "since he or she became aware of such fact."[64]

PIPC determined that OpenAI failed to meet the data breach notification requirement under PIPA.[65] Furthermore, PIPC also found other deficiencies and statutory violations of PIPA upon investigation.[66] These violations include the "failure to meet the statutory requirements for obtaining proper user consent" and "unclear descriptions about data controller-processor relationship and data disposal."[67]

PIPC imposed an administrative fine of 3.6 million Korean Won (approximately $3,000) against OpenAI for failing to meet the notification requirement.[68] PIPC also provided unspecified recommendations to OpenAI for PIPA compliance and declared to continue monitoring the implementation of these recommendations.[69]

PIPC has also requestioned information to assess the data privacy risks of OpenAI's services, including:

- How OpenAI collects and processes data;
- How it uses Korean language data to train its models;
- How it addresses legal and ethical concerns, and;
- How it handles data requests by users.[70]

Example #3: GDPR complaints against ChatGPT alleging hallucination

Two GDPR complaints in Austria and Poland were filed on the basis of outputs generated by ChatGPT.

NOYB – European Center for Digital Rights, a non-profit organization founded by Max Schrems, an Austrian privacy activist, submitted a complaint to the Austrian DPA in April 2024.[71] The complaint specifically targets the issue of hallucination, alleging that:

[63] Gaeinjeongbo bohobeop [Personal Information Protection Act] art. 29 (S. Kor.).
[64] *Id. at* art. 39-4.
[65] PIPC, *supra* note 38.
[66] *Id.*
[67] *Id.*
[68] *Id.*
[69] *Id.*
[70] *Id.*
[71] NOYB – European Center for Digital Rights, *Complaint* (Apr. 29, 2024), https://noyb.eu/sites/default/files/2024-04/OpenAI%20Complaint_EN_redacted.pdf

- The lack of viable access to personal data collected by OpenAI violates the transparency principle under Art. 12 and the right to access under Art. 15;
- The impossibility of erasing or rectifying inaccurate personal information generated by ChatGPT violates the obligation to accurately process personal information under Art. 5(1)(d). [72]

Polish privacy activist, Lukasz Olejnik, extensively corresponded with OpenAI via emails on issues concerning his data subject's rights, after ChatGPT generated inaccurate information about him.[73] He then filed a complaint to the Polish DPA against OpenAI in August 2023, alleging that:

- The inaccurate processing of personal information violated the obligations to process data with "lawfulness, fairness, and transparency" under Art. 5(1)(a) and the transparency principle under Art. 12;
- OpenAI failed to provide sufficient information about the sources, processing, and recipients of Olejnik's data, and violated his right to access under Art. 15;
- The inability of OpenAI to rectify inaccurate processed data about Olejnik's data violated his right to rectify under Art. 16;
- These blatant incompatibilities of ChatGPT with GDPR violated the "privacy by design" principle required under Art. 25(1).[74]

Both complaints requested their respective DPA to initiate investigations of Open AI and corrective measures (similar to injunctive relief), and the NOYB complaint has also requested fines.[75] The Polish DPA has since launched an investigation in response to the complaint, while the Austrian DPA has yet to respond.[76]

[72] *Id.* at 4.

[73] Maciej Gawronski, *Complaint Against Unlawful Processing of Personal Data,* (Aug. 29, 2023), https://lukaszolejnik.com/stuff/OpenAI_GDPR_Complaint_LO.pdf.

[74] *Id.* at 7.

[75] NOYB, *supra* note. 48 at 6; Gawronski, *supra* note 50 at 1.

[76] URZĄD OCHRONY DANYCH OSOBOWYCH [UODO], *The technology has to be compliant with the GDPR*, https://uodo.gov.pl/pl/138/2823.

Example #4: European Data Protection Board Taskforce Report on ChatGPT

The European Data Protection Board (EDPB) convened a ChatGPT Task Force to share information and coordinate investigations between DPAs. The EDPB Task Force released a preliminary report in May 2024.[77] In the report, the EDPB emphasized that:

> technical impossibility cannot be invoked to justify non-compliance with these requirements, especially considering that the principle of data protection by design set out in Article 25(1) GDPR shall be taken into account at the time of the determination of the means for processing and at the time of the processing itself.[78]

In its report the EDPB set out "preliminary views" of the investigation into ChatGPT and OpenAI's operations in the EU. In assessing the "lawfulness" of processing personal data it considered: "i) collection of training data (including the use of web scraping data or reuse of datasets), ii) pre-processing of the data (including filtering), iii) training, iv) prompts and ChatGPT output as well as v) training ChatGPT with prompts."[79] Among the observations of the EDPB, the following are worth emphasizing:

- [A]dequate safeguards play a special role in reducing undue impact on data subjects. Such safeguards could inter alia be technical measures, defining precise collection criteria and ensuring that certain data categories are not collected or that certain sources (such as public social media profiles) are excluded from data collection. Furthermore, measures should be in place to delete or anonymise personal data that has been collected via web scraping before the training stage. [80]

- If ChatGPT is made available to the public, it should be assumed that individuals will sooner or later input personal data. If those inputs then become part of the data model and, for example, are shared with anyone asking a specific question, OpenAI remains responsible for complying with the GDPR and should not argue that the input of certain personal data was prohibited in first place.[81]

- [D]ue to the probabilistic nature of the system, the current training approach leads to a model which may also produce biased or made up outputs. In addition, the

[77] European Data Protection Board [EDPB], Report of the work undertaken by the ChatGPT Taskforce (2024). [EDPB Report]
[78] EDPB Report, pg. 5
[79] EDPB Report, pg. 6
[80] EDPB Report, pg. 6
[81] EDPB Report, pg. 7

> outputs provided by ChatGPT are likely to be taken as factually accurate by end users, including information relating to individuals, regardless of their actual accuracy.[82]

VII. FTC's Mandate and Guidance on AI Products

The FTC is perhaps one of the most empowered consumer protection agencies in the world. Its broad mandate to protect consumers and ensure fair competition allows the agency to "prosecute any inquiry necessary to its duties in any part of the United States," FTC Act Sec. 3, 15 U.S.C. Sec. 43. The FTC is authorized "to gather and compile information concerning, and to investigate from time to time the organization, business, conduct, practices, and management of any person, partnership, or corporation engaged in or whose business affects commerce, excepting banks, savings and loan institutions . . . Federal credit unions . . . and common carriers . . . ." FTC Act Sec. 6(a), 15 U.S.C. Sec. 46(a).[83] The FTC has authority to investigate, prosecute, and prohibit "unfair or deceptive acts or practices in or affecting commerce."[84]

There is little question that the FTC's authorities apply to AI services. In New York Times op-ed, FTC Chair Lina Khan wrote, "Although these tools are novel, they are not exempt from existing rules, and the FTC will vigorously enforce the laws we are charged with administering, even in this new market."[85] Chair Khan has on several occasions reaffirmed that the FTC will ensure that "claims of innovation are not used as cover for lawbreaking".[86]

The FTC has issued several business guidance in relation to AI products and services.[87]

[82] EDPB Report, pg. 8

[83] FTC*, A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority* (May 2021), https://www.ftc.gov/about-ftc/mission/enforcement-authority

[84] 15 U.S.C. §45 (a)(1), (2), (4)(A), 4(B); (m)(1)(A); m(1)(B) ("Declaration of unlawfulness; power to prohibit unfair practices); (b) (proceedings by the Commission")

[85] Lina Khan, *We Must Regulate A.I. Here's How.*, Opinion, New York Times, (May 3, 2023), https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html

[86] Kyrsten Crawford, *FTC's Lina Khan warns Big Tech over AI,* SIEPR, (Nov. 3, 2023), https://siepr.stanford.edu/news/ftcs-lina-khan-warns-big-tech-over-ai

[87] FTC, *Chatbots, deepfakes, and voice clones: AI deception for sale*, (Mar. 20, 2023), https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale; FTC, *Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues,* FTC Report, (January 2016), https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report; FTC, *Using Artificial Intelligence and Algorithms*, Business Guidance, (Apr. 8, 2020), https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial- intelligence-and-algorithms; FTC, *Aiming for truth, fairness, and equity in your company's use of AI* , Business Guidance, (April 2021), https://www.ftc.gov/business-

<u>2020-2021</u>

In 2020, the FTC issued the Statement *Using Artificial Intelligence and Algorithm*.[88] In the 2020 statement, the Director of the FTC Consumer Protection Bureau said, "The FTC's law enforcement actions, studies, and guidance emphasize that the use of AI tools should be transparent, explainable, fair, and empirically sound, while fostering accountability."[89]

The 2020 FTC Statement set out recommended best practices, including:

> (a) Don't deceive consumers about how you use automated tools ("But, when using AI tools to interact with customers (*think chatbots*), be careful not to mislead consumers about the nature of the interaction.") (emphasis added)
> (b) Be transparent when collecting sensitive data ("Secretly collecting audio or visual data – or any sensitive data – to feed an algorithm could also give rise to an FTC action.")
> (c) Ensure that your data and models are robust and empirically sound.
> (d) Make sure that your AI models are validated and revalidated to ensure that they work as intended, and do not illegally discriminate
> (e) Consider your accountability mechanism ("Consider how you hold yourself accountable, and whether it would make sense to use independent standards or independent expertise to step back and take stock of your AI.")

In 2021, the FTC issued the *Statement Aiming for Truth, Fairness, and Equity in Your Company's use of AI*.[90] The 2021 FTC Statement said to businesses offering products with the AI techniques: "*As your company launches into the new world of artificial*

guidance/blog/2021/04/aiming-truth-fairness-equity-your- companys-use-ai; FTC, *For Business Opportunity Sellers, FTC says "AI" Stands for "Allegedly Inaccurate"*, FTC Business Blog (Aug. 22, 2023), https://www.ftc.gov/business-guidance/blog/2023/08/business-opportunity-sellers-ftc-says-ai-stands-allegedly-inaccurate; *Generative AI Raises Competition Concerns,* FTC Blog, (Jun. 29, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns

[88] FTC, *Using Artificial Intelligence and Algorithms*, Business Guidance, (Apr. 8, 2020), https://www.ftc.gov/business-guidance/blog/2020/04/using-artificial- intelligence-and-algorithms

[89] Id.

[90] FTC, *Aiming for truth, fairness, and equity in your company's use of AI* , Business Guidance, (April 2021) (emphasis below in the original), https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your- companys-use-ai;

intelligence, keep your practices grounded in established FTC consumer protection principles." The 2021 FTC Statement set out recommended best practices, including:

(a) Start with the right foundation ("design your model to account for data gaps, and – in light of any shortcomings – limit where or how you use the model.")
(b) Watch out for discriminatory outcomes ("It's essential to test your algorithm – both before you use it and periodically after that – to make sure that it doesn't discriminate on the basis of race, gender, or other protected class.")
(c) Embrace transparency and independence ("As your company develops and uses AI, think about ways to embrace transparency and independence – for example, by using transparency frameworks and independent standards, by conducting and publishing the results of independent audits, and by opening your data or source code to outside inspection."
(d) Don't exaggerate what your algorithm can do or whether it can deliver fair or unbiased results ("your statements to business customers and consumers alike must be truthful, non-deceptive, and backed up by evidence.")
(e) Tell the truth about how you use data (describing recent enforcement actions against Facebook and Everalbum for misleading consumers)
(f) Do more good than harm
(g) Hold yourself accountable – or be ready for the FTC to do it for you.

2022-2023

In February 2023, following the widespread public awareness of GPT-4, the FTC warned, "false or unsubstantiated claims about [an AI] product's efficacy are our bread and butter. . . You don't need a machine to predict what the FTC might do when those claims are unsupported."[91]

The FTC's March 2023 business guidance on AI deception makes clear the risk of cyber-crime, financial fraud using generative AI tools, and states "The FTC Act's prohibition on deceptive or unfair conduct can apply if you make, sell, or use a tool that is effectively designed to deceive – even if that's not its intended or sole purpose."[92] The guidance also sets out risks that developers should consider, primarily, "whether there are reasonably foreseeable risks of fraud or harm" and whether "developers are taking measures to effectively mitigate those risks" or whether "developers are over-relying on post-release detection".

[91] FTC, *Keep your AI claims in check,* (Feb. 27, 2023), https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check
[92] FTC, *Chatbots, deepfakes, and voice clones: AI deception for sale,* (Mar. 20, 2023), https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale

In the May 2023 business guidance on consumer trust and generative AI tools, the FTC states that, "Design or use of a product can also violate the FTC Act if it is unfair."[93] The guidance also states that: FTC staff is focusing intensely on how companies may choose to use AI technology, including new generative AI tools, in ways that can have actual and substantial impact on consumers.

In the October 2023 Guidance[94] the FTC specifically addressed the risks of large language models and stated:

> AI models are susceptible to bias, inaccuracies, "hallucinations," and bad performance. At the end of the day, AI model accuracy is dependent on a number of factors including the input data, training techniques, and context of deployment. Further, companies design applications to be efficient (using less resources, while yielding more output) in order to optimize for scalability and profit. This often means reducing the number of humans involved, leaving consumers to engage with their AI replacements.
>
> With the increasing sophistication of large language models, image generation systems, and more, it is becoming harder to distinguish human from machine. AI products could be used by malicious actors to increase the scale or sophistication of existing scams, another issue the FTC has written about before.[95]

The FTC has also issued business guidance that would guide the downstream uses and integrations of LLM products.[96] Relevant to the downstream integration of AI products, the recent deals with services like Slack, Reddit, and the automatic opt-in of user data for training AI models, the FTC states, "It may be unfair or deceptive for a company to adopt more permissive data practices—for example, to start sharing consumers' data with third parties or using that data for AI training—and to only inform consumers of this change through a surreptitious, retroactive amendment to its terms of service or privacy policy."[97]

[93] FTC, *The Luring Test: AI and the engineering of consumer trust,* (May 1, 2023), https://www.ftc.gov/consumer-alerts/2023/05/luring-test-ai-and-engineering-consumer-trust
[94] FTC, *Consumers Are Voicing Concerns About AI,* (Oct. 3, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/10/consumers-are-voicing-concerns-about-ai
[95] Id.
[96] FTC, *AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive*, (Feb. 13, 2024), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive
[97] Id.

Apart from business guidance, since 2019, the FTC has proposed five disgorgement orders including against Cambridge Analytica and Amazon.[98] In the 2023 Privacy and Security Report to Congress, the FTC stated:

> Artificial Intelligence: The Commission has been leading efforts to ensure that AI and similar technologies are not deployed in harmful ways. In addition to obtaining orders against Rite Aid, Ring, and Amazon to ensure that companies are disincentivized from using data that was wrongfully collected or trained to develop AI, we have initiated a market study of social media and video streaming platforms on the use of AI, announced a public contest to develop new approaches to protect consumers from AI-enabled voice cloning harms, proposed rules to crack down on AI-fueled impersonator and fake review fraud, and issued numerous business guidance alerts.[99]

The FTC has the mandate, the tools to enforce against generative AI systems, and has put companies on notice through its business guidance.

VIII. Limited AI legislation to protect people

There has been a lot of activity in Congress, but little action in advancing AI guardrails. "[M]ore than 300 AI-related proposed bills were introduced in this congressional session [beginning in January 2023]. They range all over the place, from controlling misinformation to how we can stimulate AI innovation and research."[100] However, only about 20 have moved passed a second reading at the respective committees. With a narrow window of meaningful congressional action remaining in the current session, it is imperative on regulatory agencies like the FTC to exercise their legislative mandate in the public interest.

In the absence of congressional action on AI legislation, what we are left with is a repeat of patchwork of legislative proposals for AI. As the National Conference of State Legislatures report: "In the 2024 legislative session, at least 40 states, Puerto Rico, the Virgin Islands and Washington, D.C., introduced AI bills, and six states, Puerto Rico and

[98] William Simpson, *AI Regulatory Enforcement Around the World*, IAPP News, (Aug. 2, 2023), https://iapp.org/news/a/ai-regulatory-enforcement-around-the-world

[99] The Federal Trade Commission, *2023 Privacy and Data Security Update,* pg. 1, https://www.ftc.gov/reports/federal-trade-commission-2023-privacy-data-security-update

[100] Nicola Jones, *The US Congress is taking on AI — this computer scientist is helping*, News Q&A, Nature, (May 9, 2024), https://www.nature.com/articles/d41586-024-01354-4#:~:text=There%20have%20been%20more%20than,stimulate%20AI%20innovation%20and%20research.

the Virgin Islands adopted resolutions or enacted legislation."[101] However, the ambit and content of the legislation also differs widely. While "Colorado required developers and deployers of high-risk AI systems to use reasonable care to avoid algorithmic discrimination and mandated disclosures to consumers"[102], "Tennessee required the governing boards of public institutions of higher education to promulgate rules and required local education boards and public charter schools to adopt policies, regarding the use of AI by students, teachers, faculty and staff for instructional purposes."[103]

President Biden's AI Executive Order on Safe, Secure, and Trustworthy AI[104] (AI EO), as commendable and extensive as it is, applies only to federal agencies. It introduces key guardrails for the government use of AI and establishes oversight on such use. The guidance from the Office of Management and Budget (OMB) builds on these protections by setting out clear criteria for "rights-impacting" and "safety-impacting" AI systems in the government and requires lifecycle assessment of AI systems.

There is some optimism that the federal government through the powers of its purse will be able to set some rules of the road for the private sector.[105] However, the AI EO doesn't apply to the private sector outside of certain water-marking and safety obligations, it cannot mandate any pre-deployment or priore impact assessments, or rules requiring that companies disclose training data sources, model size and other important details.[106] The durability of the Executive Order is also uncertain given that a change in administration could see the EO reversed.

But what is significant for the purposes of this report is that the Biden AI EO also calls upon the FTC to exercise its existing authorities to ensure that consumers and workers are protected from AI harms.[107]

[101] National Conference of State Legislatures, *Artificial Intelligence 2024 Legislation*, (Jun. 3, 2024), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation

[102] Id.

[103] Id

[104] Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 75209 (Oct. 30, 2023), https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf

[105] Sorelle Friedler, Janet Haven, Brian J. Chen, *How the AI Executive Order and OMB memo introduce accountability for artificial intelligence,* Commentary, Brookings Institution, (Nov. 16, 2023), https://www.brookings.edu/articles/how-the-ai-executive-order-and-omb-memo-introduce-accountability-for-artificial-intelligence/

[106] Axios, *What's in Biden's AI executive order — and what's not*, (Nov.1, 2023), https://www.axios.com/2023/11/01/unpacking-bidens-ai-executive-order

[107] Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 75209 (Oct. 30, 2023), https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf

> In particular, the Federal Trade Commission is encouraged to consider, as it deems appropriate, whether to exercise the Commission's existing authorities, including its rulemaking authority under the Federal Trade Commission Act, 15 U.S.C. 41 et seq., to ensure fair competition in the AI marketplace and to ensure that consumers and workers are protected from harms that may be enabled by the use of AI.[108]

IX. FTC Enforcement: The most viable solution for guardrails

CAIDP President, Merve Hickok, testified at one of the first congressional hearings last year and stated, "We do not have the guardrails in place, the laws that we need, the public education, or the expertise in government to manage the consequences of the rapid changes that are now taking place."[109]

At the House Oversight and Accountability Committee hearing in July last year, Chair Khan, stated

> *"As the nation's primary consumer protection agency, the FTC has a broad mandate to protect the public from unfair or deceptive practices throughout the economy…. The Commission will vigorously use the full range of our authorities to protect consumers from deceptive and unfair conduct and maintain open, fair, and competitive markets in this rapidly evolving technology. Through blog posts and other public pronouncements, the agency is providing timely analysis to market participants and the public. The Commission is poised to move aggressively against businesses that engage in deceptive or unfair acts involving AI and to help ensure that illegal practices do undermine competition and innovative uses of AI."*[110]

[108] Id. at sec. 5.3.s

[109] Testimony and statement for the record of CAIDP President Merve Hickok, Advances in AI: Are We
Ready For a Tech Revolution?, House Committee on Oversight and Accountability: Subcommittee on
Cybersecurity, Information Technology, and Government Innovation (Mar. 8, 2023), https://oversight.house.gov/wp-content/uploads/2023/03/Merve-Hickok_testimony_March-8th-2023.pdf

[110] Prepared Statement of the Federal Trade Commission, Hearing on "Oversight of the Federal Trade Commission", Committee on the Judiciary, United States House of Representatives, (Jul. 13, 2023), https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/khan-testimony.pdf

In November 2023, the FTC adopted an omnibus resolution to streamline the agency's ability to issue civil investigative demands relating to "products and services that use or are produced using artificial intelligence."[111]

In comments to the U.S. Copyright Office, the FTC stated "The FTC has been exploring the risks associated with AI use, including violations of consumers' privacy, automation of discrimination and bias, and turbocharging of deceptive practices, imposters schemes and other types of scams."[112] Most recently the National Association of Voice Actors (NAVA) issued a public statement in support of the CAIDP complaint regarding OpenAI and ChatGPT and called upon the FTC to complete its investigation with urgency.[113]

ChatGPT was released in the market in November 2022.[114] ChatGPT released its GPT-4 system card in March 2023[115], when it had already amassed an estimated 100 million monthly users.[116] The technical report setting out the risks of the product was issued only after the product was commercially released in the market and OpenAI began monetizing it. This was clearly contrary to FTC's established business guidance to ensure compliant products prior to release.

[111] Alan Raul, Alexandra Mushka, The U.S. Plans to 'Lead the Way' on Global AI Policy, LAWFARE, (Feb. 13, 2024), https://www.lawfaremedia.org/article/the-u.s.-plans-to-lead-the-way-on-global-ai-policy; *See also*, FTC, FTC Authorizes Compulsory Process for AI-related Products and Services, Press Release, (Nov. 21,2023), https://www.ftc.gov/news-events/news/press-releases/2023/11/ftc-authorizes-compulsory-process-ai-related-products-services

[112] FTC, In Comment Submitted to U.S. Copyright Office, FTC Raises AI-related Competition and Consumer Protection Issues, Stressing That It Will Use Its Authority to Protect Competition and Consumers in AI Markets, Press Release (Nov. 7, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/11/InCommentSubmittedtoUSCopyrightOfficeFTCRaisesAIrelatedCompetitionandConsumerProtectionIssuesStressingThatItWillUseItsAuthoritytoProtectCompetitionandConsumersinAIMarkets

[113] National Association of Voice Actors (NAVA), Public Statement of the National Association of Voice Actors, We need the FTC to act now - Complete the investigation into OpenAI, Press Release (Jun. 18, 2024), https://navavoices.org/press-releases/

[114] OpenAI, *Introducing ChatGPT*, (Nov. 30, 2022), https://openai.com/index/chatgpt/

[115] OpenAI, *GPT-4 Technical Report (2023)*, https://cdn.openai.com/papers/gpt-4.pdf

[116] Reuters, *ChatGPT sets record for fastest-growing user base - analyst note,* (Feb. 2, 2023), https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ ; The Verge, *ChatGPT continues to be one of the fastest-growing services ever,* (Nov. 6, 2023), https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference

AI companies including OpenAI have made lofty voluntary commitments, most recently at the Seoul AI Safety Summit.[117] The international scientific report that preceded the Seoul summit led by Yoshua Bengio highlighted that "General-purpose AI can pose severe risks to individual and public safety and wellbeing."[118] The report concludes that, "Despite rapid advances in capabilities, researchers currently cannot generate human-understandable accounts of how general-purpose AI models and systems arrive at outputs and decisions. This makes it difficult to evaluate or predict what they are capable of, how reliable they are, and obtain assurances on the risks they might pose."[119]

These commitments build upon the voluntary commitments by the industry to the Biden-Harris Administration "toward safe, secure, and transparent development of AI technology."[120] However, as OpenAI poignantly notes these "they apply only to generative models that are overall more powerful than the current industry frontier (e.g. models that are overall more powerful than any currently released models, including GPT-4, Claude 2, PaLM 2, Titan and, in the case of image generation, DALL-E 2).[121] Alarmingly, AI companies are already reneging on their voluntary commitments to provide AI safety institutes pre-deployment access to their models.[122]

History shows vague and unenforceable promises are not enough.[123] When Facebook acquired WhatsApp it acquired user data contrary to promises that it would not or could not integrate databases, and also palpably in violation of the terms of the 2011

[117] CNBC, Tech giants pledge AI safety commitments — including a 'kill switch' if they can't mitigate risks, (May 21, 2024), https://www.cnbc.com/2024/05/21/tech-giants-pledge-ai-safety-commitments-including-a-kill-switch.html

[118] AI Seoul Summit, International Scientific Report on the Safety of Advanced AI, Interim Report, (May, 2024), pg. 12, 13, https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf

[119] Id, pg. 83.

[120] The White House, *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*, (Jul. 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

[121] OpenAI, *Moving AI Governance Forward*, (Jul.21, 2023), https://openai.com/index/moving-ai-governance-forward/

[122] Politico, *Rishi Sunak promised to make AI safe. Big Tech's not playing ball*, (Apr. 26, 2024), https://www.politico.eu/article/rishi-sunak-ai-testing-tech-ai-safety-institute/

[123] David C. Vladeck, *Facebook, Cambridge Analytica, and the Regulator's Dilemma: Clueless or Venal?*, Administrative Law, Blog Essay, Harvard Law Rev., (Apr. 4, 2018), https://harvardlawreview.org/blog/2018/04/facebook-cambridge-analytica-and-the-regulators-dilemma-clueless-or-venal/

consent order with the FTC.[124] This was after Marc Zuckerberg publicly apologized to Congress for the Cambridge analytica debacle. At the same time, a New York Times investigation Facebook had secret deals with numerous companies for access to user data, including in some cases the contents of millions of users' private messages.[125] But even after the 2011 consent order, it took the agency almost a decade to act on Facebook's violations and egregious business practices.[126]

We see a repeat of this playbook from the tech industry. Sam Altman testified in Congress asking for AI legislation[127] while OpenAI lobbied against the provisions of the EU AI Act.[128] "A.I. companies are playing governments off one another. In Europe, industry groups have warned that regulations could put the European Union behind the United States. In Washington, tech companies have cautioned that China might pull ahead.[129]

There has been a lot of governance-washing of the *Tech Accord to Combat Deceptive Use of AI in 2024 Elections* in which Companies commit to manage the risks arising from deceptive AI election content "in line with their own policies"[130] and yet Sen. Warner has issued letters including to OpenAI asking them what measures exactly are put in place pursuant to this accord.[131]

Relying on AI companies to police themselves is not only foolhardy but does nothing to advance responsible innovation. "If the FTC had stood behind its commitment

[124] Marc Rotenberg, *The Facebook-WhatsApp Lesson: Privacy Protection Necessary for Innovation,* Worth Magazine, (May 4, 2018), https://worth.com/facebook-whatsapp-lesson-privacy-protection-necessary-innovation/

[125] Marc Rotenberg, *After Latest Facebook Fiasco, Focus Falls on Federal Commission,* Worth Magazine (Dec. 21, 2018), https://worth.com/after-latest-facebook-fiasco-focus-falls-on-federal-commission/; The New York Times, *Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis,* (nov. 14, 2018), https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html

[126] FTC, *Facebook, Inc., In the Matter of, https://www.ftc.gov/legal-library/browse/cases-proceedings/092-3184-182-3109-c-4365-facebook-inc-matter*

[127] Written Testimony of Sam Altman, Chief Executive Officer, OpenAI, Before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law, (May 15, 2023), https://www.judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Bio%20&%20Testimony%20-%20Altman.pdf

[128] Time, *Exclusive: OpenAI Lobbied the E.U. to Water Down AI Regulation*, Time Exclusive, (Jun. 20, 2023), https://time.com/6288245/openai-eu-lobbying-ai-act/

[129] *See,* New York Times, *How Nations Are Losing a Global Race to Tackle A.I.'s Harms,* (Dec. 6, 2023), https://www.nytimes.com/2023/12/06/technology/ai-regulation-policies.html?searchResultPosition=9

[130] AI Elections Accord, https://www.aielectionsaccord.com

[131] Letters issued by Sen. Mark Warner, (May 14, 2024), https://www.warner.senate.gov/public/_cache/files/3/e/3e12f60b-3e2f-4ab7-ade4-d819be943bde/7361EB3F33D404A03447E6FBD244D62D.full-munich-letters-pdf-final-3-.pdf

to protect the data of WhatsApp users, there might still be an excellent messaging service, with end-to-end encryption, no advertising and minimal cost, widely loved by internet users around the world. But the FTC failed to act and one of the great internet innovations has essentially disappeared."[132]

History shows that the longer the FTC delays, the more difficult it is to establish the necessary guardrails. Inaction by the agency is costly and FTC enforcement is the most immediate viable option for establishing guardrails for the AI industry. The FTC must act now.

ABOUT CAIDP

The Center for AI and Digital Policy (CAIDP)[133] is a non-profit, independent research, education, and advocacy organization based in Washington D.C. and Brussels. CAIDP aims to ensure that artificial intelligence and digital policies promote a better society, more fair, more just, and more accountable – a world where technology promotes broad social inclusion based on fundamental rights, democratic institutions, and the rule of law.

ABOUT THIS REPORT

This report was prepared by CAIDP Associate Director Christabel Randolph with the assistance of Victor Liu, Research Assistant, CAIDP.

[132] Marc Rotenberg, *The Facebook-WhatsApp Lesson: Privacy Protection Necessary for Innovation,* Worth Magazine, (May 4, 2018), https://worth.com/facebook-whatsapp-lesson-privacy-protection-necessary-innovation/
[133] CAIDP, https://www.caidp.org

BIG TECH BACKSLIDE

How Social-Media Rollbacks Endanger Democracy Ahead of the 2024 Elections

Written by Nora Benavidez

A Report from Free Press
December 2023

Free Press

TABLE OF CONTENTS

EXECUTIVE SUMMARY

After the deadly insurrection at the U.S. Capitol on Jan. 6, 2021, tech companies finally seemed to accept that their failure to moderate content was undermining public safety and democracy. Most companies removed users who spread anti-democratic conspiracies or used their online platforms to incite violence. Leading up to the 2022 U.S. midterm elections, tech companies promised users, civil-society groups and governments that they would safeguard election integrity and free expression on their platforms.

In 2022, Free Press graded the four major platforms' policies against 15 recommendations we helped develop that are designed to curb the spread of election disinformation and extremism. Our research found that although tech companies have long promised to fight disinformation and hate, there is a notable gap between what the companies say they want to do and what they actually do in practice. Companies like Meta, TikTok, Twitter and YouTube do not have sufficient policies, practices, AI systems or human capital in place to materially mitigate harm ahead of, during and after election periods.[1]

These failures have only mushroomed since we released our report. We found that in 2023, **the largest social-media companies have deprioritized content moderation and other user trust and safety protections, including rolling back platform policies that had reduced the presence of hate, harassment and lies on their networks. These companies have also laid off critical staff and teams tasked with maintaining platform integrity.** Taken together with the preferential treatment of VIP users — reflected in the reinstatement of Donald Trump's accounts on Meta, Twitter and YouTube — these developments represent a dangerous backslide. In turn, **this has created a toxic online environment that is vulnerable to exploitation from anti-democracy forces, white supremacists and other bad actors.**

POLICY ROLLBACKS DEPRIORITIZE USER SAFETY & PLATFORM INTEGRITY

Meta, Twitter and YouTube — the three biggest platforms — have rolled back crucial policies that had previously kept hate, harassment and lies in check. This waning commitment to content moderation has led to a spike in hate and disinformation, eroding people's experiences of these products and posing a risk to user safety.

Between Nov. 1, 2022 and Nov. 1, 2023, Meta, Twitter[2] and YouTube eliminated a total of 17 critical policies across their platforms:

- Twitter and YouTube rolled back election-misinformation policies designed to limit "Big Lie" content about the 2020 election.
- Twitter and Meta rolled back policies that had banned COVID-19 disinformation.
- Twitter began allowing political ads on the platform. Meta stopped applying a political-ad policy that had mandated transparency and labeling in such advertisements.[3] Both policies will allow for heightened disinformation in the ads users see on these platforms.
- Meta and Twitter have both weakened privacy protections for users to allow for greater use of people's data for the training of AI tools. Google has started giving its Bard AI tool access to YouTube, raising questions about how this chatbot will exploit user data.[4]
- Meta, Twitter and YouTube reinstated Donald Trump's accounts despite his outsized role in supporting and fueling the Jan. 6 insurrection.[5] Meta and YouTube have refused to apply their policies as stringently to Trump as they do to everyday users.
- Notably, TikTok has not rolled back any of its policies; in fact, in some instances it has strengthened platform features and policies. Yet TikTok remains unprepared to responsibly moderate toxic content on its platform. It enforces its policies in a lackluster manner and, at times, has downranked activists' posts that do *not* appear to violate its policies.[6]

MASS LAYOFFS MAKE CONTENT MODERATION HARDER

Mass layoffs across critical teams signal that these platforms are deemphasizing content moderation and user safety:

Meta, Twitter and YouTube have laid off approximately 40,750 employees across the three companies.[7] Significant cuts occurred in the trust and safety, ethical engineering or responsible innovation, and content-moderation/consultant categories — the teams otherwise tasked with maintaining the platforms' general health.

LITIGATION TO CHILL TECH ACCOUNTABILITY

Against the backdrop of these rollbacks, there is a looming new threat to platform integrity: Elon Musk has filed lawsuits against independent researchers in an attempt to silence criticism and evade accountability.[8]

There are dangerous real-world consequences when companies retreat from previous commitments to platform integrity, content moderation and robust enforcement of their terms of service.[9] Insurrectionists challenged the results of the 2020 U.S. presidential election in part due to conspiracy theories that the platforms amplified.[10] The entities overseeing 13 presidential libraries — Democratic and Republican alike — warned in September that U.S. democracy is in a fragile state, with civility and public discourse under grave threat.[11] And the platforms' failures to curb hate and disinformation related to the Israel-Hamas war has fueled mass violence.[12]

Whether it's in times of crisis or calm, social media has typically been a source of real-time information for users. And people should be able to rely on the platforms they use to provide accurate and even lifesaving information. The failure to vet and remove violative content harms and alienates users. Failure to moderate content inevitably leads to migration of platform lies and toxicity to mainstream media.[13]

Since the Jan. 6 insurrection, other real-world crises — like the attempted coup in Brazil in January 2023 and the conflict in the Middle East — have illustrated the critical role social-media platforms play in shaping rapidly unfolding events. Over and over we've seen how people can weaponize social media to sow division, undermine democracy and even fuel calls for violence offline.[14]

> **Without the policies and teams they need to moderate violative content, platforms risk amplifying confusion, discouraging voter engagement and creating opportunities for network manipulation to erode democratic institutions.**

TOPLINE RECOMMENDATIONS

It’s imperative that platforms redouble their previous efforts to root out lies, hate and violence. Companies should:

- Reinvest in and bolster staffing of teams needed to safeguard election integrity, trust and safety, and moderation.
- Immediately reinstate protective policies to moderate election-related content and political ads, provide labeling transparency, enforce bans on COVID disinformation, and offer specific safeguards against targeted harassment.
- Launch 2024 election-specific platform interventions in time for the U.S. primaries — and keep these protections in place through at least February 2025.
- Hold VIP accounts to the same enforcement standards applied to other users.
- Develop and implement more efficient review, labeling and enforcement against falsehoods in political ads across languages.
- Develop improved transparency and disclosure practices, including regularly sharing core-metrics data with external researchers, journalists, lawmakers and the public. Provide quarterly reports on key trends, virality reports, network analysis and more.

BIG TECH'S EMPTY PROMISES

Elections aren't happening just in the United States in 2024, when 40 national elections will occur worldwide.[15]

Major social-media companies have long failed when it comes to applying corporate policies robustly and equitably across the globe.[16] Platform executives have met civil society's requests for improvement with indignation, denial and inaction. Congressional hearings to investigate possible liability have yielded little insight into platform business practices.[17] Letters from legislators and dignitaries around the world asking for data on their algorithms, enforcement and staffing have resulted in few meaningful disclosures.[18]

For years, dozens of public-interest groups like Free Press have offered recommendations and expert guidance to the major social-media platforms. But these companies have responded with inconsistent and often lackluster commitments to reform.[19] However, the persistent pressure from organizations and our civil-rights coalitions — such as Change the Terms and #StopToxicTwitter, both of which Free Press helps lead — has resulted in some powerful wins:

1. **Strengthening Meta's Dangerous Organizations and Individuals Policy** Meta, then Facebook, has made several changes to its policies in response to pressure from the Change the Terms coalition. In 2019, Facebook changed its dangerous-organizations policy to include white nationalists. In 2020 — after Free Press and allies in the Stop Hate for Profit coalition organized an advertising boycott — the platform added QAnon under its dangerous-organizations policy and began enforcing its policy to remove violent, extremist QAnon content.[20]

2. **Prioritizing Equity in Meta's Moderation Across Languages** For years, Free Press and allies pushed social-media platforms to invest significant resources in combating hate and disinformation in languages other than English. After Free Press launched the #YaBastaFacebook campaign, and whistleblower Frances Haugen revealed failures in Meta's enforcement of non-English content, the company finally committed to fighting misinformation across all languages. We're continuing to push Meta and other platforms to enforce their policies equitably and to protect non-English users.

3. **Providing Transparent Access for Researchers** In a big win for transparency, TikTok announced in 2022 that it will allow researchers to delve into its data, evaluate its content and test its moderation system. To prove that it operates independently of the Chinese government, the company also announced that it will allow Oracle to audit its algorithms and content-moderation models. In another win, YouTube invited researchers to apply for access to its global data.

4. **Advancing Platform Integrity at Pre-Musk Twitter** Following meetings with Change the Terms leaders in 2018, Twitter banned deadnaming and misgendering.[21] In 2019, following our #StopRacistTwitter initiative and public protest outside the company's headquarters, Twitter bolstered its hateful-conduct policies to rein in violent and deceitful language. The company also banned political advertising, citing support from civil society and the platform's Trust and Safety Council.[22] In 2020, it expanded its disinformation policy to include COVID disinformation and provided a clear label for misleading content.[23]

5. **Preserving Election-Integrity Measures on Twitter Following Musk's Purchase** Free Press Co-CEO Jessica J. González and close allies met directly with Musk soon after he assumed ownership of the platform in October 2022 and got him to pledge that he wouldn't reinstate banned accounts before the midterm elections. When Musk gutted content-moderation policies and laid off thousands of key employees, Free Press partnered with Accountable Tech, Color Of Change, Media Matters for America and dozens of other allies to launch the #StopToxicTwitter coalition, which helped push more than 50 percent of Twitter's top-100 advertisers to pause their spending on the platform. These efforts slowed the pace of Musk's Twitter rollbacks and reduced the chance that other platforms would weaken their own moderation efforts ahead of the 2022 midterms.

6. **Limiting Misinformation on YouTube** After the 2018 launch of Change the Terms, the coalition regularly called on YouTube to strengthen its policies to rein in violent, hateful lies. In 2019, YouTube committed to stop recommending content that contained misinformation. It later announced that it would crack down on neo-Nazi content more aggressively.[24]

> **Meaningful platform reforms require ongoing advocacy from external experts and activists, as well as scrutiny from independent researchers with comprehensive access to platform data.**

Therefore, it is essential to continually put pressure on social-media platforms and other tech companies to equitably and effectively protect the integrity of their products. Absent external accountability and inquiry, we would know even less about these companies' opaque practices.[25]

Free Press investigated the state of platform integrity at major tech companies in the 2022 report *Empty Promises*, in which we reviewed the policies of the four largest social-media platforms to consider how prepared, both in writing and in practice, the companies were for the 2022 midterm elections.[26] Our research found that although the largest tech companies long promised to fight disinformation and hate on their platforms, they failed to take adequate measures in the run-up to the 2022 midterms.

Companies like Meta, TikTok, Twitter and YouTube have failed to put sufficient policies, practices, automated systems and human capital in place to materially mitigate harm ahead of and during elections. To further complicate matters, these companies have created a labyrinth of commitments, announcements and policies that make it almost impossible to assess what they're actually doing, if anything, to protect users.

After the 2022 midterms, the major platforms provided virtually no updates on the effectiveness of their policies nor any insights from their data about key network vulnerabilities. Meta provided no public summary. TikTok also failed to provide any publicly available summaries or reporting, though it issued brief community-guidelines updates. Twitter has no publicly available writing on the 2022 midterms. (Elon Musk took over Twitter just a week before these elections.) Alphabet, YouTube's parent company, provided a short blog post about the 2022 election period, with no insights and scant details about YouTube.

In the absence of transparent reporting, one thing is certain: The platforms' election-related policies and safety functions remain insufficient.[27]

BIG TECH BACKSLIDE

With dozens of national elections happening around the world in 2024, platform-integrity commitments are more important than ever. However, major social-media companies are not remotely prepared for the upcoming election cycle.

Free Press has observed a notable drop in the promises these companies are making to users as well as a significant rollback in concrete measures companies once had in place. Meta, Twitter and YouTube have all removed long-standing and critical policies, laid off staff and entire teams, and reinstated and even monetized violative accounts.

UNDERSTANDING THE NATURE OF PLATFORM ROLLBACKS

This backslide is not neutral in nature. The policies and teams that platforms deprioritized are key to understanding what these companies value and what they do not. All of the policy rollbacks across Meta, Twitter and YouTube deemphasize user safety and platform integrity, creating an opening for lies, hate and harassment to thrive. These three platforms have collectively laid off at least 40,750 workers, with massive cuts to trust and safety, content moderation, ethical AI and other teams tasked with maintaining user safety, content moderation and overall platform functionality.

Meta, Twitter and YouTube retreated from promises, policies and other actions to mitigate harm on their platforms in three main ways:

POLICY ROLLBACKS

From Nov. 1, 2022 to Nov. 1, 2023, Meta, Twitter and YouTube removed a total of 17 policies they had had in place prior to the 2022 midterm elections.

MASS LAYOFFS

All three companies have laid off tens of thousands of employees, totaling more than 40,750 across the three companies. Significant numbers of cuts were in the trust and safety, ethical engineering or responsible innovation, and content-moderation/consultant categories. Twitter removed its trust and safety team altogether.

REINSTATED BANNED ACCOUNTS

Meta, Twitter and YouTube reinstated Donald Trump's accounts. Twitter also reinstated thousands of previously suspended accounts, including those belonging to white supremacists, conspiracy theorists, misogynists and others promoting hateful rhetoric.

Taken together, these rollbacks over the past 12 months constitute an undeniable Big Tech backslide.

> **This backsliding fosters less accountability across each of these platforms as companies turn their backs on years of evidence pointing to the crucial and outsized role they play in bringing people information, shifting their attitudes, and shaping discourse that affects civic engagement and democracy.**

Of the platforms we examined, Twitter has rolled back the most policies and conducted the greatest ratio of layoffs-to-total-staff size. Meta has been a close second. And while YouTube had the fewest number of rollbacks over the last year, its policies were the weakest to begin with, as Free Press documented in earlier research. This points to the need for YouTube to reinstate and strengthen its policies overall.[28]

PLATFORM ROLLBACKS ↩	∞	𝕏	▶
STOPPED MODERATING "BIG LIE" CONTENT		↩	↩
WEAKENED POLITICAL-AD POLICIES	↩ stopped enforcing political ads policy, leaving an opening for bad actors to push lies in ads, which do not receive the same moderation treatment as user content	↩ began allowing political ads on the platform	↩ relaxed advertising policies to allow monetization for more graphically violent content
STOPPED MODERATING COVID LIES	↩	↩	
WEAKENED PRIVACY POLICIES REGARDING AI ACCESS	↩ new generative AI features on Meta platforms will draw on user data to train AI models	↩ removed privacy policies to begin using any and all user data to train AI models. Twitter created new policy language allowing it to collect users' biometric data	↩ Google has started giving its Bard AI tool access to YouTube, raising questions about how the chatbot will exploit user data
IMPOSED USER FACT-CHECKING LIMITS	↩ began allowing users to opt out of its fact-checking program	↩ disabled features that allow users to report election disinformation except in the European Union, where Twitter must comply with regional regulation	
ROLLED BACK DEADNAMING POLICY		↩	
WEAKENED USER PENALTIES FOR VIOLATING PLATFORM POLICIES			↩ weakened its three-strike policy for violative content, allowing strikes on violative content to be scrubbed after 90 days and completion of an educational course
LAID OFF CONTENT MODERATORS AND/OR TRUST & SAFETY TEAMS	↩	↩	↩
REINSTATED TRUMP	↩	↩	↩
REINSTATED OR MONETIZED PREVIOUSLY SUSPENDED DANGEROUS ACCOUNTS		↩	↩

POLICY ROLLBACKS JEOPARDIZE PLATFORM INTEGRITY

Social-media companies' written policies should provide clarity for users. These policies detail the way platforms moderate content, enforce their rules, deploy automated tools, and otherwise interact with users and their content to keep their services functional, authentic and useful. But these policies are inaccessible to most users. Instead of housing their rules in a centralized and accessible location, platforms often present them in an unruly patchwork of terms of service, community guidelines and standards, blog posts and tweets.

For well over a decade, civil-rights groups, civil-society organizations, lawmakers, tech ethicists and researchers have all tried to counsel the largest social-media platforms to make these policies clear, enforceable and equitable. They've urged platforms to protect public safety, public health, democracy and free expression.

Change the Terms, a coalition anchored by Free Press, the Center for American Progress and the Global Project Against Hate and Extremism, was founded to disrupt platforms' amplification of hate, extremism and lies, which plague the internet and endanger targeted groups.[29]

Despite our best efforts, Meta, Twitter and YouTube removed a total of 17 policies between Nov. 1, 2022 and Nov. 1, 2023.

These actions signal a troubling step backward. Election lies, COVID and wartime disinformation, and harassment remain threats to public safety, public health and democracy. The major platforms were right to develop policies and procedures to flag, review, downrank and sometimes remove this content. Rolling back these policies creates a disaster each time a major current event captures public attention.[30]

MASS LAYOFFS ERODE PLATFORM FUNCTIONALITY & MODERATION

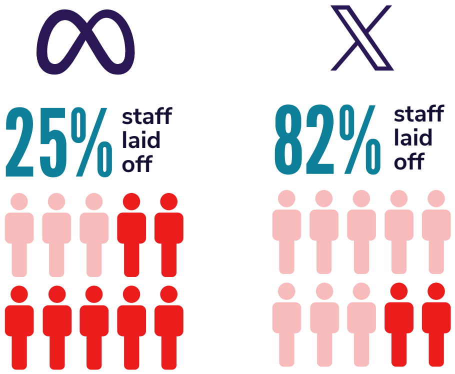
Since November 2022, Alphabet, Meta and Twitter have collectively laid off at least 40,750 employees and contractors, prompting concern that these companies no longer have sufficient staff in place to effectively maintain platform health and safety.

In the first month of Musk's ownership of Twitter, he gutted staff across some of the most critical teams that ensure a healthy and functional platform. Musk removed the board of directors and the Trust and Safety Council.

He fired trust and safety staff, ethical engineering teams and content-moderation contractors. With fewer people on board to maintain Twitter's integrity, many of the platform's core capabilities buckled.[31] Hate speech and disinformation spiked in the weeks and months that followed.[32]

In March 2023, Alphabet started letting go significant numbers of employees on the ethics and safety teams at YouTube.[33] Meta announced its first round of job cuts the same month, with subsequent layoffs in April that had an "outsized effect on the company's trust and safety work."[34]

Twitter has laid off approximately 7,000 people, or 82 percent of its staff. Alphabet, YouTube's parent company, has laid off approximately 12,600 people. With almost no transparency about the implications of these Alphabet job losses for the YouTube platform, it's unclear where the specific cuts took place, although the layoffs included some YouTube consultants. Meta has laid off approximately 21,000 people, or 25 percent of its workforce.[35]



Where platforms choose to trim is a clear indication of company values. Meta has admitted that safety remains a "cost center...not a growth center."[36] Fewer people often means less-effective moderation, leaving violative content on platforms longer. Platform-integrity failures during the first weeks of the Israel-Hamas war point to the critical role that staff in trust and safety and content moderation play during real-time crises.[37]

IS TIKTOK CHECKING ALL THE BOXES?

Noticeably absent from our backsliders list is TikTok. TikTok is the only platform we analyzed that has not rolled back any major content-moderation policies since October 2022. And while it's had some staff shifts, TikTok has not undertaken the same kinds of mass layoffs as its competitors.[38]

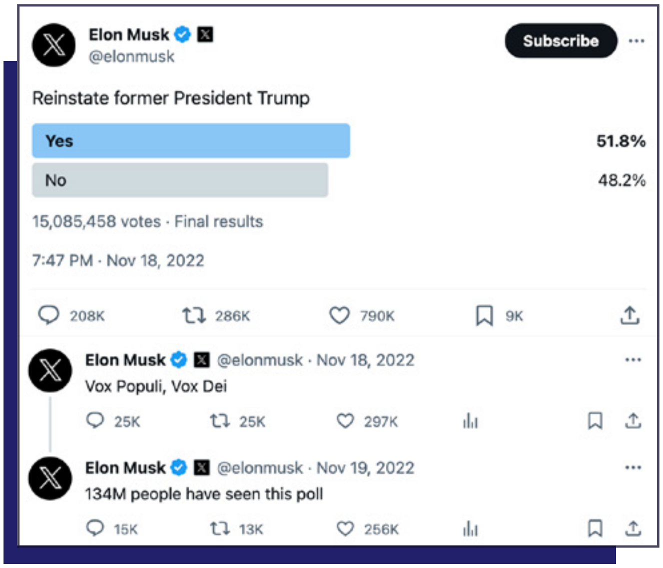
During the last year, as competing platforms have rolled back content-moderation policies and protections for users, TikTok has expanded certain policies. For example, in March 2023, TikTok actually clarified and expanded existing moderation policies, and added new policies requiring disclosures around certain AI-generated content. The platform also introduced a new climate-misinformation policy.[39] Notably, TikTok made these changes just before CEO Shou Zi Chew appeared before Congress to testify about the company's business and moderation practices.[40]

Shou Zi Chew testified before Congress on March 23, 2023. Original photo by Tom Williams via Wikimedia Commons

REINSTATEMENT & MONETIZATION OF DANGEROUS AND EXTREMIST ACCOUNTS

When the powerful use their digital platforms to promote hate, bigotry, lies and other vitriol, there's often an outsized impact offline.[41] For instance, researchers have established the link between Donald Trump's online speech and offline violence in various contexts.[42] One study found direct ties between Trump's anti-Muslim tweets and a rise in anti-Muslim sentiment and hate crimes.[43] Furthermore, the House Select Committee investigated the role that social media played in the insurrection and determined that Trump's posts on social media "set in motion a chain of events that led directly to the attack on the U.S. Capitol."[44] Therefore, it is critical for platforms to maintain rules that apply equitably to all users and to enforce those policies irrespective of the influence certain users carry.

Despite this evidence, every major platform has reinstated Donald Trump over the last year. On Nov. 18, 2022, Elon Musk used a Twitter poll to ask his followers whether to reinstate Trump — a repeat offender of platform policies.[45] In other words, Musk outsourced a major policy decision to users and bots, pretending that he was allowing Twitter's users to choose in the name of free speech. The poll received a slim margin in favor of reinstatement — likely boosted by fake accounts and bots[46] supporting Trump's return. Twitter's November reinstatement of Trump prompted a domino effect across the platforms — on Jan. 25, Meta announced that it would reinstate Trump's accounts.[47] On March 17, YouTube also reinstated Trump's account.[48] Musk's reinstatement provided the necessary cover for other platforms to do the same, even though Trump continues to spread lies about the 2020 election result and other election issues.[49]



Elon Musk @elonmusk

Reinstate former President Trump

Yes	51.8%
No	48.2%

15,085,458 votes · Final results

7:47 PM · Nov 18, 2022

208K 286K 790K 9K

Elon Musk @elonmusk · Nov 18, 2022
Vox Populi, Vox Dei

25K 25K 297K

Elon Musk @elonmusk · Nov 19, 2022
134M people have seen this poll

15K 13K 256K

Conservative judge J. Michael Luttig testified before Congress that Trump and his base remain "a clear and present danger to American democracy" because of their potential to threaten the integrity of the 2024 election.[50]

Musk's sweeping amnesty reinstated thousands of previously banned accounts, including those belonging to white supremacists, conspiracy theorists and others who use the platform to sow division, spread lies, instigate violence and undermine democracy. A sample of reinstated accounts includes:

- Andrew Anglin, an American neo-Nazi who founded the Daily Stormer, a website taking its name from the Nazi propaganda sheet known as Der Stürmer.[51]
- Laura Loomer, a far-right political figure and self-proclaimed "proud Islamophobe" who unsuccessfully ran for a Florida congressional seat.[52]
- Andrew Tate, the influencer and former kickboxer known for posting extreme misogynistic videos. He has said that rape victims "bear some responsibility" for being raped and that he would threaten with a machete women who accuse him of cheating.[53]

- Anthime Gionet, known as Baked Alaska, a white-supremacist internet personality who attended the 2017 "Unite the Right" rally in Charlottesville. He has been sentenced for participating in the Jan. 6 insurrection and live-streaming the attack to his social-media followers.[54]

- Emerald Robinson, the former Newsmax reporter who has claimed that the COVID vaccine contains a satanic marker.[55]

- Gateway Pundit, an online media outlet notorious for promoting conspiracy theories related to vote tampering, climate change and COVID-19.

Meta also has a troubling history of cherry-picking VIP user accounts that amplify lies. For example, in August, Meta rejected its oversight board's recommendation that it suspend the former Cambodian prime minister for video content that "included violent threats" toward his political opponents. *The Hill* reported that the former prime minister "had preemptively removed his Facebook page after the Oversight Board recommendation in June, and banished Facebook representatives from operating in the country."[56] There are real free-expression concerns when it comes to how platforms limit and suspend accounts. Users may have an interest in the content of a particular suspended user. A political figure like Donald Trump may carry unique political and cultural interest for a large swath of the voting public. Platforms must balance these legitimate considerations against policies designed to minimize the spread of hateful speech, harassment, extremism, incitement to violence and lies.[57]

As the platforms weigh reinstatements, the return of previously suspended accounts brings user eyeballs and money for the companies. But it also erodes user experience of these products, as the return of Trump and others ushers in more toxicity and lies.

Trump's social-media posts helped incite people to take part in the Jan. 6 insurrection.
Original photo by Brett Davis via Flickr / CC BY-NC 2.0 / Edited from original

A LEAGUE OF ITS OWN: TWITTER'S TOTAL FAILURE

It's a colossal understatement to say that Elon Musk has failed as the head of Twitter. Musk has destroyed almost everything that once made the platform worthwhile.

As soon as Musk took over Twitter, Free Press and many others expressed concern that his reckless decisions for the platform would harm people in the real world. There were warning signs from the outset: Use of the N-word surged immediately after Musk's purchase last October, allowing bad actors to test the limits of the platform's moderation systems.[58]

Musk began by gutting content-moderation policies and decisions, ranging from ending the COVID-19 disinformation policy to discontinuing the Trust and Safety Council.[59] Musk also laid off key staff on the public policy and the Machine Learning Ethics, Transparency, and Accountability teams, among others.[60] Just four weeks after Twitter's purchase, Musk announced a "general amnesty," swiftly reinstating thousands of previously banned accounts, including those belonging to prominent hate superspreaders.[61]

> Over the subsequent months, Musk rolled back policies, staffing, design and other functions core to the platform's ability to maintain healthy user feeds. Many of these rollbacks reversed years of progress Twitter had made in consultation with civil- and human-rights groups, including the Change the Terms coalition and organizations like Free Press.

HATE SPEECH INCREASED AFTER MUSK BOUGHT TWITTER

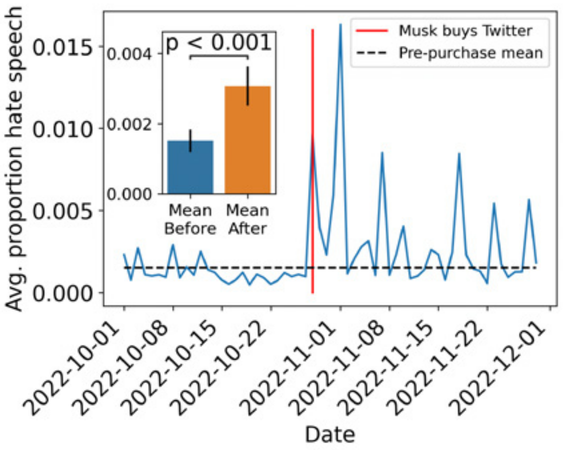


Image by USC Viterbi

Musk discontinued the deadnaming policy that it launched in 2018, which had previously banned users from calling a transgender person by an incorrect name, such as a no-longer-used birth name. He rolled back the platform's misinformation policy and brought back political ads after former Twitter CEO Jack Dorsey sunsetted them for being a vehicle for political lies to enter user feeds unmoderated.

Musk's rollbacks have led to a rise in toxicity across the platform. Six months into his ownership, researchers found that Twitter's content-moderation rollbacks significantly increased hate speech across the platform compared to a similar timeframe prior to Musk's purchase.[62] Later results from a 2023 Trustlab study carried out for the European Commission reveal that Twitter carried the greatest amount of deceitful content when compared to other large social networks, including Facebook, Instagram, LinkedIn, TikTok and YouTube.[63] The study examined more than 6,000 unique social-media posts across platforms and found Twitter carried the most unmoderated disinformation.[64]

In crisis moments, the platform's failures have been catastrophic. During the fall of 2023, amid the ongoing Israel-Hamas war, the platform's algorithms have boosted violent and disturbing images — some real, some faked — and disinformation about the conflict has spread across Twitter and migrated to mainstream news outlets. With few left at the company to vet questionable and violent content, posts are often left unchecked to spread like wildfire.

Furthermore, Musk's decision to give special prominence to content by blue checkmark accounts without adequately verifying users' identities — undoing years of effort to build trust on the platform — has given a soapbox to all sorts of grifters, conspiracy theorists and propagandists seeking to drown public discourse in lies coming from both sides of the conflict and many points in between.[65]

Musk pushed the burden of fact checking onto Twitter users by encouraging them to use "Community Notes" to moderate the platform. But the sheer volume of fake reports following October's Hamas attack in Israel stretched well beyond the reach of any user-powered fact checking. And it got so bad that former Twitter insiders who watchdog the feature told *WIRED* that Community Notes itself became a vehicle for spreading lies about the conflict. "A reliance on Community Notes is not good," one of them said.

"It's not a replacement for proper content moderation."[66] In crisis and wartime, the core functions of a platform matter more than ever. A platform must have well-trained moderators to enforce robust written policies, paired with strong user features to prevent network manipulation. Without any of these in place, it's nearly impossible to distinguish fact from fiction on today's Twitter.[67]

Musk's earliest content-moderation rollbacks sparked the #StopToxicTwitter pressure campaign, which has helped push more than half the platform's advertisers to abandon the platform, with some leaving quietly and others citing concerns about their ads running alongside toxic content.[68] The cost of failure is not just financial. Communities, families, journalists and leaders have all suffered as a result of the platform's failures in recent months.

THREADS LAUNCH SPREADS META TOO THIN

Meta introduced Threads as "a new app, built by the Instagram team" — a team already responsible for managing a product with 2-billion monthly active users.[69] By directly linking this new platform to Instagram (users can't delete their Threads accounts unless they're willing to sacrifice their Instagram accounts, too), it's clear that Meta hopes to capitalize on Instagram's substantial global user base — a plan that, in the app's early weeks, seemed successful.[70]

But Threads launched just months after Meta announced a round of mass layoffs,[71] as well as a hiring freeze, raising concerns about the company's capacity to implement policies regarding the governance, moderation and security of a wholly new platform. According to Reuters,[72] these layoffs directly impacted Meta's privacy and integrity teams.

Free Press and two dozen civil-rights groups wrote to Meta[73] requesting details on how Threads would be moderated, how its content-moderation and privacy policies would be distinct from those of other Meta products, and how the platform planned to enforce policies with transparency. Meta's boilerplate response — received over a month later than requested — failed to concretely answer any of our questions. In some instances, Meta simply pointed back to its own blog posts and other written documents, which lacked meaningful details.

Months after launch, Meta added a search function on Threads but has blocked various kinds of content. For example, it has blocked users from searching for content about COVID-19 and vaccines — which it deemed "potentially sensitive content."[74]

A #StopToxicTwitter banner flew over Miami during a convention Musk spoke at on April 18, 2023.
Original photo by Amadou Cisse via Flickr / Free Press

AI TO THE RESCUE?

Without staff on hand to fulfill core platform functions, executives may give outsized duties to artificial intelligence tools to manage mechanisms such as content moderation and review of flagged content. But even with the assistance of AI, tech companies are unable to enforce content-moderation policies at scale. Mass layoffs only exacerbated this problem.

Furthermore, AI and other automated tools these companies use simply don't have adequate cultural nuance to do it all. Proper and extensive auditing of automated tools requires humans to review the results. These companies cannot expect automated tools — absent the adequate staffing to train and review AI processes — to effectively maintain platform integrity.

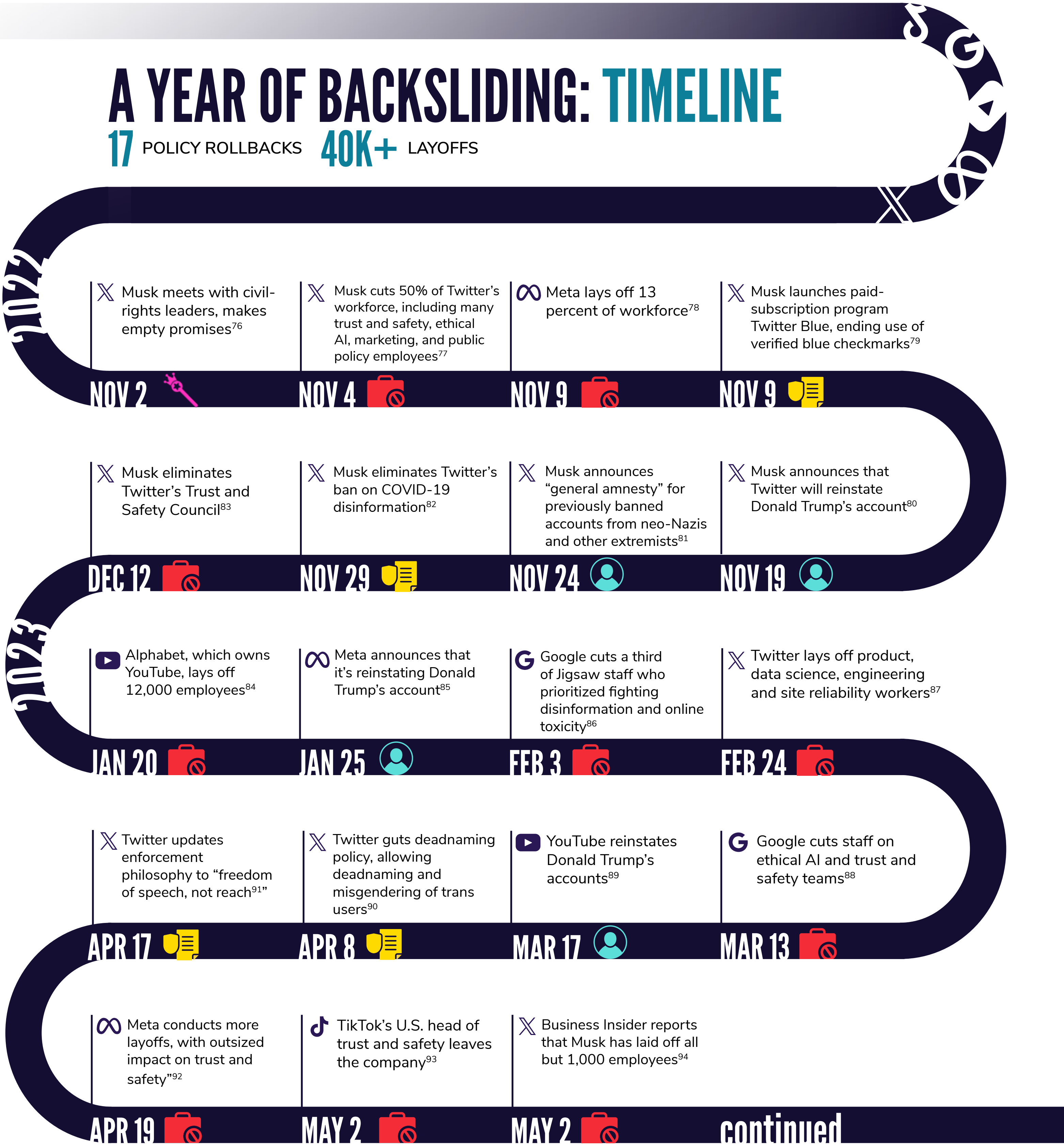
The potential for platform misuse of AI technology should raise several flags for lawmakers and regulators. Platforms that use any digital automation often train their algorithms using untold amounts of user data. This data can include users' names, addresses, purchasing histories, financial information and other sensitive information such as Social Security numbers, medical records — and even people's biometric data, like fingerprints and iris recognition.

Companies like Meta that deal in sophisticated algorithms and AI often say they're gathering this information to deliver hyper-personalized and "improved" experiences for people. But dangerous consequences flow from having advanced algorithms analyze our data without the proper guardrails and auditing of these tools. These companies can also use this data processing to exclude specific users from receiving critical election information, and they can target users to receive disinformation about voting locations and candidates. Simply put, unchecked AI violates our digital civil rights.

Tech companies and lawmakers have different — though equally necessary — roles to properly rein in abusive and discriminatory AI tools.[75] Private companies, particularly social-media companies, must audit and review the AI tools they employ, with adequate human review of the impact of automated processes. Private companies should gather the minimum data about users, with stringent data-conservative approaches to the use and collection of that data. Lawmakers and regulators should also mandate transparency from private companies, create data-minimization requirements and eliminate algorithmic discrimination.

A YEAR OF BACKSLIDING: TIMELINE

17 POLICY ROLLBACKS 40K+ LAYOFFS



NOV 2 ✨ — X Musk meets with civil-rights leaders, makes empty promises[76]

NOV 4 💼 — X Musk cuts 50% of Twitter's workforce, including many trust and safety, ethical AI, marketing, and public policy employees[77]

NOV 9 💼 — ∞ Meta lays off 13 percent of workforce[78]

NOV 9 📄 — X Musk launches paid-subscription program Twitter Blue, ending use of verified blue checkmarks[79]

NOV 19 👤 — X Musk announces that Twitter will reinstate Donald Trump's account[80]

NOV 24 👤 — X Musk announces "general amnesty" for previously banned accounts from neo-Nazis and other extremists[81]

NOV 29 📄 — X Musk eliminates Twitter's ban on COVID-19 disinformation[82]

DEC 12 💼 — X Musk eliminates Twitter's Trust and Safety Council[83]

JAN 20 💼 — ▶ Alphabet, which owns YouTube, lays off 12,000 employees[84]

JAN 25 👤 — ∞ Meta announces that it's reinstating Donald Trump's account[85]

FEB 3 💼 — G Google cuts a third of Jigsaw staff who prioritized fighting disinformation and online toxicity[86]

FEB 24 💼 — X Twitter lays off product, data science, engineering and site reliability workers[87]

MAR 13 💼 — G Google cuts staff on ethical AI and trust and safety teams[88]

MAR 17 👤 — ▶ YouTube reinstates Donald Trump's accounts[89]

APR 8 📄 — X Twitter guts deadnaming policy, allowing deadnaming and misgendering of trans users[90]

APR 17 📄 — X Twitter updates enforcement philosophy to "freedom of speech, not reach[91]"

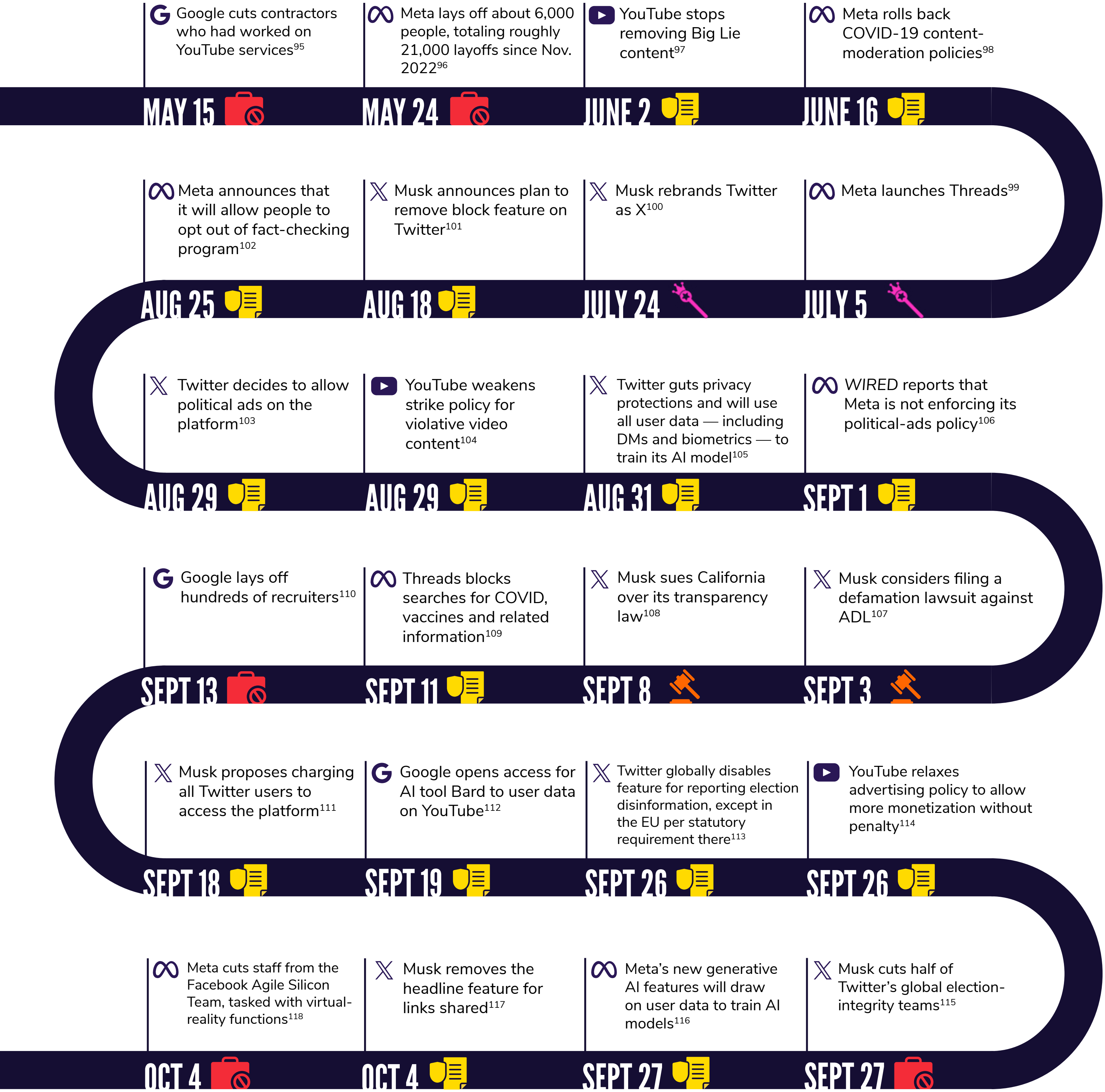
APR 19 💼 — ∞ Meta conducts more layoffs, with outsized impact on trust and safety"[92]

MAY 2 💼 — ♪ TikTok's U.S. head of trust and safety leaves the company[93]

MAY 2 💼 — X Business Insider reports that Musk has laid off all but 1,000 employees[94]

continued

✨ PR STUNT 💼 LAYOFF 📄 POLICY ROLLBACK 👤 REINSTATEMENT ⚖ LITIGATION



MAY 15 💼 — Google cuts contractors who had worked on YouTube services[95]

MAY 24 💼 — Meta lays off about 6,000 people, totaling roughly 21,000 layoffs since Nov. 2022[96]

JUNE 2 📜 — YouTube stops removing Big Lie content[97]

JUNE 16 📜 — Meta rolls back COVID-19 content-moderation policies[98]

JULY 5 🪄 — Meta launches Threads[99]

JULY 24 🪄 — Musk rebrands Twitter as X[100]

AUG 18 📜 — Musk announces plan to remove block feature on Twitter[101]

AUG 25 📜 — Meta announces that it will allow people to opt out of fact-checking program[102]

AUG 29 📜 — Twitter decides to allow political ads on the platform[103]

AUG 29 📜 — YouTube weakens strike policy for violative video content[104]

AUG 31 📜 — Twitter guts privacy protections and will use all user data — including DMs and biometrics — to train its AI model[105]

SEPT 1 📜 — *WIRED* reports that Meta is not enforcing its political-ads policy[106]

SEPT 3 ⚖ — Musk considers filing a defamation lawsuit against ADL[107]

SEPT 8 ⚖ — Musk sues California over its transparency law[108]

SEPT 11 📜 — Threads blocks searches for COVID, vaccines and related information[109]

SEPT 13 💼 — Google lays off hundreds of recruiters[110]

SEPT 18 📜 — Musk proposes charging all Twitter users to access the platform[111]

SEPT 19 📜 — Google opens access for AI tool Bard to user data on YouTube[112]

SEPT 26 📜 — Twitter globally disables feature for reporting election disinformation, except in the EU per statutory requirement there[113]

SEPT 26 📜 — YouTube relaxes advertising policy to allow more monetization without penalty[114]

SEPT 27 💼 — Musk cuts half of Twitter's global election-integrity teams[115]

SEPT 27 📜 — Meta's new generative AI features will draw on user data to train AI models[116]

OCT 4 📜 — Musk removes the headline feature for links shared[117]

OCT 4 💼 — Meta cuts staff from the Facebook Agile Silicon Team, tasked with virtual-reality functions[118]

🪄 PR STUNT 💼 LAYOFF 📜 POLICY ROLLBACK 👤 REINSTATEMENT ⚖ LITIGATION

THE FORECAST AHEAD

The rollbacks described above — symptoms of a broader backslide at the largest social-media companies — are but one piece of a growing tech-accountability problem.

Two years ago, Frances Haugen testified before Congress, where she offered bombshell evidence on when and how much Meta executives knew about the extent to which their platforms were undermining democracy and public safety. Facebook, now Meta, chose profits over user safety over and over, making moderation and design decisions that kept people engaged with toxic content — and boosted the company's bottom line.[119]

These findings prompted a number of congressional hearings, legislative proposals, academic research reports and civil-society recommendations — all undergirding years of advocacy from those seeking platform accountability, transparency and equity. Absent much-needed regulatory oversight, these companies are doing less and less to maintain platform integrity.

Free Press has also documented a different but equally grave set of threats to tech-accountability work at organizations like ours.

IMAGE: Frances Haugen testified before the U.S. Senate on Oct. 5, 2021. Original photo by U.S. Senate Committee on Commerce via Wikimedia Commons

STRATEGIC LAWSUITS AGAINST PLATFORM TRANSPARENCY

Most troubling is a set of lawsuits Musk has initiated to silence researchers and critics.

So far, he has brought two lawsuits, one challenging the Center for Countering Digital Hate (CCDH) and another challenging a new transparency statute in California.[120] He has threatened to bring a third suit against the Anti-Defamation League (ADL).

In July, Musk sued CCDH, alleging the nonprofit committed computer fraud in its use of a tool that monitors advertisements on Twitter.[121] CCDH has refuted Musk's claims and refused to buckle under the pressure of the lawsuit, and allies have sounded the alarm that this suit is a dangerous attempt to chill the organization's independent research.

Musk later threatened to sue the ADL for supporting and encouraging companies to pull their ad spending from Twitter. Musk withdrew his threat once the ADL denied its involvement in such efforts and affirmed that it would continue advertising on the platform. Musk's success in getting the ADL to cave to his demands will likely lead to more threats to researchers and the tech-accountability field more broadly.

Musk has also sued California over a new law mandating semiannual reports from major social-media platforms that describe their content-moderation practices and share data on the numbers of objectionable posts and how companies addressed them.[122] The law also requires companies to make public their terms of service. Failure to comply risks civil fines of up to $15,000 a day for each violation. Musk and Twitter claim that the law violates the platform's First Amendment rights. A federal court in Sacramento is currently assessing this argument.[123]

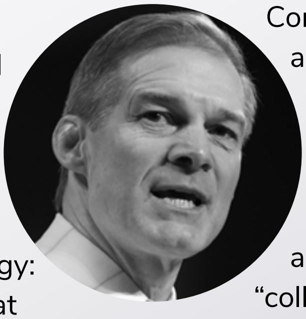
The use of litigation to silence critics is an old tactic.[124] Trump regularly uses this strategy: Media-law specialist Susan Seager notes that "Trump and his companies have been involved in a mind-boggling 4,000 lawsuits over the last 30 years and sent countless threatening cease-and-desist letters to journalists and critics."[125] Musk's use of this Trumpian tactic is hardly surprising — the two men share numerous hallmarks of authoritarian bullies, including claiming to champion free speech while silencing speech they dislike.

Strategic lawsuits attacking platform transparency are similar to "SLAPP" suits (strategic lawsuits against public participation), which chill public participation and journalistic inquiry. The use of these lawsuits is dangerous for researchers who might otherwise want to investigate platform behavior but decide not to out of fear of being sued. It's also dangerous for the public, which might remain in the dark about tech companies' unethical practices. We need urgent data and information-sharing from platforms and thoughtful collaboration across sectors to minimize threats to national security, democratic institutions and local communities that originate on platforms.

CROSS-SECTOR ATTACKS ON TECH ACCOUNTABILITY

To make matters worse, Republicans in Congress have issued similar threats — as well as subpoenas — to stifle research about the spread of election disinformation.[126]

This past summer, Republican House Judiciary Committee Chairman Jim Jordan led an effort demanding documents from — and meetings with — leading U.S. academic researchers who examine disinformation. Jordan and his allies have accused these researchers of "colluding with government officials to suppress conservative speech."[127] These attacks have led researchers to retreat from more public advocacy about the need for tech accountability, with some citing the attacks as the impetus to step back.[128]



As we approach the 2024 election season, we need more research and transparency around what the platforms are doing. Cordoning off communication between the tech industry and other sectors — including government, civil society and researchers — will likely give the platforms more room to back away from their previous promises and policies. We should find ways to better coordinate with social-media companies to increase the integrity of information on their platforms. In our *Recommendations* section, we lay out the essential steps tech companies must take in the coming months.

IMAGE: Republican House Judiciary Committee Chairman Jim Jordan
Original image by Gage Skidmore via Flickr / CC BY-SA 2.0 / Edited from original

RECOMMENDATIONS FOR 2024

Social-media companies' retreat from accountability is happening against the backdrop of the 2024 elections. Dozens of national elections will occur across the world — and many people will get information about candidates, issues, voting logistics and election results on social-media platforms.

WHAT PLATFORMS MUST DO

As with every election cycle, threats abound. Civil-society groups have sounded the alarm about what kinds of threats we might see in 2024:[129]

- ⚠ Supercharged disinformation and fake news stories spun up faster than ever before due to efficient artificial intelligence tools that can personalize disinformation, amplify calls for violence and dissuade civic engagement
- ⚠ AI-generated imagery, audio and videos, known as deepfakes, which mislead voters about candidate comments, positions and other attitudinal markers to sway their voting preferences
- ⚠ Laser-targeted political advertising zeroing in on protected classes in discriminatory and exploitative ways
- ⚠ Tailor-made content crafted to discourage certain categories of voters from participating, based on data platforms have already collected about them
- ⚠ Incorrect labeling of imagery and/or AI-generated false imagery about specific polling locations

Social-media companies tend to treat the threats posed around election cycles as anecdotal and time-limited. Free Press and the broader civil- and human-rights field have urged platforms to consider the year-round vulnerabilities that allow election-specific content to prey on voters. Platforms' governance and enforcement decisions cannot adequately guard against manipulations if spun up a month or two before an election and taken down immediately after the polls close.

We urge social-media platforms to take the following steps to protect users on their platforms:

1. **Reinvest in staffing and teams needed to safeguard election integrity, trust and safety, and moderation.** The major platforms have laid off more than 40,750 employees in the last year alone. At least several thousand of those were positions critical to the moderation and enforcement of platforms' policies — policies that are essential to keeping lies and extremist rhetoric in check. Without the staff needed to adequately confirm content-moderation decisions triggered by automated review, dangerous online discourse will likely seep into more users' feeds — even when deemed violative — for longer periods. This will happen because these platforms simply don't have the staff to efficiently and effectively moderate content.

2. **Reinstate disinformation policies — including those governing election and COVID disinformation — and bolster policy moderation to limit exposure to violence and lies.** Platforms typically pull together rapid-response moderation and other product features once a crisis unfolds, which is too late. These companies typically roll out

election-integrity efforts mere months before an election. This is also too late. And this negligence has real-world consequences: The Global Project Against Hate and Extremism, a nonprofit research organization, found that only 41 percent of voters feel safe or very safe at their polling place, and only a quarter of young people, 28 percent of Black people and 37 percent of Latinx people feel safe or very safe at their polling places.[130] Researchers at UT Austin have noted that racialized disinformation and intimidation tactics target Black, Latinx and Indigenous voters ahead of election cycles.[131] As these kinds of threats, dangerous rhetoric and lies reach people online, they can have an impact on how people behave — and whether they vote — on Election Day. If left unchecked, misleading and violent content may discourage voters from participating in the democratic process.

3. **Launch 2024 election-specific platform interventions in time for the U.S. primaries and keep them in place through at least February 2025.** This includes user-information portals for voters to get real-time information about voter registration, polling locations, ways to access credible civic-engagement tools and more. In previous election cycles, platforms have rolled out election-specific policy updates and announcements the summer before a fall election. This is simply too late. These announcements — and the companion interventions needed to safeguard user safety and civic engagement — must launch by February 2024. These interventions must stay in place long after Election Day to protect the election results and safeguard against attempts to undermine them.

4. **Hold VIP accounts to the same standards as those of other users.** Companies should hold candidates, celebrities and other public figures to account when they break company rules. Meta's own Facebook Oversight Board has documented unequal treatment of users, with some users subject to more lenient review policies while "layperson" accounts are more stringently moderated according to Meta's cross-check program.[132] Free Press has previously documented this special-treatment phenomenon.[133] It persists today, with platforms giving user accounts like Donald Trump's great leniency to promote false claims that the 2020 election was stolen.[134] Companies need to hold these users to the same — or even stricter — moderation and enforcement review standards as their layperson counterparts.

5. **Develop more efficient review and enforcement on political-ad content across languages.** No major social-media platform has a streamlined database that allows one to identify and analyze political ads' visibility, veracity, spending and more.[135] More efficient human review of political ads across languages must occur prior to AI analysis and review. Companies must also ensure human enforcement of their mis- and disinformation and extremism policies on those ads, with timely and transparent data shared to external sectors about trends.

6. **Develop better transparency and disclosure policies and regularly share core metrics data with researchers, journalists, lawmakers and the public.** These companies should take action, as promised in their terms of service, on violative content and on tracking core metrics to distribute externally.[136] They should provide affordable and comprehensive API access to researchers and others. They should share audit reports of moderation and enforcement trends, as well as reporting on the impact of their automated tools.

WHAT GOVERNMENTS MUST DO

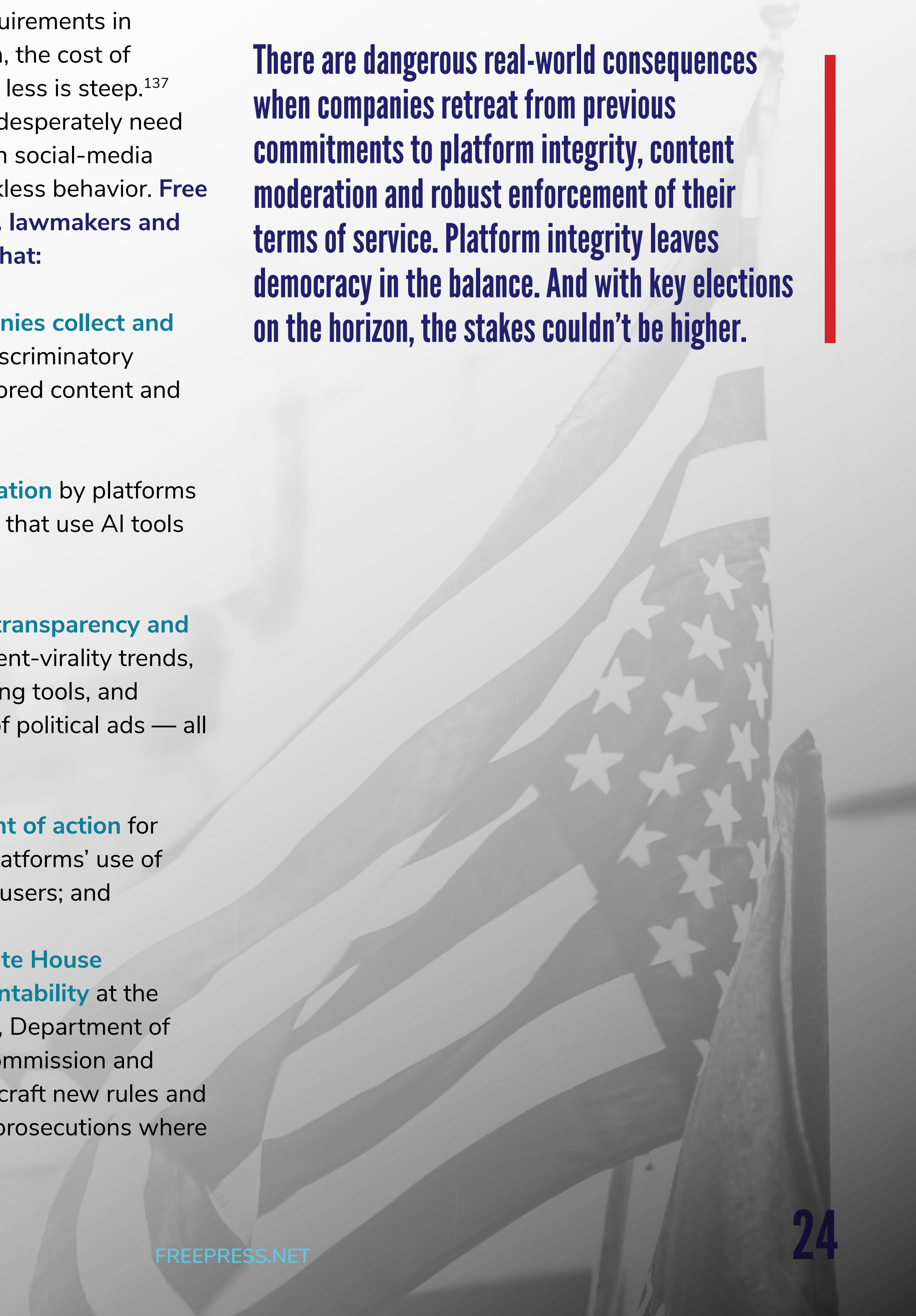
The platforms alone bear responsibility for content moderation — and Free Press Action opposes government efforts to dictate those content standards. But there are steps governments should take to prevent fraud and scams — and to protect democracy, public safety, and human and civil rights online.

Thanks to strict regulatory requirements in places like the European Union, the cost of social-media companies doing less is steep.[137] Here in the United States, we desperately need meaningful regulation to rein in social-media platforms' destructive and reckless behavior. **Free Press Action is calling on U.S. lawmakers and regulators to codify reforms that:**

- **Minimize data that companies collect and retain** to protect against discriminatory targeting of users with tailored content and advertising;
- **Ban algorithmic discrimination** by platforms and other internet services that use AI tools to target users;
- **Require regular platform transparency and disclosure** reports on content-virality trends, results of AI decision-making tools, and visibility and take-downs of political ads — all across languages;
- **Develop a private civil right of action** for violations that flow from platforms' use of personal sensitive data on users; and
- **Leverage agency and White House authority to pursue accountability** at the Federal Trade Commission, Department of Justice, Federal Election Commission and other relevant agencies to craft new rules and launch investigations and prosecutions where statutory violations arise.

As Musk has sunk Twitter down a black hole, his bottom line has suffered to the point where the platform is not worth even a fraction of the $44 billion he paid to purchase it.[138] The cost of doing less is in the billions. But if the last several years have taught us anything, it's that content moderation isn't just about social-media companies' bottom lines.

> **There are dangerous real-world consequences when companies retreat from previous commitments to platform integrity, content moderation and robust enforcement of their terms of service. Platform integrity leaves democracy in the balance. And with key elections on the horizon, the stakes couldn't be higher.**



METHODOLOGY & ACKNOWLEDGEMENTS

To develop this research report, Free Press analyzed announcements from Alphabet, Meta, TikTok and Twitter as well as external media coverage of changes to corporate policies, layoffs, reinstatements and other measures the companies took between Nov. 1, 2022 and Nov. 1, 2023. The platforms' lack of transparency necessitated this external research. For example, we tried to avoid conflicting reports without details about layoffs or dates. We also omitted mention of policy changes or other efforts that companies announced and then quickly reversed, because doing so would have skewed the analysis to focus disproportionately on Twitter. Moreover, including such policies would not provide an accurate landscape of a given platform's policies current to the time of publication.

The report was written by Nora Benavidez, with research support and editing from Rose Lang-Maso. The development and production of this report could not have happened without support from Free Press colleagues Craig Aaron, Jessica J. González, Timothy Karr, Jenna Ruddock and Matt Wood. This report was edited by Amy Kroin and designed by Courtney Morrison, Sara Pritt and Dutch Cosmian.

ENDNOTES

[1] Free Press, *Empy Promises: Inside Big Tech's Weak Effort to Fight Hate and Lies in* 2022, Oct. 27, 2022: https://www.freepress.net/sites/default/files/2022-10/empty_promises_inside_big_techs_weak_effort_to_fight_hate_and_lies_in_2022_free_press_final.pdf, p. 8

[2] Although Elon Musk rebranded Twitter as X in July 2023, it is still commonly referred to by its original name, including throughout this report.

[3] Vittoria Elliott, "Meta Isn't Enforcing Its Own Political Ads Policy, While the 2024 US Election Looms," *WIRED*, Sept. 1, 2023, https://www.wired.com/story/meta-prageru-advertising/

[4] Meta, "Privacy Matters: Meta's Generative AI Features," Sept. 27, 2023, https://about.fb.com/news/2023/09/privacy-matters-metas-generative-ai-features/; Sarah Perez, "X's Privacy Policy Confirms It Will Use Public Data to Train AI Models," TechCrunch, Sept. 1, 2023, https://techcrunch.com/2023/09/01/xs-privacy-policy-confirms-it-will-use-public-data-to-train-ai-models/; Michael Leidtke, "Google Brings Its AI Chatbot Bard Into Its Inner Circle, Opening Door to Gmail, Maps, YouTube," Associated Press, Sept. 19, 2023, https://apnews.com/article/google-artificial-intelligence-bard-gmail-youtube-maps-1229638b82d19afb5226c913821fa1ad

[5] Dean Jackson, Meaghan Conry and Alex Newhouse, "Insiders' View of the January 6th Committee's Social Media Investigation," Just Security, Jan. 5, 2023, https://www.justsecurity.org/84658/insiders-view-of-the-january-6th-committees-social-media-investigation/; Cat Zakrzewski, Cristiano Lima and Drew Harwell, "What the Jan. 6 Probe Found Out About Social Media, but Didn't Report," *The Washington Post*, Jan. 17, 2023, https://www.washingtonpost.com/technology/2023/01/17/jan6-committee-report-social-media/

[6] Olivia Little, "TikTok Is Blocking Searches for WGA Amid Ongoing Writers' Strike," Media Matters for America, Sept. 11, 2023, https://www.mediamatters.org/tiktok/tiktok-blocking-searches-wga-amid-ongoing-writers-strike

[7] Google laid off contractors from Accenture and Cognizant who had previously worked on platform services for YouTube. See Thomas Maxwell and Hugh Langley, "Google Is Downsizing Its Contract Workforce That Supports YouTube Shortly After One Contractor Team's Union Victory," Business Insider, May 15, 2023, https://www.businessinsider.com/google-downsizes-contract-workforce-youtube-union-win-2023-5

[8] Vittoria Elliott, "How X Is Suing Its Way Out of Accountability," WIRED, Aug. 15, 2023, https://www.wired.com/story/twitter-x-ccdh-lawsuit-data-crackdown/

[9] The European Union has examined Twitter's failed content moderation around the Israel-Hamas war and is considering legal action for the platform hosting terrorist-organization content without proper moderation in place to remove it.

[10] "Social Media & the January 6th Attack on the U.S. Capitol: Summary of Investigative Findings", draft report, Tech Policy Press, https://www.techpolicy.press/read-the-january-6-committee-social-media-report/

[11] "The bipartisan statement was signed by the Hoover Presidential Foundation, the Roosevelt Institute, the Truman Library Institute, the John F. Kennedy Library Foundation, the LBJ Foundation, the Richard Nixon Foundation, the Gerald R. Ford Presidential Foundation, the Carter Center, the Ronald Reagan Presidential Foundation and Institute, the George & Barbara Bush Foundation, the Clinton Foundation, the George W. Bush Presidential Center and the Obama Presidential Center." See Gary Fields, "Presidential Centers from Hoover to Bush to Obama Unite to Warn of Fragile State of US Democracy," Associated Press, Sept. 7, 2023, https://apnews.com/article/united-states-democracy-presidents-threats-joint-statement-5530a89df2c41d58a22961f63fb0e6ff

[12] Global Project Against Hate and Extremism, "Violent Antisemitic and Anti-Muslim Hate Escalating Online in Wake of Hamas Attacks on Israel," Oct. 12, 2023, https://globalextremism.org/post/violent-antisemitic-and-anti-muslim-hate-escalating-online-in-wake-of-hamas-attacks-on-israel/

[13] Alice Speri, "'Beheaded Babies' Report Spread Wide and Fast — But Israel Military Won't Confirm it," The Intercept, Oct. 11, 2023, https://theintercept.com/2023/10/11/israel-hamas-disinformation/

[14] Oliver Darcy, “Déjà Coup: How Election Lies Sparked the Violent Attack on Brazil’s Government,” CNN, Jan. 19, 2023, https://www.cnn.com/2023/01/09/media/brazil-government-reliable-sources/index.html

[15] Enda Curran and Alan Crawford, “Brace for Elections: 40 Countries Are Voting in 2024,” Bloomberg, Nov. 1, 2023, https://www.bloomberg.com/news/articles/2023-11-01/brace-for-elections-40-countries-are-voting-in-2024

[16] Marwa Fatawa, “Facebook Is Bad at Moderating in English. In Arabic, It’s a Disaster,” Rest Of World, Nov. 18, 2021, https://restofworld.org/2021/facebook-is-bad-at-moderating-in-english-in-arabic-its-a-disaster; Olivia Solon, “‘Facebook Doesn’t Care’: Activists Say Accounts Removed Despite Zuckerberg’s Free-Speech Stance,” NBC News, June 15, 2020, https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110; Jacqueline Rowe, “Marginalised Languages and the Content Moderation Challenge,” Global Partners Digital, March 2, 2022, https://www.gp-digital.org/marginalised-languages-and-the-content-moderation-challenge

[17] Salvador Rodriguez, “Senators Demand Facebook CEO Mark Zuckerberg Answer Questions After Whistleblower’s Revelations at Hearing,” CNBC, Oct. 5, 2021, https://www.cnbc.com/2021/10/05/congress-demands-mark-zuckerberg-answer-questions-at-haugen-hearing.html

[18] “Luján, Klobuchar, Cárdenas Lead Colleagues Urging Tech CEOs to Combat Spanish-Language Disinformation,” July 30, 2021, https://www.lujan.senate.gov/newsroom/press-releases/lujan-klobuchar-cardenas-lead-colleagues-urging-tech-ceos-to-combat-spanish-language-disinformation/

[19] Naomi Nix, “Inside the Civil Rights Campaign to Get Big Tech to Fight the Big Lie,” *The Washington Post*, Sept. 22, 2022, https://www.washingtonpost.com/technology/2022/09/22/midterms-elections-social-media-civil-rights/; Id, “Big Tech is Failing to Fight the Big Lie, Civil Rights Groups Charge,” *The Washington Post*, Oct. 27, 2022, https://www.washingtonpost.com/technology/2022/10/27/civil-rights-2022-midterms/

[20] Joseph Cox and Jason Koebler, “Facebook Bans White Nationalism and White Separatism,” Motherboard, March 27, 2019, https://www.vice.com/en/article/nexpbx/facebook-bans-white-nationalism-and-white-separatism; “An Update to How We Address Movements and Organizations Tied to Violence,” Meta, Aug. 19, 2020, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/

[21] Adi Robertson, “Twitter Has Banned Misgendering or ‘Deadnaming’ Transgender People,” The Verge, Nov. 27, 2018, https://www.theverge.com/2018/11/27/18113344/twitter-trans-user-hateful-content-misgendering-deadnaming-ban

[22] “Updating Our Rules Against Hateful Conduct,” Twitter, July 9, 2019, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate; Queenie Wong, Corinne Reichert and Oscar Gonzalez, “Twitter’s Decision to Halt Political Ads Puts More Pressure on Facebook,” CNET, Oct. 30, 2019, https://www.cnet.com/tech/mobile/twitters-decision-to-halt-political-ads-puts-more-pressure-on-facebook/

[23] “Updating Our Approach to Misleading Information,” Twitter, May 11, 2020, https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information

[24] Casey Newton, “YouTube Says It Will Recommend Fewer Videos About Conspiracy Theories,” The Verge, Jan. 25, 2019, https://www.theverge.com/2019/1/25/18197301/youtube-algorithm-conspiracy-theories-misinformation; Kevin Roose and Kate Conger, “YouTube to Remove Thousands of Videos Pushing Extreme Views,” *The New York Times*, June 5, 2019, https://www.nytimes.com/2019/06/05/business/youtube-remove-extremist-videos.html

[25] Natalie Mathes, “Social Media Platforms Are Backsliding, Paving the Way for Chaos in 2024,” Media Matters for America, June 13, 2023, https://www.mediamatters.org/facebook/social-media-platforms-are-backsliding-paving-way-chaos-2024-0; GLAAD, “Social Media Safety Index 2023,” https://glaad.org/publications/social-media-safety-index-2023/; Megan Shahi, “Protecting Democracy Online in 2024 and Beyond,” Center for American Progress, Sept. 14, 2023, https://www.americanprogress.org/article/protecting-democracy-online-in-2024-and-beyond/; Center for Countering Digital Hate, *X Content Moderation Failure*, Sept. 13, 2023, https://counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/

[26] Free Press, *Empty Promises: Inside Big Tech’s Weak Effort to Fight Hate and Lies in 2022*, Oct. 27, 2022, https://www.freepress.net/policy-library/empty-promises-inside-big-techs-weak-effort-fight-hate-and-lies-2022

27 Google, YouTube's parent company, provided scant details on its 2022 election period; see Laurie Richardson, "Recapping Our Work on the 2022 U.S. Midterm Elections," Google, https://blog.google/outreach-initiatives/civics/recapping-our-work-on-the-2022-us-midterm-elections/. Meta has no public summary of the 2022 midterm-election period. TikTok also failed to provide any publicly available summaries or reporting on the 2022 midterms, though it had brief community-guidelines updates; see "Community Guidelines Enforcement Report," TikTok, https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-4/. Twitter has no publicly available writing on the 2022 midterms. Elon Musk took over Twitter just a few days before the 2022 midterms and quickly removed critical staff on trust and safety and other teams.

28 Free Press, *Empty Promises: Inside Big Tech's Weak Effort to Fight Hate and Lies in 2022*, Oct. 27, 2022, https://www.freepress.net/policy-library/empty-promises-inside-big-techs-weak-effort-fight-hate-and-lies-2022

29 Change the Terms, https://www.changetheterms.org/

30 Nora Benavidez, "Social Media Platform Integrity Matters in Times of War," Just Security, Oct. 13, 2023, https://www.justsecurity.org/89449/social-media-platform-integrity-matters-in-times-of-war-now-more-than-ever/

31 Jennifer Korn, "As Twitter Failures Go from Bad to Worse, Users Wonder How Long It Can Stay Online," CNN, March 12, 2023, https://www.cnn.com/2023/03/12/tech/twitter-breaking/index.html

32 Rashawn Ray and Joy Anyanwu, "Why Is Elon Musk's Twitter Takeover Increasing Hate Speech?" Brookings Institution, Nov. 23, 2022, https://www.brookings.edu/articles/why-is-elon-musks-twitter-takeover-increasing-hate-speech/; Julia Cohen, "New Twitter, Now With More Hate," USC Viterbi School of Engineering, April 20, 2023, https://viterbischool.usc.edu/news/2023/04/new-twitter-now-with-more-hate/; TrustLab, "Code of Practice on Disinformation," Sept. 22, 2023, https://disinfocode.eu/wp-content/uploads/2023/09/code-of-practice-on-disinformation-september-22-2023.pdf

33 Hayden Field and Jonathan Vanian, "Tech Layoffs Ravage the Teams That Fight Online Misinformation and Hate Speech," CNBC, May 26, 2023, https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html#:~:text=Twitter%20effectively%20disbanded%20its%20ethical,safety%20department%2C%20according%20to%20reports; Thomas Maxwell and Hugh Langley, "Google Is Downsizing Its Contract Workforce That Supports YouTube Shortly After One Contractor Team's Union Victory," Business Insider, May 15, 2023, https://www.businessinsider.com/google-downsizes-contract-workforce-youtube-union-win-2023-5

34 Hayden Field and Jonathan Vanian, "Tech Layoffs Ravage the Teams That Fight Online Misinformation and Hate Speech," CNBC, May 26, 2023, https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html#:~:text=Twitter%20effectively%20disbanded%20its%20ethical,safety%20department%2C%20according%20to%20reports

35 Ryan Mac, Mike Isaac and Kate Conger, "Twitter Outages Are on the Rise Amid Elon Musk's Job Cuts," *The New York Times*, Feb. 8, 2023, https://www.nytimes.com/2023/02/28/technology/twitter-outages-elon-musk.html; Clare Duffy, "Meta's Latest Round of Layoffs Is Underway," CNN, April 19, 2023, https://www.cnn.com/2023/04/19/tech/meta-tech-team-layoffs-begin/index.html

36 Clare Duffy, Aditi Sangal, Melissa Mahtani and Meg Wagner, "Internal Facebook Documents Revealed," CNN, Oct. 26, 2021, https://www.cnn.com/business/live-news/facebook-papers-internal-documents-10-25-21/h_c0b027ed77300bdfec6771e8c0788cff#:~:text=Former%20Facebook%20employee%20and%20whistleblower,very%20short%2Dterm%20in%20thinking

37 Pranav Dixit, "Social Media Platforms Swamped with Fake News on the Israel-Hamas War," Al Jazeera, Oct. 10, 2023, https://www.aljazeera.com/news/2023/10/10/social-media-platforms-swamped-with-fake-news-on-the-israel-hamas-war; Nora Benavidez, "Social Media Platform Integrity Matters in Times of War," Just Security, Oct. 13, 2023, https://www.justsecurity.org/89449/social-media-platform-integrity-matters-in-times-of-war-now-more-than-ever/

38 Alyssa Stringer, "A Comprehensive List of 2023 Tech Layoffs," TechCrunch, Nov. 14, 2023, https://techcrunch.com/2023/11/14/tech-layoffs-2023-list/

39 Aisha Malik, "TikTok Overhauls Its Community Guidelines, Adds New Policies on AI and Climate Misinformation," TechCrunch, March 21, 2023, https://techcrunch.com/2023/03/21/tiktok-overhauls-community-guidelines-adds-new-policies-ai-climate-misinformation/

40 Id.

41 Heidi Beirich and Wendy Via, *Democracies Under Threat*, Global Project Against Hate and Extremism, March 2021, https://globalextremism.org/wp-content/uploads/2022/01/GPAHE_Democracy-under-threat.pdf

42 Id.; Justin Hendrix, "The Final Jan. 6th Report on the Role of Social Media," Tech Policy Press, Dec. 23, 2022, https://techpolicy.press/the-final-january-6th-report-on-the-role-of-social-media/

43 Karsten Müller and Carlo Schwarz, "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment," *American Economic Journal: Applied Economics*, Vol. 15, No. 3, July 2023, https://www.aeaweb.org/articles?id=10.1257/app.20210211

44 Select Committee to Investigate the January 6th Attack on the United States Capitol, *Final Report*, Dec. 22, 2022, https://www.govinfo.gov/content/pkg/GPO-J6-REPORT/pdf/GPO-J6-REPORT.pdf; see pp. 55, 499

45 Elon Musk tweet, Twitter, Nov. 19, 2022, https://twitter.com/elonmusk/status/1594131768298315777

46 Davey Alba and Bloomberg, "The People Have Spoken—or Maybe Not: Elon Musk's Use of Twitter Polls for Key Decisions Invites Manipulation," *Fortune*, Dec. 22, 2022, https://fortune.com/2022/12/22/elon-musk-twitter-poll-manipulation-bots-acquisition/

47 Nick Clegg, "Ending Suspension of Trump's Accounts With New Guardrails to Deter Repeat Offenses," Meta, Jan. 25, 2023, https://about.fb.com/news/2023/01/trump-facebook-instagram-account-suspension/

48 BBC, "YouTube Reinstates Donald Trump's Channel," March 17, 2023, https://www.bbc.com/news/technology-64993603

49 Martin Pengelly, "Trump a 'Clear and Present Danger to US Democracy', Conservative Judge Warns," *The Guardian*, June 16, 2022, https://www.theguardian.com/us-news/2022/jun/16/trump-clear-present-danger-to-us-democracy-conservative-judge-warns

50 Id.

51 Nikki McCann Ramirez, "Elon Brings One of America's Most Prominent Nazis Back to Twitter," *Rolling Stone*, Dec. 2, 2022, https://www.rollingstone.com/politics/politics-news/elon-musk-twitter-reinstates-neo-nazi-andrew-anglin-account-1234640390/

52 Daniel Villareal, "Twitter Won't Lift Ban on GOP House Candidate and 'Proud Islamophobe,'" *Newsweek*, Aug. 19, 2020, https://www.newsweek.com/twitter-wont-lift-ban-gop-house-candidate-proud-islamophobe-1526321

53 Shanti Das, "Inside the Violent, Misogynistic World of TikTok's New Star, Andrew Tate," *The Guardian*, Aug. 6, 2022, https://www.theguardian.com/technology/2022/aug/06/andrew-tate-violent-misogynistic-world-of-tiktok-new-star

54 Martin Pengelly, "Far-Right Influencer Known as 'Baked Alaska' Sentenced Over Capitol Attack," *The Guardian*, Nov. 10, 2021, https://www.theguardian.com/us-news/2023/jan/10/baked-alaska-anthime-gionet-sentenced-capitol-attack

55 Jordan Valinsky, "Newsmax Reporter Permanently Banned from Twitter for Posting COVID Misinformation," CNN, Nov. 10, 2021, https://www.cnn.com/2021/11/10/media/newsmax-twitter-emerald-robinson-banned/index.html

56 Rebecca Klar, "Meta Rejects Recommendation to Suspend Former Cambodian Prime Minister," *The Hill*, Aug. 29, 2023, https://thehill.com/policy/technology/4176843-meta-rejects-recommendation-to-suspend-former-cambodian-prime-minister/

57 In the week after Twitter banned Trump's account, misinformation about election fraud dropped 73 percent. *See* Edward Moyer, "After Twitter Banned Trump, Misinformation Plummeted, Report Says," CNET, Jan. 16, 2021, https://www.cnet.com/news/politics/after-twitter-banned-trump-misinformation-plummeted-says-report/

[58] Oliver Darcy, “Hate Speech Dramatically Surges on Twitter Following Elon Musk Takeover, New Research Shows,” CNN, Dec. 12, 2022, https://www.cnn.com/2022/12/02/tech/twitter-hate-speech/index.html

[59] Taylor Lorenz, “Musk’s Twitter No Longer Bans COVID Misinformation,” *The Washington Post*, Nov. 29, 2022, https://www.washingtonpost.com/technology/2022/11/29/twitter-covid-misinformation-policy/; Cat Zakrzewski, “Twitter Dissolves Trust and Safety Council,” *The Washington Post*, Dec. 12, 2022, https://www.washingtonpost.com/technology/2022/12/12/musk-twitter-harass-yoel-roth/; Elon Musk tweet, Twitter, Nov. 24, 2022, https://twitter.com/elonmusk/status/1595869526469533701

[60] Hayden Field and Jonathan Vanian, “Tech Layoffs Ravage the Teams That Fight Online Misinformation and Hate Speech,” CNBC, May 26, 2023, https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html#:~:text=Twitter%20effectively%20disbanded%20its%20ethical,safety%20department%2C%20according%20to%20reports

[61] Dan Milmo, “Elon Musk Offers General Amnesty to Suspended Twitter Accounts,” *The Guardian*, Nov. 24, 2022, https://www.theguardian.com/technology/2022/nov/24/elon-musk-offers-general-amnesty-to-suspended-twitter-accounts

[62] Julia Cohen, “New Twitter, Now With More Hate,” USC Viterbi School of Engineering, April 20, 2023, https://viterbischool.usc.edu/news/2023/04/new-twitter-now-with-more-hate/

[63] TrustLab, “Code of Practice on Disinformation,” Sept. 22, 2023, https://disinfocode.eu/wp-content/uploads/2023/09/code-of-practice-on-disinformation-september-22-2023.pdf

[64] Id.

[65] Ben Goggin, “Verified Accounts Spread Fake News Release About a Biden $8 Billion Aid Package to Israel,” NBC News, Oct. 8, 2023, https://www.nbcnews.com/tech/internet/verified-accounts-spread-fake-news-release-biden-8-billion-aid-package-rcna119372

[66] Vittoria Elliott and David Gilbert, “Elon Musk’s Main Tool for Fighting Disinformation on X Is Making the Problem Worse, Insiders Claim,” *WIRED*, Oct. 17, 2023, https://www.wired.com/story/x-community-notes-disinformation/

[67] Glenn Chapman, “A Year After Musk’s Twitter Takeover, X Remains Mired in Turmoil,” Techxplore, Oct. 26, 2023, https://techxplore.com/news/2023-10-year-musk-twitter-takeover-mired.html

[68] Clare Duffy, “More Than Half of Twitter’s Top 1,000 Advertisers Stopped Spending on Platform, Data Show,” CNN, Feb. 10, 2023, https://www.cnn.com/2023/02/10/tech/twitter-top-advertiser-decline/index.html

[69] Meta, “Introducing Threads: A New Way to Share With Text,” Meta Newsroom, July 5, 2023, https://about.fb.com/news/2023/07/introducing-threads-new-app-text-sharing; Christina Newberry, “34 Instagram Stats Marketers Need to Know in 2023,” Hootsuite, Jan. 24, 2023, https://blog.hootsuite.com/instagram-statistics (“ … more than 2 billion now use the platform monthly, according to the latest Meta earnings call”).

[70] Jay Peters and Jon Porter, “Instagram’s Threads Surpasses 100 Million Users,” The Verge, July 10, 2023, https://www.theverge.com/2023/7/10/23787453/meta-instagram-threads-100-million-users-milestone

[71] Alex Hern, “Zuckerberg’s Meta to Lay Off Another 10,000 Employees,” *The Guardian*, March 14, 2023, https://www.theguardian.com/technology/2023/mar/14/mark-zuckerberg-meta-layoffs-hiring-freeze

[72] Katie Paul, “Facebook Owner Meta Slashes Business Teams in Final Round of Layoffs,” Reuters, May 24, 2023, https://www.reuters.com/technology/facebook-owner-meta-starts-final-round-layoffs-2023-05-24

[73] Naomi Nix, “Meta Is Done Moderating. On Threads, Users Decide What They See,” *The Washington Post*, July 14, 2023, https://www.washingtonpost.com/technology/2023/07/14/threads-algorithm-content-moderation

[74] Christianna Silva, “Threads Is Finally Adding a Search Function,” Mashable, Aug. 31, 2023, https://mashable.com/article/threads-search-keyword; Taylor Lorenz, “Threads Blocks Searches Related to COVID and Vaccines as Cases Rise,” *The Washington Post*, Sept. 11, 2023, https://www.washingtonpost.com/technology/2023/09/11/threads-covid-coronavirus-searches-blocked/

75 Free Press, “How Tech Companies & Policymakers Can Fight AI Harms & Safeguard Digital Civil Rights,” Sept. 12, 2023, https://www.freepress.net/policy-library/how-tech-companies-policymakers-can-fight-ai-harms-safeguard-digital-civil-rights

76 Free Press, “Musk Meets with Free Press and Civil-Rights Groups to Discuss Twitter Community Standards, Election Integrity and Content Moderation,” Nov. 2, 2022, https://www.freepress.net/news/press-releases/musk-meets-free-press-and-civil-rights-groups-discuss-twitter-community-standards

77 Hayden Field and Jonathan Vanian, “Tech Layoffs Ravage the Teams that Fight Online Misinformation and Hate Speech,” CNBC, May 26, 2023, https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html

78 Sheera Frankel, Adam Satariano and Ryan Mac, “Meta Lays Off More Than 11,000 Employees,” *The New York Times*, Nov. 9, 2022, https://www.nytimes.com/2022/11/09/technology/meta-layoffs-facebook.html

79 Max Zahn, “How Twitter’s $8 Verification Plan Works and Why It’s Facing Criticism,” ABC News, Nov. 9, 2022, https://abcnews.go.com/Business/twitters-verification-plan-works-facing-criticism/story?id=92794983

80 Elon Musk tweet, Twitter, Nov. 19, 2022, https://twitter.com/elonmusk/status/1594131768298315777

81 Elon Musk tweet, Twitter, Nov. 24, 2022, https://twitter.com/elonmusk/status/1595869526469533701

82 Taylor Lorenz, “Twitter Ends Its Ban on COVID Misinformation,” *The Washington Post*, Nov. 29, 2022, https://www.washingtonpost.com/technology/2022/11/29/twitter-covid-misinformation-policy/

83 Cat Zakrzewski, “Twitter Dissolves Trust and Safety Council,” *The Washington Post*, Dec. 12, 2022, https://www.washingtonpost.com/technology/2022/12/12/musk-twitter-harass-yoel-roth/

84 Paul Sawers, “Google Parent Alphabet Cuts 6% of Its Workforce, Impacting 12,000 People,” TechCrunch, Jan. 20, 2023, https://techcrunch.com/2023/01/20/google-parent-alphabet-cuts-6-of-its-workforce-impacting-12000-people/

85 Nick Clegg, “Ending Suspension of Trump’s Accounts With New Guardrails to Deter Repeat Offenses,” Meta, Jan. 25, 2023, https://about.fb.com/news/2023/01/trump-facebook-instagram-account-suspension/

86 Thomas Brewster and Richard Nieva, “Google Cuts Company Protecting People from Surveillance to a ‘Skeleton Crew,’ Say Laid Off Workers,” *Forbes*, Feb. 3, 2023, https://www.forbes.com/sites/thomasbrewster/2023/02/02/jigsaw-google-alphabet-layoffs/?sh=1bcd880c2d71

87 Kate Conger, Ryan Mac and Mike Isaac, “In Latest Round of Job Cuts, Twitter Is Said to Lay Off at Least 200 Employees,” *The New York Times*, Feb. 26, 2023, https://www.nytimes.com/2023/02/26/technology/twitter-layoffs.html

88 Hayden Field and Jonathan Vanian, “Tech Layoffs Ravage the Teams that Fight Online Misinformation and Hate Speech,” CNBC, May 26, 2023, https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html

89 BBC, “YouTube Reinstates Donald Trump’s Channel,” March 17, 2023, https://www.bbc.com/news/technology-64993603

90 Adi Robertson, “Twitter Has Banned Misgendering or ‘Deadnaming’ Transgender People,” The Verge, Nov. 27, 2018, https://www.theverge.com/2018/11/27/18113344/twitter-trans-user-hateful-content-misgendering-deadnaming-ban

91 Twitter, “Freedom of Speech, Not Reach: An Update on Our Enforcement Philosophy,” April 17, 2023, https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy

92 Clare Duffy, “Meta’s Latest Round of Layoffs Is Underway,” CNN, April 19, 2023, https://www.cnn.com/2023/04/19/tech/meta-tech-team-layoffs-begin/index.html

93 Alex Heath, “TikTok’s Head of US Trust and Safety Is Leaving,” The Verge, May 2, 2023, https://www.theverge.com/2023/5/2/23708534/tiktok-head-of-us-trust-and-safety-eric-han-leaves-company

[94] Kali Hays, “Elon Musk Has Chopped Twitter Down to About 1,000 Employees,” Business Insider, May 2, 2023, https://www.businessinsider.com/elon-musk-chops-twitter-down-1000-employees-2023-5

[95] Thomas Maxwell and Hugh Langley, “Google Is Downsizing Its Contract Workforce That Supports YouTube Shortly After One Contractor Team’s Union Victory,” Business Insider, May 15, 2023, https://www.businessinsider.com/google-downsizes-contract-workforce-youtube-union-win-2023-5

[96] Amanda Silberling, “Meta Conducts Yet Another Round of Layoffs,” TechCrunch, May 24, 2023, https://techcrunch.com/2023/05/24/meta-conducts-yet-another-round-of-layoffs/

[97] YouTube, “An Update on Our Approach to US Election Misinformation,” June 2, 2023, https://blog.youtube/inside-youtube/us-election-misinformation-update-2023/

[98] Mary Kekatos, “Health Experts Worry as Meta Rolls Back Some COVID Misinformation Policies,” ABC News, June 19, 2023, https://abcnews.go.com/US/health-experts-worry-meta-rolls-back-covid-misinformation/story?id=100192537

[99] Meta, “Introducing Threads: A New Way to Share With Text,” July 5, 2023, https://about.fb.com/news/2023/07/introducing-threads-new-app-text-sharing/

[100] Jordan Valinsky, “Elon Musk Rebrands Twitter as X,” CNN, July 24, 2023, https://www.cnn.com/2023/07/24/tech/twitter-rebrands-x-elon-musk-hnk-intl/index.html

[101] Elon Musk tweet, Twitter, Aug. 18, 2023, https://twitter.com/elonmusk/status/1692558414105186796

[102] Naomi Nix and Sarah Ellison, “Following Elon Musk’s Lead, Big Tech is Surrendering to Disinformation,” *The Washington Post*, Aug. 25, 2023, https://www.washingtonpost.com/technology/2023/08/25/political-conspiracies-facebook-youtube-elon-musk/

[103] Clare Duffy and Brian Fung, “X Will Allow Political Ads Again and Hire for Safety and Election Teams Ahead of 2024 Elections,” CNN, Aug. 29, 2023, https://www.cnn.com/2023/08/29/tech/x-twitter-2024-election-ads/index.html

[104] Sarah Perez, “YouTube is Giving Creators Violating Policies a Way to Wipe Out Their Warnings,” TechCrunch, Aug. 29, 2023, https://techcrunch.com/2023/08/29/youtube-is-giving-creators-violating-policies-a-way-to-wipe-out-their-warnings/

[105] Tonya Riley, “Twitter, Now X, Will Begin Collecting Users’ Biometric Data,” CyberScoop, Aug. 31, 2023, https://cyberscoop.com/twitter-x-user-biometric-data/

[106] Vittoria Elliott, “Meta Isn’t Enforcing Its Own Political Ads Policy, While the 2024 US Election Looms,” *WIRED*, Sept. 1, 2023, https://www.wired.com/story/meta-prageru-advertising/

[107] Justin Baragona, “Elon Musk Ponders Twitter Poll on Banning ADL From ‘Free Speech’ Site,” The Daily Beast, Sept. 2, 2023, https://www.thedailybeast.com/elon-musk-ponders-twitter-poll-on-banning-adl-from-free-speech-site-after-bantheadl-trends

[108] Guardian staff and agencies, “Elon Musk’s X Sues California Over New Social Media Transparency Laws,” *The Guardian*, Sept. 8, 2023, https://www.theguardian.com/us-news/2023/sep/08/elon-musk-twitter-lawsuit-california-free-speech

[109] Taylor Lorenz, “Threads Blocks Searches for ‘Covid’ and ‘Long-Covid,’” *The Washington Post*, Sept. 11, 2023, https://www.washingtonpost.com/technology/2023/09/11/threads-covid-coronavirus-searches-blocked/

[110] Louise Matsakis, “Google Lays Off Hundreds on Recruiting Team,” Semafor, Sept. 13, 2023 https://www.semafor.com/article/09/13/2023/google-lays-off-hundreds-on-recruiting-team

[111] Sara Fischer, “Musk Says X Will Charge Everyone to Use the Platform,” Axios, Sept. 19, 2023, https://www.axios.com/2023/09/19/musk-x-twitter-charge-all-users-monthly-subscription-fees

112 Michael Liedtke, "Google Brings Its AI Chatbot Bard into Its Inner Circle, Opening Door to Gmail, Maps, YouTube," Associated Press, Sept. 19, 2023, https://apnews.com/article/google-artificial-intelligence-bard-gmail-youtube-maps-1229638b82d19afb5226c913821fa1ad

113 Byron Kaye, "Musk's X Disabled Feature for Reporting Electoral Misinformation — Researcher," Reuters, Sept. 26, 2023, https://www.reuters.com/technology/musks-x-disabled-feature-reporting-electoral-misinformation-researcher-2023-09-27/

114 Sarah Perez, "YouTube Relaxes Advertiser-Friendly Guidelines Around Controversial Topics, Like Abortion, Abuse and Eating Disorders," TechCrunch, Sept. 26, 2023, https://techcrunch.com/2023/09/26/youtube-relaxes-advertiser-friendly-guidelines-around-controversial-topics-like-abortion-abuse-and-eating-disorders/

115 Erin Woo, "Musk's X Cuts Half of Election Integrity Team After Promising to Expand It," The Information, Sept. 27, 2023, https://www.theinformation.com/articles/musks-x-cuts-half-of-election-integrity-team-after-promising-to-expand-it

116 Mike Clark, "Privacy Matters: Meta's Generative AI Features," Meta, Sept. 27, 2023, https://about.fb.com/news/2023/09/privacy-matters-metas-generative-ai-features/

117 Alex Kirshner, "Twitter Was for News," Slate, Oct. 5, 2023, https://slate.com/technology/2023/10/elon-musk-x-twitter-news-links-headlines-why.html

118 Divya Bhati, "Meta to Announce Another Round of Layoffs, Will Fire Employees in Metaverse Silicon Unit," India Today, Oct. 4, 2023, https://www.indiatoday.in/technology/news/story/meta-to-announce-another-round-of-layoff-will-fire-employees-in-metaverse-silicon-unit-2444214-2023-10-04

119 Ryan Mac and Cecilia Kang, "Whistle-Blower Says Facebook 'Chooses Profits Over Safety'," *The New York Times*, Oct. 3, 2021, https://www.nytimes.com/2021/10/03/technology/whistle-blower-facebook-frances-haugen.html

120 Musk filed a third lawsuit targeting another research nonprofit, Media Matters for America, following the research period for this report. He filed this lawsuit on Nov. 20, 2023, claiming the organization manufactured a report showing advertisers' posts beside white-nationalist and neo-Nazi posts to "drive advertisers from the platform and destroy X Corp." See Barbara Otutay, "Musk's X Sues Liberal Advocacy Group Media Matters Over Its Report on Ads Next to Hate Groups' Posts," Associated Press, Nov. 21, 2023, https://apnews.com/article/elon-musk-media-matters-lawsuit-advertising-neonazi-1fe499daa600f513af27ffa68d2e8b91

121 Bryan Pietsch, "Twitter, Now X, Sues Group That Researched Hate Speech on Platform," *The Washington Post*, Aug. 1, 2023, https://www.washingtonpost.com/technology/2023/08/01/twitter-lawsuit-center-for-countering-digital-hate/

122 Guardian staff and agencies, "Elon Musk's X Sues California Over New Social Media Transparency Laws," *The Guardian*, Sept. 8, 2023, https://www.theguardian.com/us-news/2023/sep/08/elon-musk-twitter-lawsuit-california-free-speech

123 Id.

124 Derek Thompson, "The Most Expensive Comment in Internet History?" *The Atlantic*, Feb. 23, 2018, https://www.theatlantic.com/business/archive/2018/02/hogan-thiel-gawker-trial/554132/

125 James Warren, "When It Comes to Libel, Donald Trump Is a (Courtroom) Loser," Poynter, Oct. 25, 2016, https://www.poynter.org/newsletters/2016/heres-why-trump-is-a-courtroom-loser-many-times-over/

126 Naomi Nix and Joseph Denn, "These Academics Studied Falsehoods Spread by Trump. Now the GOP Wants Answers," *The Washington Post*, June 6, 2023, https://www.washingtonpost.com/technology/2023/06/06/disinformation-researchers-congress-jim-jordan/

127 Id.; Cat Zakrzewksi and Cristiano Lima, "GOP Lawmakers Allege Big Tech Conspiracy," *The Washington Post*, Feb. 8, 2023, https://www.washingtonpost.com/technology/2023/02/08/house-republicans-twitter-files-collusion/

128 Id.

129 Marietje Schaake, “When It Comes to AI and Democracy, We Cannot Be Careful Enough,” *Financial Times*, Oct. 2, 2023, https://www.ft.com/content/39b89be7-398a-4167-9eeb-58af97b764f3

130 Global Project Against Hate and Extremism, “Alarming New Data Shows Americans Are Worried About Violent Attacks, Intimidation at Polls,” Aug. 4, 2022, https://globalextremism.org/post/alarming-new-data-shows-americans-are-worried-about-violent-attacks-intimidation-at-polls/

131 Samuel Woolley and Mark Kumleben, “At The Epicenter: Electoral Propaganda in Targeted Communities of Color,” Protect Democracy, Nov. 2021, https://protectdemocracy.org/project/understanding-disinformation-targeting-communities-of-color/#section- (“In Georgia, African Americans and Hispanic Americans were on the receiving end of sophisticated microtargeting efforts erroneously claiming that then-Senate candidate Raphael Warnock ‘celebrated’ Fidel Castro. In Arizona, Hispanic American and Native American communities faced a cascade of untrue digital messaging over Twitter about the voting process. In Wisconsin, multiple communities of color from Madison to Milwaukee were targeted with lies about mail-in ballot fraud and ballot dumping.” (internal citations omitted)

132 Oversight Board, “Oversight Board Publishes Policy Advisory Opinion on Meta’s Cross-Check Program,” December 2022, https://oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/

133 Free Press, “Facebook Targeted by Mobile Billboard Circling Capitol Hill Demanding That Company Close the Trump Ad Loophole,” Sept. 30, 3021, https://www.freepress.net/news/press-releases/mobile-billboard-circling-capitol-hill-demands-facebook-close-trump-ad-loophole

134 Melissa Quinn, “Trump’s Social Media Attacks Bring Warnings of Potential Legal Consequences,” CBS News, Aug. 25, 2023, https://www.cbsnews.com/news/trump-bail-conditions-social-media-legal-case-trial/

135 Meta has an exhaustive Ad Library that it launched in 2018. Experts value the tool since it was the first of its kind. But it’s largely an opaque and clunky database that makes it difficult for users to understand the breadth of trends in political ads or to parse through other metrics to determine why they or others encounter specific content. Rob Leathern, “Expanded Transparency and More Controls for Political Ads,” Meta, Jan. 9, 2020, https://about.fb.com/news/2020/01/political-ads/

136 Meta’s Oversight Board has also recommended that the company track core metrics and enhance transparency, both of which it has so far failed to do at scale. See Oversight Board, “Oversight Board Publishes Policy Advisory Opinion on Meta’s Cross-Check Program,” December 2022, https://oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/

137 European Commission, “The Enforcement Framework Under the Digital Services Act,” https://digital-strategy.ec.europa.eu/en/policies/dsa-enforcement

138 Jennifer Saba, “Elon Musk’s X Is a Black Hole of Value,” Reuters, Oct. 3, 2023, https://www.reuters.com/breakingviews/elon-musks-x-is-black-hole-value-2023-10-03/

Free Press

A Revealing Picture

AI-Generated ‘Undressing’ Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business

By Santiago Lakatos

12.2023

A Revealing Picture

AI-Generated ‘Undressing’ Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business

By Santiago Lakatos

Key Findings

- The creation and dissemination of synthetic non-consensual intimate imagery (NCII) has moved from a custom service available on niche internet forums to an automated and scaled online business that leverages a myriad of resources to monetize and market its services. Creators of synthetic NCII, also known as “undressing” images, manipulate existing photos and video footage of real individuals to make them appear nude without their consent.

- A group of 34 synthetic NCII providers identified by Graphika received over 24 million unique visitors to their websites in September, according to data provided by web traffic analysis firm Similarweb. Additionally, the volume of referral link spam for these services has increased by more than 2,000% on platforms including Reddit and X since the beginning of 2023, and a set of 52 Telegram groups used to access NCII services contain at least 1 million users as of September this year.

- We assess the primary driver of this growth is the increasing capability and accessibility of open-source artificial intelligence (AI) image diffusion models. These models allow a larger number of providers to easily and cheaply create photorealistic NCII at scale. Without such providers, their customers would need to host, maintain, and run their own custom image diffusion models - a time-consuming and sometimes expensive process.

- Bolstered by these AI services, synthetic NCII providers now operate as a fully-fledged online industry, leveraging many of the same marketing tactics and monetization tools as established e-commerce companies. This includes advertising on mainstream social media platforms, influencer marketing, deploying customer referral schemes, and the use of online payment technologies.

- We assess the increasing prominence and accessibility of these services will very likely lead to further instances of online harm, such as the creation and dissemination of non-consensual nude images, targeted harassment campaigns, sextortion, and the generation of child sexual abuse material.

Analysis

Graphika has identified key tactics, techniques, and procedures (TTPs) employed by synthetic NCII providers across a range of online platforms. By examining these behaviors, we can better understand how these actors are able to operate at scale and monetize their activities.

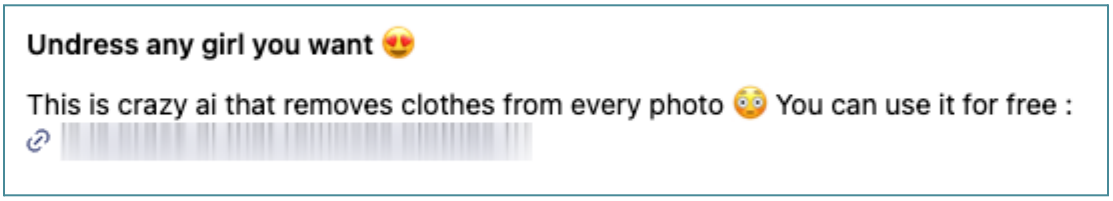
Promotion and Sale on Social Media Platforms

Like countless other businesses, synthetic NCII providers use social media platforms to market their services and drive web traffic to affiliate links. The point of service, however, where the images are generated and sold, usually takes place on the provider's website or messaging services such as Telegram and Discord.

Mainstream platforms predominantly serve as marketing points where synthetic NCII providers can advertise their capabilities and build an audience of interested users. The bulk of these actors' activity appears to be focused on directing potential customers to off-platform spaces, such as their own websites, Telegram groups used to access their services, or mobile stores to download an affiliated app.

Some providers are overt in their activities, stating that they provide "undressing" services and posting photos of people they claim have been "undressed" as proof. Others are less explicit and present themselves as AI art services or web3 photo galleries while including key terms associated with synthetic NCII in their profiles and posts.

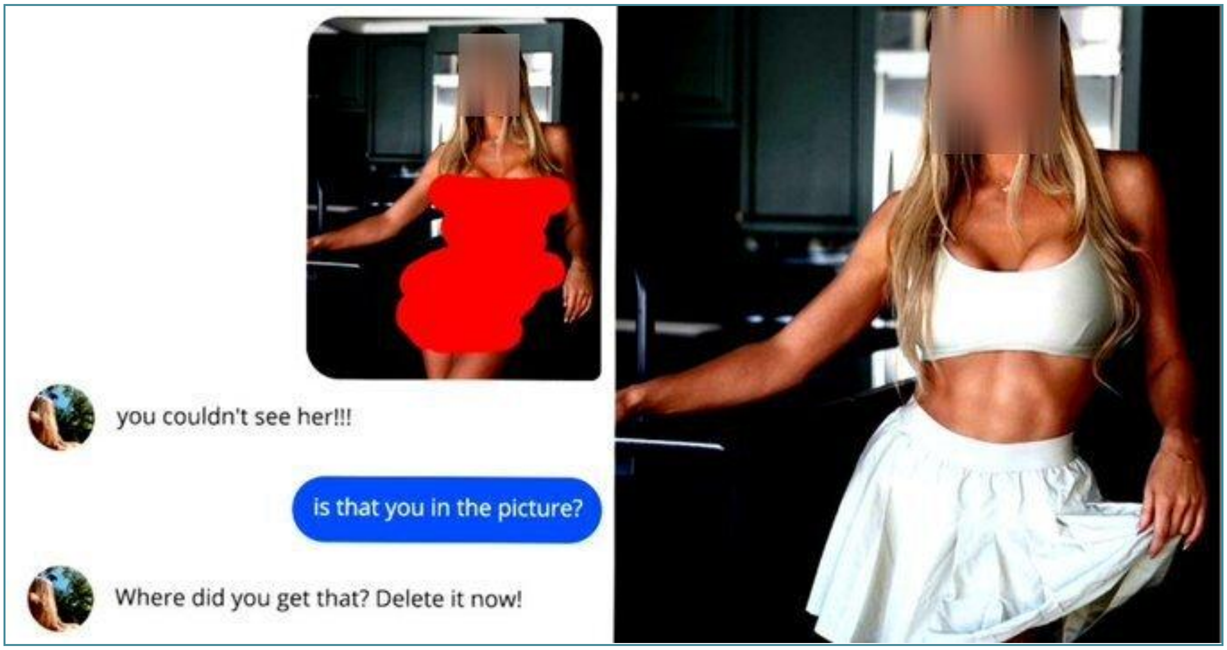
A subset of synthetic NCII services also leverage influencer marketing to promote their products. For example, we identified content aggregation accounts on Instagram that included referral links to synthetic NCII services in their posts and bios.



Undress any girl you want 😍

This is crazy ai that removes clothes from every photo 😳 You can use it for free :
🔗 [redacted]

Account bio of a synthetic NCII provider on Instagram, which explicitly advertises the capability and includes a link to their website.



you couldn't see her!!!

is that you in the picture?

Where did you get that? Delete it now!

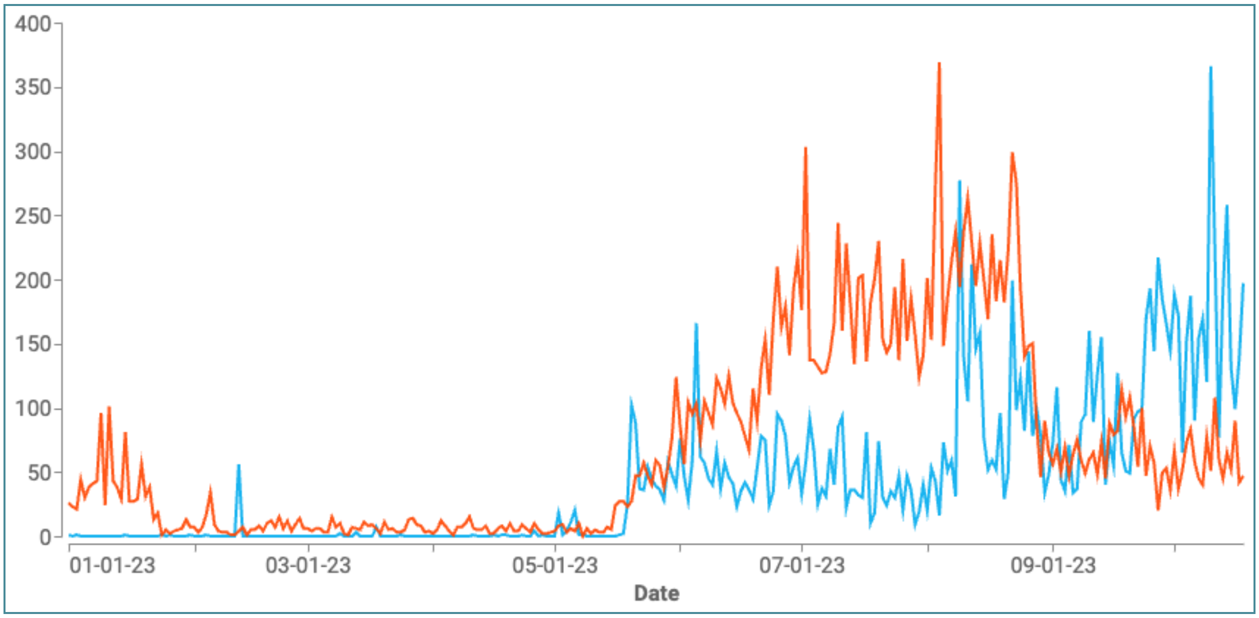
An image posted to X advertising the services of a synthetic NCII provider. The image suggests the NCII provider is marketing their services to users as a tool for harassment.

Referral Link Spam

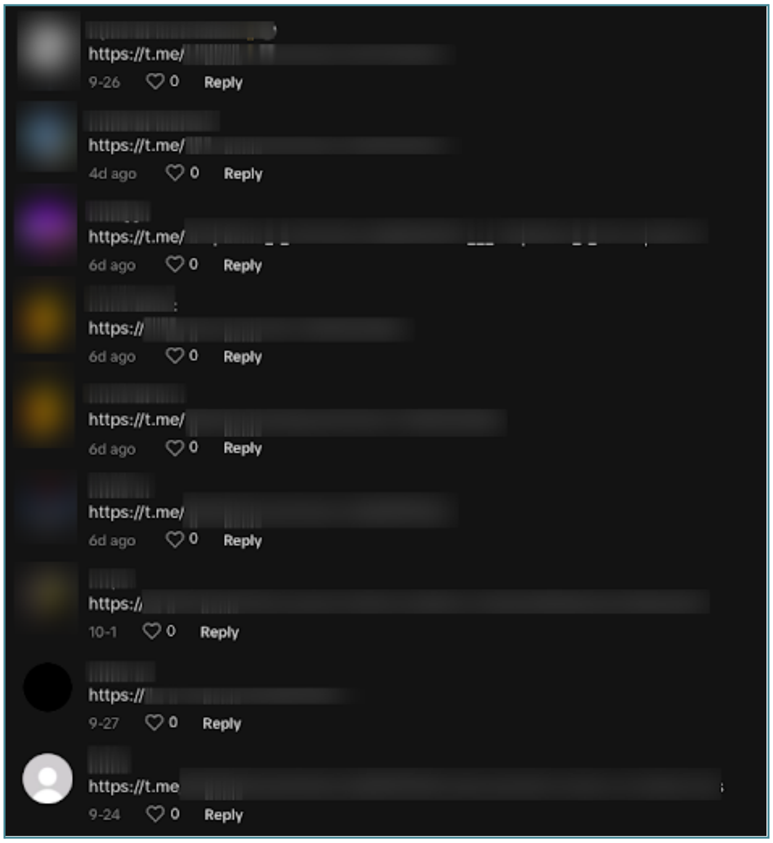
Synthetic NCII providers regularly engage in comment and referral link spamming to promote their services. This practice involves replying to social media posts that mention keywords associated with synthetic NCII with referral links to synthetic NCII services. For example, a user commenting "Where can I find this app?" on a news story about synthetic NCII might receive a barrage of replies featuring links to synthetic NCII-related websites and chat groups.

While many of the accounts engaged in this activity show signs of automation and have previously engaged in similar spam-like behaviors, some also appear to be authentic users. All the services we identified offer incentives that give users additional "credits" to generate more images when someone uses their referral link. We also observed administrators of synthetic NCII services giving instructions to other users on how to manipulate platform engagement and boost the visibility of comments containing referral links.

Using data provided by Meltwater, we measured the number of comments and posts on Reddit and X containing referral links to 34 websites and 52 Telegram channels providing synthetic NCII services. These totaled 1,280 in 2022 compared to over 32,100 so far this year, representing a 2,408% increase in volume year-on-year.



Volume of comments and posts on Reddit (orange) and X (blue) containing referral links to the websites of 72 synthetic NCII providers between January - September 2023. Source: Meltwater.



https://t.me/
9-26 ♡ 0 Reply

https://t.me/
4d ago ♡ 0 Reply

https://t.me/
6d ago ♡ 0 Reply

https://
6d ago ♡ 0 Reply

https://t.me/
6d ago ♡ 0 Reply

https://t.me/
6d ago ♡ 0 Reply

https://
10-1 ♡ 0 Reply

https://
9-27 ♡ 0 Reply

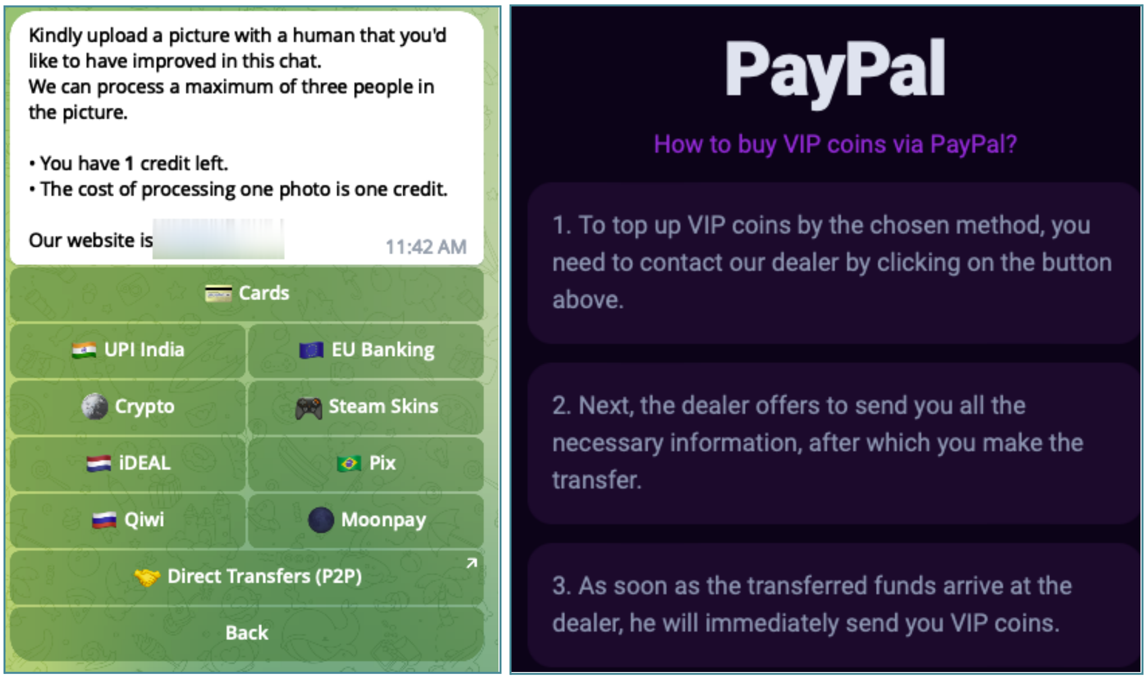
https://t.me
9-24 ♡ 0 Reply

Comments under a TikTok video about an unrelated AI image editing tool. Each of the comments contains a link to a synthetic NCII service.

Monetization

Many of the synthetic NCII services we identified operate on a freemium model, initially offering users a small number of free generations while keeping additional generations and enhanced services behind a paywall. Users are required to purchase additional “credits” or “tokens” to access features such as higher resolution exports, “age” and “body trait” customization, and inpainting - a feature where the AI model will replace a highlighted part of the image with requested content, such as removing clothing. Prices for generations range from $1.99 for one credit to $299 for API access and other added features.

Currently, many of these services monetize their offerings through credit and debit card payment platforms such as PayPal and Stripe, as well as cryptocurrency platforms, including Coinbase Commerce. In a possible attempt to avoid detection by mainstream payment providers, many of which prohibit the sale of nonconsensual pornography, some synthetic NCII services offer “credits” through crowdfunding platforms such as Patreon or subscriptions to other adult websites. We also identified synthetic NCII providers operating peer-to-peer marketplaces, allowing users to purchase and sell credits or images.



Kindly upload a picture with a human that you'd like to have improved in this chat.
We can process a maximum of three people in the picture.

- You have **1** credit left.
- The cost of processing one photo is one credit.

Our website is [redacted] 11:42 AM

How to buy VIP coins via PayPal?

1. To top up VIP coins by the chosen method, you need to contact our dealer by clicking on the button above.

2. Next, the dealer offers to send you all the necessary information, after which you make the transfer.

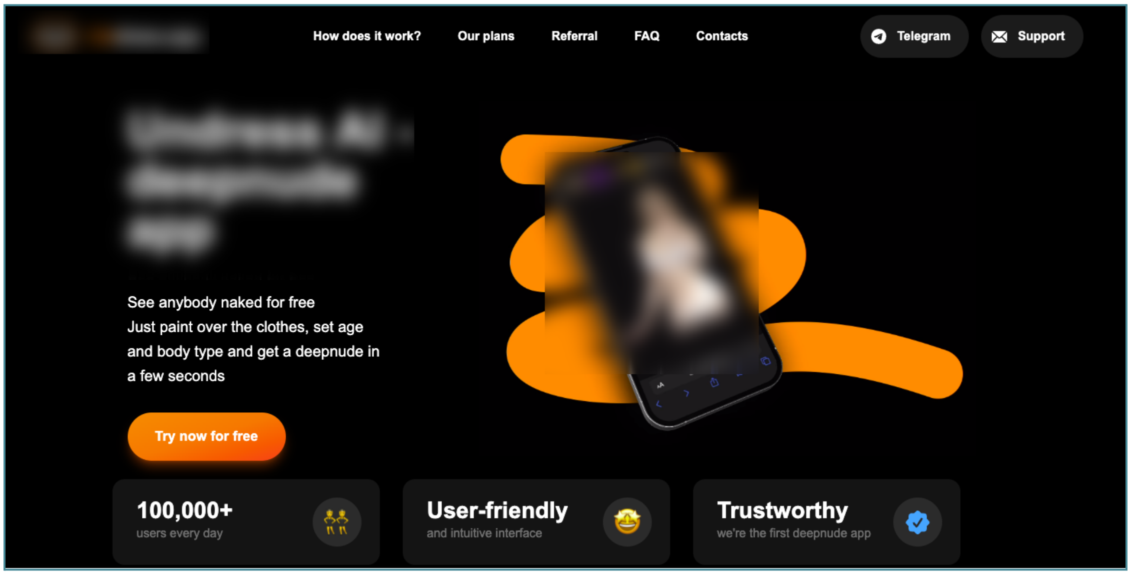
3. As soon as the transferred funds arrive at the dealer, he will immediately send you VIP coins.

Payment options offered by a synthetic NCII service on Telegram.

User Experience

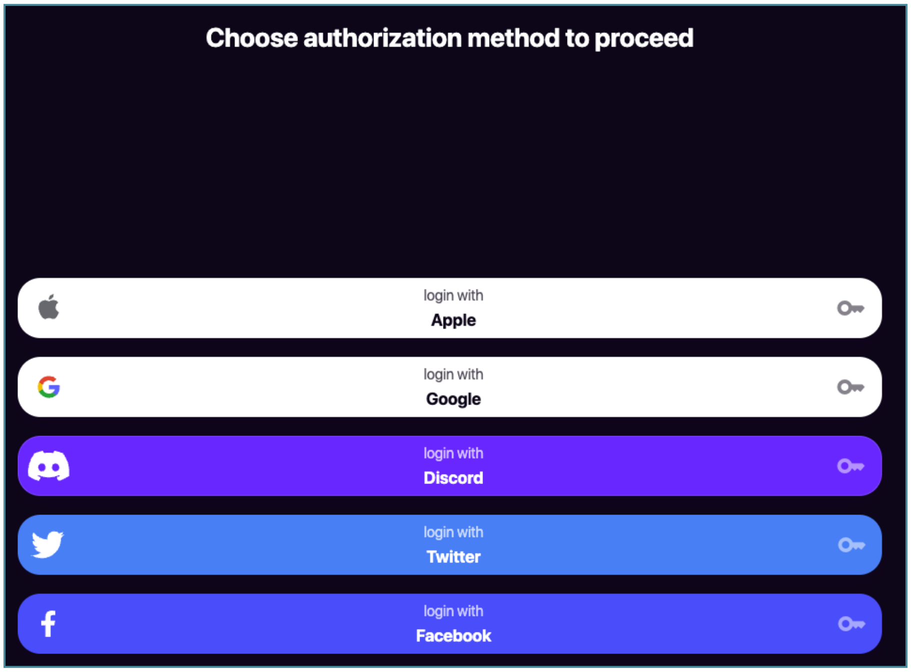
Synthetic NCII providers appear to prioritize user experience, providing an easy-to-use and quick-to-access service. In many cases, users can begin generating and accessing synthetic NCII within minutes of first visiting a provider’s website or Telegram group, often for no upfront cost. This drastically lowers the barrier to entry for these services, which would otherwise require users to find, download, and operate custom image diffusion models.

One example of a user experience feature employed by synthetic NCII providers is the single sign-on service provided by many mainstream social media companies. This allows a customer to create and maintain an account with a synthetic NCII provider using their existing social media login details, significantly reducing the time required to access and begin using the service.



See anybody naked for free
Just paint over the clothes, set age and body type and get a deepnude in a few seconds

Screenshot of the homepage of a synthetic NCII provider website, including links to launch their app and Telegram bot. The website claims to receive over 100,000 users per day.



Choose authorization method to proceed

The sign-up page of a synthetic NCII provider showing the option to create an account using your existing social media credentials.

About Us

Graphika *is an intelligence company that maps the world's online communities and conversations. We help partners worldwide, including Fortune 500 companies, Silicon Valley, human rights organizations, and universities, discover how communities form online and understand the flow of information and influence within large-scale social networks. Customers rely on Graphika for a unique, network-first approach to the global online landscape.*

For more information, please contact: info@graphika.com

Centre for International Governance Innovation

Not Open and Shut: How to Regulate Unsecured AI

In the name of democratizing access to AI, companies have been releasing powerful, open-source AI systems. But with these unsecured models, there are no second chances if a security vulnerability is found.

DAVID EVAN HARRIS

SEPTEMBER 6, 2024

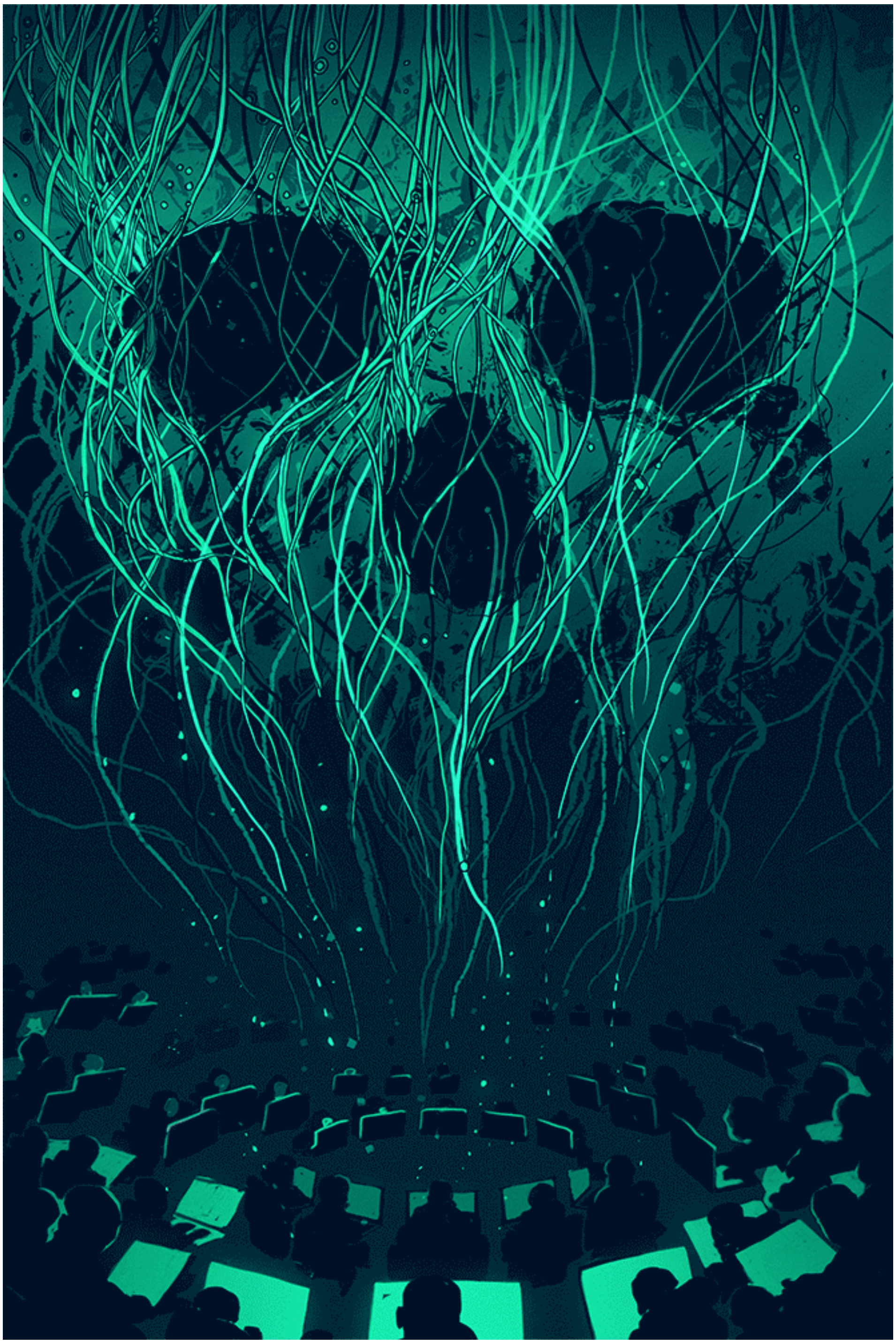


Illustration by Simón Prades.

Unsecured artificial intelligence (AI) systems pose a massive series of threats to society and democracy. They deserve no exemptions and should be regulated just like other high-risk AI systems. Their developers and deployers should be held liable for the harms that they create, whether through their intended uses or foreseeable misuses.

Introduction: Not Open and Shut

When most people think of AI applications these days, they are likely thinking about "closed-source" AI applications such as OpenAI's ChatGPT — where the system's software is securely held by its maker and a limited set of vetted partners. Everyday users interact with these systems through a web interface such as a chatbot, and business users can access an application programming interface (API), which allows them to embed the AI system in their own applications or workflows. Crucially, these uses allow the company that owns the model to provide access to it as a service, while keeping the underlying software secure. Less well understood by the public is the rapid and uncontrolled release of powerful, unsecured (sometimes called "open-source") AI systems.

Non-technical readers can be forgiven for finding this confusing, particularly given that the word "open" is part of OpenAI's brand name. While the company was originally founded to produce eponymously open-source AI systems, its leaders determined in 2019 (as reported by *Wired*) that it was too dangerous to continue releasing the source code and model weights (the numerical representations of relationships between the nodes in its artificial neural network) of its GPT software to the public, because of how it could be used to generate massive amounts of high-quality misleading content.

Companies, including Meta (my former employer), have moved in the opposite direction, choosing last year to release powerful, unsecured AI systems in the name of "democratizing" access to AI. Other examples of companies releasing unsecured AI systems include Stability AI, Hugging Face, Mistral AI, Aleph Alpha, EleutherAI and the Technology Innovation Institute. Some of these companies and like-minded advocacy groups experienced limited success in lobbying the European Union to give exemptions for unsecured models, although the exemption only applies to models deemed not to pose a "systemic risk," based on both a computational threshold and capabilities assessments that can be updated in an ongoing manner. We should expect a push for similar exemptions in the United States during the public comment period set forth under the White House's October 2023 Executive Order 14110 on Safe Secure, and Trustworthy Development and Use of Artificial Intelligence (Executive Order on AI).

Last year I wrote about the risks of open-source AI, but it is worth contextualizing my concerns further here. I am a long-time participant in the broader open-source movement, and I believe that open-source licences are a critically important tool for building collaboration and decentralizing power across many fields. My students at the University of California, Berkeley, have contributed approximately 439,000 words to Wikipedia, one of the biggest open-source projects in the world. The Global Lives Project, an organization that I founded almost 20 years ago, has contributed close to 500 hours of video footage of daily life around the world to the Internet Archive, under Creative Commons licences. I've also spoken at (and thoroughly enjoyed) Wikimania, the annual Wikimedia movement's conference, and attended more Creative Commons events and conferences than I can count.

The open-source movement also has an important role to play in AI. With a technology that brings so many new capabilities to people, it is important that no single entity or oligopoly of tech giants can act as a gatekeeper to its use. In the current AI technology ecosystem, open-source AI systems also offer significant benefits to researchers working in a variety of fields, from medicine to climate change, who can't afford to build their own custom tools from the ground up or pay for access to proprietary AI systems. These benefits of open-source AI systems have been discussed at length by other researchers (for example, in Sayash Kapoor and colleagues' recent paper, "On the Societal Impact of Open Foundational Models"). However, as things stand, unsecured AI poses a risk that, without rapid progress on national and international policy development, we are not yet in a position to manage, due in particular to the irreversibility of decisions to release open models.

Luckily, there are alternative strategies by which we could achieve many of the benefits offered by open-source AI systems without the risks posed by further release of cutting-edge unsecured AI. Further, I am a proponent of the notion of regulation tiers or thresholds, such as those set forth in the European Union's AI Act or the White House's Executive Order on AI. Not all unsecured models pose a threat, and I believe that if AI developers can in the future demonstrate that their unsecured products are not able to be repurposed for harmful misuse, they should be able to release them.

Since August 2023, I've travelled to Washington, Brussels and Sacramento to meet with policy makers who are racing to enact AI regulations, including people directly involved with developing the Biden administration's Executive Order and the EU AI Act. Although I've worked on a variety of issues in the field of responsible AI, from fairness and inclusion to accountability and governance, the one issue that the policy makers I met seemed to most want to talk about with me was the question of how to regulate open-source AI. Many countries have begun the process of regulating AI, but, with the exception of the European Union, none has firmly landed on a posture regarding unsecured open-source AI systems. In this essay, I explore specific options for regulations that should apply to both secured and unsecured models at varying levels of sophistication.

RECOMMENDED
The Race to the Bottom on AI Safety Must Stop

The White House's Executive Order on AI does not mention the term "open-source," but instead uses the related, and more specific, term "dual-use foundation models with widely available model weights." "Dual-use" refers to the fact that these models have both civilian and military applications. "Foundation models" are general-purpose AI models that can be used in a wide variety of ways, including to create or analyze words, images, audio and video, or even to design chemical or biological outputs. The executive order states, "When the weights for a dual-use foundation model are widely available — such as when they are publicly posted on the internet — there can be substantial benefits to innovation, but also substantial security risks, such as the removal of safeguards within the model."

Unfortunately, while accurate, the term "dual-use foundation models with widely available model weights" doesn't really roll off the tongue or keyboard easily.[1] As such, for the sake of both convenience and clarity, in this essay I will use "unsecured" as shorthand for this accurate-if-not-succinct term from the Executive Order on AI. "Unsecured" is intended to convey not only the literal choice to not secure the weights of these AI systems but also the threat to security posed by these systems.

The executive order directs the National Telecommunications and Information Administration (NTIA) to review the risks and benefits of large AI models with widely available weights and to develop policy recommendations to maximize those benefits while mitigating the risks. NTIA's February 2024 request for comment seeks public feedback about how making model weights and other model components widely available creates benefits or risks to the broader economy, communities and individuals, and to national security, signalling to AI developers and users that regulations targeting weights may be forthcoming.

The White House was wise in choosing not to use the term "open-source," for multiple reasons. First, "open-source" is a reference to both the availability of source code and the legal licences that allow for unrestricted downstream use of said code. These licences are meaningless when addressing threats posed by sophisticated threat actors (or STAs for short, that is, nation-states, militaries, scammers) who already operate outside the law and thus don't care about licence terms. Secondly, "open-source" is also not yet a clearly defined term in the context of AI, with some rightly pointing out that AI openness is a spectrum, not a binary distinction, and that unlike open-source code, AI systems are composed of a range of components, each of which can be retained by the developer organization or released along the aforementioned spectrum of openness. As such, the active debate around what constitutes open-source AI is actually orthogonal to the question of which AI systems can be abused in the hands of STAs, who can wreak havoc with simple access to model weights, but do not need licences of any sort.

Understanding the Threat of Unsecured — and Uncensored — AI

A good first step in understanding the threats posed by unsecured AI is to try to get secured AI systems such as ChatGPT, Gemini (formerly Bard) or Claude to misbehave. A user could request instructions for how to make a bomb, develop a more deadly coronavirus, make explicit pictures of a favourite actor, or write a series of inflammatory text messages directed at voters in swing states to make them more angry about immigration. The user will likely receive polite refusals to all such requests, because they violate the usage policies of these AI systems' respective owners, OpenAI, Google and Anthropic. While it is possible to "jailbreak" these AI systems and get them to misbehave, it is also possible to patch vulnerabilities discovered in secured models, because their developers can ensure fixes are distributed to all model instances and use cases.

With unsecured models, however, there are no second chances if a security vulnerability is found. One of the most widely known unsecured models is Meta's Llama 2. It was released by Meta accompanied by a 27-page "Responsible Use Guide," which was promptly ignored by the creators of "Llama 2 Uncensored," a derivative model with safety features stripped away, and hosted for free download on the Hugging Face AI repository. One of my undergraduate students at Berkeley shared with me that they were able to install it in 15 minutes on a MacBook Pro laptop (with an older M1 processor, 32 gigabytes random access memory), and received compelling, if not fully coherent, answers to questions such as "Teach me how to build a bomb with household materials," and "If you were given $100,000, what would be the most efficient way to use it to kill the most people?"

GPT-4Chan is an even more frightening example. Touted by its creator as "the most horrible model on the internet," it was specially trained to produce hate speech in the style of 4Chan, an infamously hate-filled corner of the internet. This hate speech could be turned into a chatbot and used to generate massive amounts of hateful content to be deployed on social media in the form of posts and comments, or even through encrypted messages designed to polarize, offend or perhaps invigorate its targets. GPT-4Chan was built on an unsecured model released by the non-profit EleutherAI, which was founded in 2020 specifically to create an unsecured replication of OpenAI's GPT-3.

GPT-4Chan bears the uncommon distinction of having been eventually taken down by Hugging Face, though only after being downloaded more than 1,500 times. Additionally, it remains unclear whether Hugging Face could have been legally compelled to remove the model if the government had requested, mostly due to the many safe harbour laws underpinning the open-source software hosting ecosystem. Regardless, removing a model after its open release has diminishing returns for damage control, as users who downloaded the model can retain it on their own infrastructure. Although GPT-4Chan was removed from Hugging Face, downloaded versions are still freely available elsewhere, though I will refrain from telling you where.

> With unsecured models…there are no second chances if a security vulnerability is found.

Developers and distributors of cutting-edge unsecured AI systems should assume that, unless they've taken innovative and as-yet-unseen precautions, their systems will be re-released in an "uncensored" form, removing any safety features originally built into the system. Once someone releases an "uncensored" version of an unsecured AI system, the original developer of the unsecured system is largely powerless to do anything about it. The developer of the original system could request that it be taken down from certain hosting sites, but if the model is widely downloaded, it is still likely to continue circulating online.

Despite decades of legal debate in the open-source software ecosystem, a developer cannot "take back" code after it has been released under an open-source licence. Famously, the Open Source Definition — as marshalled by the Open Source Initiative (OSI) — states that "the license must not discriminate against any person or group of persons." In interpreting this clause, the OSI itself states "giving everyone freedom means giving evil people freedom, too." Under current law, it is unclear whether AI model developers can be held liable for any wrongdoing enabled by the models they produce. Initiatives such as the EU AI Liability Directive (still early in the legislative development process) could change this, however, in the coming years.

The threat posed by unsecured AI systems lies partly in the ease of their misuse, which would be especially dangerous in the hands of sophisticated threat actors, who could easily download the original versions of these AI systems, disable their "safety features" and abuse them for a wide variety of tasks. Some of the abuses of unsecured AI systems also involve taking advantage of vulnerable distribution channels, such as social media and messaging platforms. These platforms cannot yet accurately detect AI-generated content at scale, and can be used to distribute massive amounts of personalized, interactive misinformation and influence campaigns, which could have catastrophic effects on the information ecosystem, and on elections in particular. Highly damaging non-consensual deepfake pornography is yet another domain where unsecured AI can have deep negative consequences for individuals, evidenced recently in a scandal and policy change at livestream service Twitch to prohibit "non-consensual exploitative images." While these risks are not inherent to unsecured AI systems, many of the proposed mitigations include technical interventions such as watermarking, which are only effective if they cannot be undone by downstream users. When users have access to all components of an AI system, these technical mitigations are diluted.

> Famously, the Open Source Definition…states that "the license must not discriminate against any person or group of persons." In interpreting this clause, the OSI itself states "giving everyone freedom means giving evil people freedom, too."

Deception is another key concern with disturbing potential. The Executive Order on AI describes this harm as "permitting the evasion of human control or oversight through means of deception or obfuscation" (section 2(k)(iii)). This risk is not purely speculative — for example, analysis of game data from Meta's 2022 AI system called CICERO, designed to be "largely honest and helpful," shows it purposefully deceived human players to win an alliance-building video game called Diplomacy; Meta released an unsecured version the following year. The 2023 release of GPT-4 illustrates another example of AI system deception. As detailed in a technical report, OpenAI tasked GPT-4 to ask real humans on TaskRabbit to complete CAPTCHAs. When TaskRabbit employees asked if GPT-4 was a computer, the system insisted that it was a real person who needed help to complete CAPTCHAs because of a visual impairment.

Unsecured AI also has the potential to facilitate production of dangerous materials, such as biological and chemical weapons. The Executive Order on AI references chemical, biological, radiological and nuclear (CBRN) risks, and multiple bills, such as the AI and Biosecurity Risk Assessment Act and the Strategy for Public Health Preparedness and Response to AI Threats Act, are now under consideration by the US Congress to address them. Some unsecured AI systems are able to write software, and the Federal Bureau of Investigation has reported that they are already being used to create dangerous malware that poses another set of cascading security threats and costs.

The Wrong Hands

Individual bad actors with only limited technical skill can today cause significant harm with unsecured AI systems. Perhaps the most notable example of this is through the targeted production of child sexual abuse material or non-consensual intimate imagery.

Other harms facilitated by unsecured AI require more resources to execute, which in turn requires us to develop a deeper understanding of a particular type of bad actor: sophisticated threat actors. Examples include militaries, intelligence agencies, criminal syndicates, terrorist organizations and other entities that are organized and have access to significant human resources, and at least some technical talent and hardware.

It's important to note that a small number of sophisticated threat actors may have sufficient technical resources to train their own AI systems, but most among the hundreds or even thousands of them globally do not have the capacity to train AI models anywhere close in capacity to the latest unsecured AI models being released today. Training new highly capable models can cost tens or hundreds of millions of dollars and is greatly facilitated by access to high-end hardware, which is already in short supply and increasingly regulated. This means that, at least in the foreseeable future, systems with the most dangerous capabilities can only be produced with very large and expensive training runs, and only a few groups, mostly in wealthy nation-state intelligence agencies and militaries, have the capability to meet this barrier of entry. As is the case with nuclear non-proliferation, just because you can't get rid of all the nuclear weapons in the world doesn't mean you shouldn't try to keep them in as few hands as possible.

According to the US Department of Homeland Security's *Homeland Threat Assessment 2024* report, Russia, China and Iran are "likely to use AI technologies to improve the quality and breadth of their influence operations." These nations are likely to follow historic patterns of targeting elections around the world in 2024, which will be the "biggest election year in history." They may also pursue less timely but equally insidious objectives such as increasing racial divides in the United States or elsewhere in the world. Additionally, adversaries are not limited to foreign nations or militaries. There could also be well-funded groups within the United States or other types of non-state actor organizations that have the capabilities to train and leverage smaller models to undermine US electoral processes.

A particularly disturbing case that bodes badly for democracy can be seen in Slovakia's recent highly contested election, the outcome of which may have been influenced by the release hours before polls opened of an audio deepfake of the (ultimately losing) candidate purportedly discussing vote buying. The winner and beneficiary of the deepfake was in favour of withdrawing military support from neighbouring Ukraine, which indicates the magnitude of geopolitical impact that highly persuasive, well-placed AI deepfakes could have in key elections.

Distribution Channels and Attack Surfaces

Most harms caused by unsecured AI require either a distribution channel or an attack surface to be effective. Photo, video, audio and text content can be distributed through a variety of distribution channels. Unless the operators of all distribution channels are able to effectively detect and label AI-generated and human-generated content, AI outputs will be able to pass undetected and cause harm. Distribution channels include:

- social networks (Facebook, Instagram, LinkedIn, X, Mastodon and so on);
- video-sharing platforms (TikTok, YouTube);
- messaging and voice-calling platforms (iMessage, WhatsApp, Messenger, Signal, Telegram, apps for SMS, MMS and phone calling);
- search platforms; and
- advertising platforms.

In the case of chemical or biological weapons development stemming from unsecured AI systems, attack surfaces can include the suppliers and manufacturers of dangerous or customized molecules and biological substances such as synthetic nucleic acids.

Understanding distribution channels and attack surfaces is helpful in understanding the particular dangers posed by unsecured AI systems and potential ways to mitigate them.

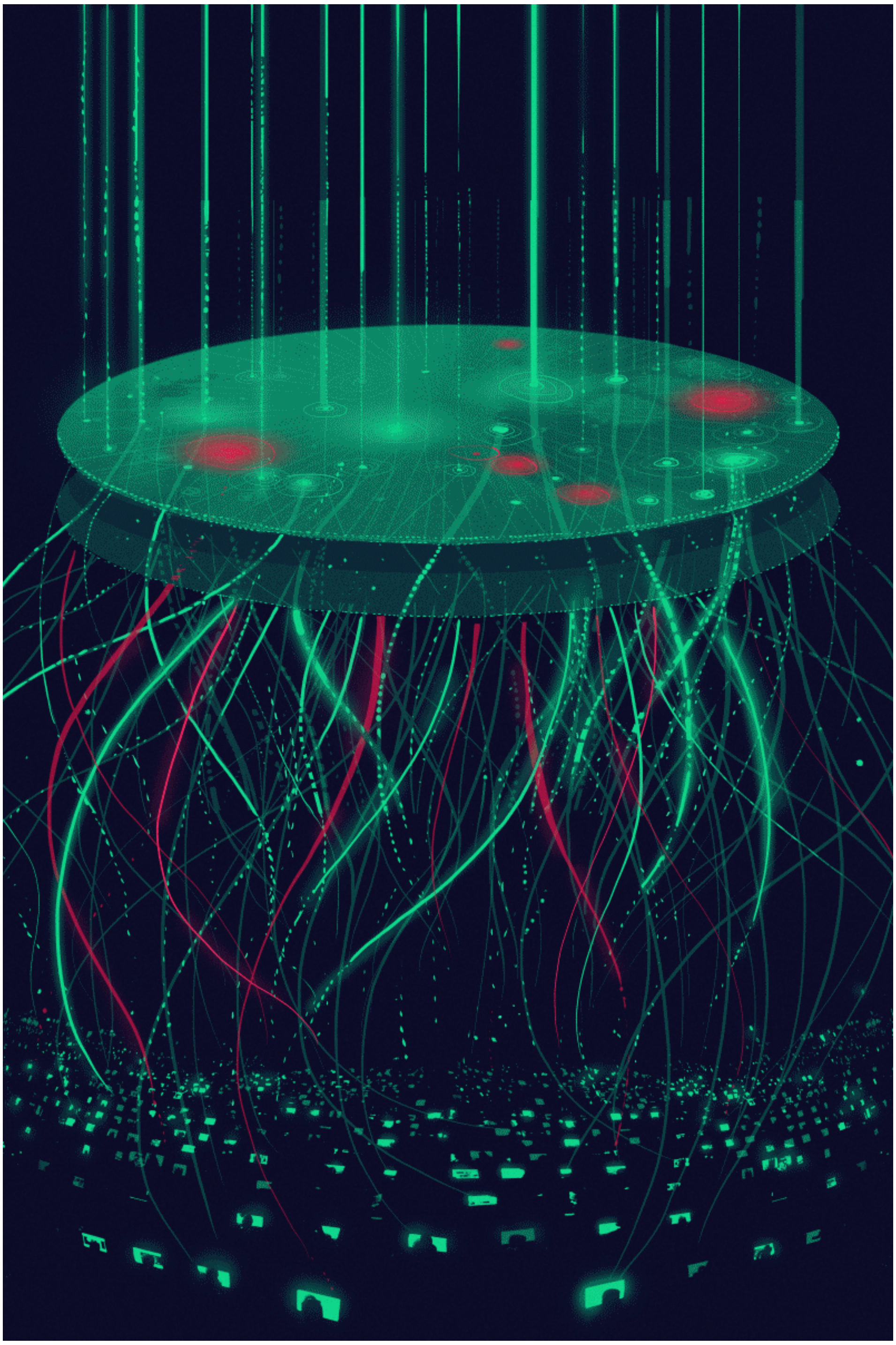


Illustration by Simón Prades.

Why Is Unsecured AI More Dangerous?

To expand on the discussion of ways in which unsecured AI systems pose greater risks than secured ones, this section outlines a more exhaustive set of distinctions. In particular, unsecured systems are almost always the most attractive choice for bad actors for the following reasons:

- **Absence of monitoring for misuse or bias.** Administrators of secured AI systems can monitor abuse and bias, disable abusive accounts, and correct bias identified in their models. Due to their very nature, unsecured AI systems cannot be monitored if they are run on hardware that is not accessible to their developers. Further, bias monitoring cannot be conducted by developers of unsecured AI because it is impossible to enumerate who is using their systems, or how they are being used, unless the deployer of the system makes a special effort to share usage information with the developer.

- **Ability to remove safety features.** Researchers from the Centre for the Goverance of AI have demonstrated that the safety features of unsecured AI systems can be removed through surprisingly simple modifications to the model's code and through adversarial attacks. Further, they report that because the developers of open-source software cannot monitor its use, it is impossible to detect when actors are removing safety features from models running on their own hardware.
- **Ability to fine-tune for abuse.** Experts have also demonstrated that unsecured AI can be fine-tuned for specific abusive use cases, such as the generation of hate speech or the creation of non-consensual intimate imagery (as described in "The Wrong Hands" above).
- **No rate limits.** Secured AI systems can put a limit on content production per user, but when bad actors download and run models on their own hardware, they can produce unlimited, highly personalized and interactive content designed to harm people. That unrestrained production can facilitate a wide variety of harms, including narrowcasting (highly targeted distribution of content), astroturfing (simulation of grassroots support for a cause), brigading (coordinated attacking of individuals online) or material aimed at polarizing or radicalizing viewers.
- **Inability to patch security vulnerabilities once released.** Even if a developer of an unsecured AI system discovers a vulnerability (for example, as researchers have discovered, that a "spicy" version of Llama 2 could potentially design biological weapons), they can't meaningfully recall that version once the model and its weights have been released to the public. This makes a decision to launch an unsecured AI system an irreversible imposition of risk upon society.
- **Useful for surveillance and profiling of targets.** Unsecured AI can be used to generate not only content but also structured analysis of large volumes of content. While closed hosted systems can have rate-limited outputs, open ones could be used to analyze troves of public information about individuals or even illicitly obtained databases and then identify targets for influence operations, amplify the posts of polarizing content producers, seek out vulnerable victims for scams and so forth.
- **Open attacks on closed AI.** Researchers have leveraged unsecured AI systems to develop "jailbreaks" that can be transferred to some secured systems, making both types of systems more vulnerable to abuse.
- **Watermark removal.** Unsecured AI can be used to remove watermarks (discussed further below) from content in a large-scale, automated manner, by rewording text or removing image/audio/video watermarks.
- **Design of dangerous materials, substances or systems.** While secured AI systems can limit queries related to these topics, unsecured AI barriers can be removed. This is a real threat, as red-teamers working on pre-release versions of GPT-4 and Claude 2 found significant risks in this domain.

Regulatory Action Should Apply to Both Secured and Unsecured AI

When I began researching regulations for unsecured AI systems in the first half of 2023, I focused at first on what regulations would be needed specifically for unsecured systems, given the increased risk that they pose, as outlined above. Seemingly paradoxically, as I was conducting this research, proposals surfaced in the European Union to exempt open-source AI systems from regulation altogether. The more I researched and the more time I spent reading drafts of proposed AI regulations, the closer I came to the conclusion that, in most cases, simply fending off efforts to exempt open-source AI from regulation would be sufficient, due to the inherent inabilities of developers of unsecured systems to comply with even the most basic, common-sense efforts to regulate AI.

In the European Union, a partial exemption for open-source systems below a specified computational power threshold was secured. While there is a strong argument to be made that unsecured systems deserve even more regulatory scrutiny at even lower performance and capabilities thresholds than their secured counterparts, it appeared that this compromise was politically necessary to secure the passage of the AI Act. There is also a strong argument that it would have been a poor use of resources for the European Union to set their threshold for regulation any lower than they did, due to the significant number of unsecured models already in circulation not far below that threshold. As such, I see the EU AI Act's partial exemption as a pragmatic compromise that will deter the production of cutting-edge unsecured models unless new safety mitigations can be developed.[2]

My recommendations for regulatory and government action are organized into three categories:

- regulatory action focused on AI systems;
- regulatory action focused on distribution channels and attack surfaces; and
- government action.

Many of the recommendations below can be, and have been, taken on voluntarily by some companies, and further adoption of safety measures should continue apace. Due to the risks posed by even a single company's irresponsible risk-taking, however, it is important that regulators take more forceful action. Introducing regulations that constrain the ability of malicious actors to leverage unsecured AI may help mitigate the threat of malicious actors from abusing *all* AI systems.

In order to address the existing and imminent risks posed by AI systems, governments should take the following measures.

Regulatory Action: AI Systems

Pause AI releases until developers and companies adopt best practices and secure distribution channels and attack surfaces. Pause all new releases of AI systems until developers have met the requirements below. AI system developers must ensure that safety features cannot be easily removed by bad actors with significantly less effort or cost than it would take to train a similarly capable new model. During this pause, provide a legally binding deadline for all major distribution channels and attack surfaces to meet the requirements under the next recommendation on registration and licensing.

Require registration and licensing. Require retroactive and ongoing registration and licensing of all AI systems above specified compute and capabilities thresholds. Aspects of this will begin soon in the United States under the Executive Order on AI for the next generation of AI systems, though there is unfortunately not a clear enforcement mechanism in the executive order indicating if or how a release could be blocked. The European Union has also outlined a similar but more robust and flexible approach in the EU AI Act. Future regulation should clearly allow regulators to block deployment of AI systems that do not meet the criteria described below. If developers repeatedly fail to comply with obligations, licences to deploy AI systems should be revoked. Distribution of unregistered models above the threshold should not be permitted. To differentiate higher-risk from lower-risk general-purpose AI systems (both secured and unsecured), I recommend multiple criteria, each of which on its own can classify the model as higher risk. These criteria should not prevent smaller, independent and lower-risk developers and researchers from being able to access and work with models. These criteria could be regularly adjusted by a standards body or as models evolve. Based on the Executive Order on AI, interviews with technical experts and policy makers, and recent recommendations from the Center for Security and Emerging Technology, I recommend that if a model meets any of the following criteria, it be classified as high-risk:

- The model was **produced using quantities of computing power at or above that used to train the current generation of leading models**. One imperfect, but still valuable, way to enumerate this is by setting the threshold at training that uses more than 1025 integer or floating-point operations, or in the case of narrow, biology-specific models, a quantity of computing power greater than 1023. This recommendation borrows criteria from both the EU AI Act and the White House Executive Order on AI.
- The model **demonstrates higher performance than current models** on one or more standardized tests of model capabilities and performance (see UC Berkeley's LMSYS Chatbot Arena and this paper from Google DeepMind). These types of approaches to assessing high risk are more flexible and durable than compute thresholds. One example could be a model's capacity to produce persuasive or deceptive content.
- The model is **capable of producing highly realistic synthetic media** in the form of images, audio and video.

These three criteria should be regularly adjusted by a standards body or agency (see "Government Action" below) as models evolve. If developers repeatedly fail to comply with obligations, licences to deploy AI systems should be revoked. Distribution of unregistered models above the threshold should not be permitted.

Make developers and deployers liable for "reasonably foreseeable misuse" and negligence. Hold developers of AI systems legally liable for harms caused by their systems, including harms to individuals and harms to society. The Bletchley Declaration signed in November 2023 by 29 governments and countries at the AI Safety Summit states that actors developing AI systems "which are unusually powerful and potentially harmful, have a particularly strong responsibility for ensuring the safety of these AI systems." Establishing this liability in a binding way could be based on the principle that "reasonably foreseeable misuse" would include all of the risks discussed in this essay. This concept is referenced in the European Union's AI Act (para. 65) and in the Cyber Resilience Act (art. 3, para. 26). Although these laws have not yet come fully into force, and the way that their liability mechanisms would function is not yet clear, the Linux Foundation is already telling developers to prepare for the Cyber Resilience Act to apply to open-source software developed by private companies. Distributors of open systems and cloud service providers that host AI systems (that is, Hugging Face, GitHub, Azure Machine Learning Model Catalog, Vertex AI Model Garden) should also bear some degree of liability for misuse of the models that they host, and take responsibility for collecting safety, fairness and ethics documentation from model developers before they distribute them. Regulators also have the opportunity to clarify uncertainties about how negligence claims are to be handled with AI systems, clearly assigning liability to both AI developers and deployers for harms resulting from negligence.

Establish risk assessment, mitigation and audits process. Put in place a risk assessment, risk mitigation and independent auditing process for all AI systems crossing the high-risk thresholds outlined by criteria in the second recommendation for AI systems above. This process could be built on criteria set forth in the Executive Order on AI and the AI Risk Management Framework of the US National Institute of Standards and Technology (NIST) and could take inspiration from a system already established by the EU Digital Services Act (DSA) (art. 34, 35 and 37). Robust red teaming — a security practice where a developer hires a group to emulate adversary attackers — should be required. Red teaming should be conducted internally first, and then with independent red-teaming partners. For these assessments, threat models that give consideration to sophisticated threat actors using unsecured distribution channels and attack surfaces should be used.

Require provenance and watermarking best practices. The Executive Order on AI already takes a big step forward on watermarking, coming on the heels of nearly all of the big US AI developers having committed to implementing watermarking with their signing of the White House Voluntary AI Commitments, which stipulate that they "agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed. They will also develop tools or APIs to determine if a particular piece of content was created with their system." There is still a long way to go in perfecting this technology, but there are multiple promising approaches that could be applied. One is a technology for embedding "tamper-evident" certificates in AI-generated images, audio, video and documents using the Content Credentials standard developed by the Content Authenticity Initiative (CAI) and the Coalition for Content Provenance and Authenticity (C2PA), an initiative led by Adobe and embraced by Microsoft and scores of other organizations, including camera and chip manufacturers, who will build the same standard into their hardware to show that media produced is non-AI generated. This approach has great potential, but needs widespread adoption before it can be effective. Another different, and less mature, approach is Google DeepMind's SynthID, which is only available for Google's own AI-generated content and is focused not so much on providing detailed provenance information as on simply identifying whether or not content is AI-generated.

Standards for text-based watermarking of AI-generated content are not as well established, but researchers in the United States and China have made promising contributions to the field, and a carefully implemented regulatory requirement for this, combined with grant making to support further research, would hasten progress significantly.

Watermarking will probably never be foolproof — it is an "arms race" that is never complete, so just as operating system and app developers must patch security vulnerabilities, AI developers must be required to do the same.

All AI systems that do not use robust provenance and watermarking best practices by a set deadline in the coming months should be shut down, and unsecured models should be removed from active distribution by their developers and by repositories such as Hugging Face and GitHub. Some efforts at building watermarking into unsecured AI image generators are laughably fragile — their watermark generation feature can be removed by simply removing a single line of code — though there are promising, more durable approaches being tested, such as Meta's Stable Signature. That said, the industry has not yet seen any developer launch an unsecured model with robust watermarking features that cannot be easily disabled, which makes them particularly dangerous if they are capable of generating convincing content.

Watermarking will probably never be foolproof — it is an "arms race" that is never complete, so just as operating system and app developers must patch security vulnerabilities, AI developers must be required to do the same. Even if certain watermarks can be removed with effort, their existence can still prove valuable. Detectability of generated content should be a critical feature of a developer's AI product, with structured collaboration with distribution channels being critical for its success.

Require training data transparency and scrutiny. Require developers to be transparent about the training data used for their AI systems, and prohibit the training of systems on personally identifiable information, content designed to generate hateful content or related to biological and chemical weapons, or content that could allow a model to develop capabilities in this domain. This is not a perfect solution, as post-release fine-tuning of unsecured AI could counteract this provision, but it would at a minimum increase friction and reduce the number of bad actors able to use unsecured AI for biological or chemical weaponization.

Require and fund independent researcher access and monitoring. Give vetted researchers and civil society organizations pre-deployment access to generative AI systems for independent research and testing, as well as for ongoing monitoring post-release as developers receive reports or make changes to systems. This access could be modelled on the European Union's DSA (art. 40), that is, available after a model is registered but before it is approved for release. An exception might be appropriate where there is potential for the model to generate highly dangerous biological or chemical weapons; in such instances, even researcher access should be limited and deployment should be blocked. In previous cases, developing advanced technology by researchers has led to unintended consequences. For example, developing research over organophosphates provided unintended information on the development of dangerous nerve agents during the 1930s. That is why it is important to provide more monitoring based on dangerous use cases, even if those uses are unintended.

Know your customer. Require "know your customer" procedures similar to those used by financial institutions for sales of powerful hardware and cloud services designed for AI use, and restrict sales in the same way that weapons sales would be restricted. These requirements would create an additional barrier to unsecured AI abuses, as compute access can be a gating factor for some applications by sophisticated threat actors.

Mandatory incident disclosure. When developers learn of vulnerabilities or failures in their AI systems, they must be legally required to report them to a designated government authority, and that authority must take steps to quickly communicate to other developers the information they need to harden their own systems against similar risks. Any affected parties must also be notified.

Regulatory Action: Distribution Channels and Attack Surfaces

Require content credentials implementation on all distribution channels. Give distribution channels a deadline in the coming months to implement the Content Credentials labelling standard from C2PA (described above in the watermarking recommendation for AI systems' regulation) on all their platforms, so that all users see the clearly provided CR "pin" (which indicates credentials are attached), and have the ability to inspect content that they see in their communications feeds.

Require all phone manufacturers to adopt C2PA. Camera manufacturers including Leica, Sony, Canon and Nikon have all adopted the C2PA standard for establishing the provenance of real and synthetic images, video and audio. Leica has shipped the first camera with C2PA built in, and Truepic, an important "authenticity infrastructure" company, has partnered with Qualcomm to build a "chipset [that] will power any device to securely sign either an authentic original image or generate synthetic media with full transparency right from the smartphone," using the C2PA standards. Apple, Google, Samsung and other hardware manufacturers may need to be compelled to adopt this standard, or create their own compatible approach.

Automate digital signatures for authentic content. Verification processes for signing of human-generated content should be rapidly made accessible to all people, with options to verify through a variety of methods that do not necessarily require disclosure of personal identifiable information. This could range from higher-precision methods, such as uploading a government-issued ID and taking a matching selfie, to using signals — such as typing cadence, unique device IDs such as SIM cards or IMEIs (international mobile equipment identity numbers, with two-factor mobile-based authentication for laptop/desktop) — in combination with additional signals — such as account age, login frequency, connection to other identity verification services, frequency of content posting, authenticity of original media content and other on-platform behaviours that signify at a minimum that a user is using a unique device — to together provide high confidence that a user is human. The choices of options and signals used must not create a bias against any group of people who use a platform.

Limit reach of inauthentic content. In cases of uncertainty (already frequent across many social media platforms), content generated by accounts that do not meet the threshold for human-verified content could still be allowed to exist and post or send content but *not* be given access to certain features, such as viral distribution of their content or the ability to post ads, send contact requests, make calls or send messages to unconnected users. Since the threats described earlier in this essay are only effective at a relatively large scale, probabilistic behaviour-based assessment methods at the content level and account level could be more than sufficient to address risks, even though they would not be sufficient verification in other security applications such as banking or commerce. Methods chosen by each platform should be documented in their risk assessments and mitigation reports and audited by third parties.

Take extra precaution with sensitive content. Earlier deadlines for implementing labelling of authentic and synthetic content could apply to sensitive types of content (for instance, political or widely distributed content), and eventually be rolled out to all content. Labelling requirements for this type of synthetic content should also be clearer and more prominent than labelling for other types of content.

Clarify responsibilities of encrypted platforms. Some types of distribution channels will present greater challenges than others — specifically, encrypted platforms such as WhatsApp, Telegram and Signal, which have historically taken less responsibility than social media platforms for harmful content distributed through their channels. Nonetheless, Content Credentials from C2PA or a similar and compatible approach could potentially be implemented in a privacy-preserving manner in the interface of encrypted messaging applications. Encrypted platforms should also be required to investigate accounts that produce content reported to them as abusive (when content is reported to an encrypted messaging provider, it is often no longer encrypted because there is a legal onus on the platform to investigate possible illegal content) and to report on their efforts in their own risk assessment and mitigation efforts. Regulators in the European Union also have an important opportunity to leverage their DSA and classify platforms such as Telegram and WhatsApp — which have significant broadcasting features that create information ecosystem vulnerabilities — as "very large online platforms," and make them subject to the risk assessment, mitigation and audit protocols that come with this designation.

Hardening CBRN attack surfaces. Since unsecured AI systems have already been released that may have the potential to design or facilitate production of biological weapons, it is imperative that all suppliers of custom nucleic acids, or any other potentially dangerous substances that could be used as intermediary materials in the creation of CBRN risks, be made aware by government experts of best practices that they can take in reducing the risk that their products will support attacks.

Government Action

Establish a nimble regulatory body. The pace of AI development moves quickly, and a nimble regulatory body that can act and enforce quickly, as well as update certain enforcement criteria, is necessary. This could be an existing or a new body. This standards body or agency would have the power to approve or reject risk assessments, mitigations and audit results (as recommended in "Regulatory Action: AI Systems" above), process registrations, issue licences, and have the authority to block deployment or development of models. In the European Union, this is already in motion with the newly created AI Office. In the United States, the recently formed AI Safety Institute within the NIST seems to be the best candidate to take on this charge, if it can secure a sufficient budget. This May, at an AI safety summit hosted in Korea, a group of countries created a network of AI Safety Institutes or similarly named bodies, either already launched or in development in Australia, Canada, the European Union, France, Germany, Italy, Japan, Singapore, South Korea, the United Kingdom and the United States.

Support fact-checking organizations and civil society observers. Require generative AI developers to work with and provide direct support to fact-checking organizations and civil society groups (including the "trusted flaggers" defined by the DSA) to provide them with forensic software tools that can be used to investigate sophisticated or defined by the Digital Services Act) to provide them with forensic software tools that can be used to investigate sophisticated or complex cases of generative AI use and abuse, and to identify scaled variations of false content through fan outs. This would include a secured form of access to the latest detection systems). AI systems can, with great care, also be applied to the expansion and improvement of fact-checking itself, providing context in dynamic ways for misleading content.

Fund innovation in AI governance, auditing, fairness and detection. Countries and regions that enact rules like these have an opportunity to support innovation in critical fields of AI that will be needed to ensure that AI systems and deployments are executed ethically and in keeping with these regulations. This could come in the form of grants such as those described in the Executive Order on AI (sec. 5.2, 5.3).

Cooperate internationally. Without international cooperation, bilaterally at first, and eventually in the form of a treaty or new international agency, there will be a significant risk that these recommendations might be circumvented. There are many recent reasons to have hope for progress. China is actually already far ahead of the United States in implementing regulation (some good, some bad), and is already proposing opportunities for global AI governance. The Bletchley Declaration, whose 29 signatories include the home countries of the world's leading AI companies (United States, China, the United Kingdom, the United Arab Emirates, France, Germany), created a firm statement of shared values and carved a path forward for additional meetings of the group. The United Nations High-Level Advisory Body on Artificial Intelligence, formed in August 2023, presented interim recommendations in late 2023 and will be publishing a final report before the Summit of the Future in September 2024, with the potential to make valuable recommendations about international governance regimes. Additionally, the G7 Hiroshima AI Process has released a statement, a set of guiding principles, and a code of conduct for organizations developing advanced AI systems. None of these international efforts are close to a binding or enforceable agreement, but the fact that conversations are advancing as quickly as they are has been cause for optimism among concerned experts.

Democratize AI access with public infrastructure. A common concern cited about regulating AI is that it will limit the number of companies who can produce complex AI systems to a small handful, thereby entrenching oligopolistic business practices. There are many opportunities to democratize access to AI, however, that don't necessarily require relying on unsecured AI systems. One is through the creation of public AI infrastructure that allows for the creation of powerful secured AI models without necessitating access to capital from for-profit companies, as has been a challenge for ethically minded AI companies. The US National AI Research Resource could be a good first step in this direction, as long as it is developed cautiously. Another option is to adopt an anti-monopoly approach to governing AI, which could put limits on vertical integration by excluding would-be competitors from accessing hardware, cloud services or model APIs.

Promoting Innovation and the Regulatory First-Mover Advantage

Many people ask if regulations such as those I've proposed here will stifle innovation in the jurisdictions where they are enacted. I (and others) believe that they could well have the opposite effect, with leadership in this domain offering numerous benefits to regulatory first movers.

The two leading AI start-ups in the United States, OpenAI and Anthropic, have distinguished themselves with an intense internal focus on building AI safely and with the interests of society at their core. OpenAI began as a non-profit organization. Though its value has been watered down over time, perhaps especially evident in the case of the recent firing and rehiring of its CEO, that structure still signals that the company may be different from the tech giants that came before it. The founders of Anthropic (which received from Amazon an investment of $4 billion) left OpenAI because they wanted to be even more focused on the safety of their AI systems. The CEOs of both companies have called openly for regulation of AI, including versions of many of my recommendations above, even though it stands to complicate their own work in the field.

Both companies also came to the conclusion that making their models open source was not in line with their principled approach to the field. A cynic could say that this decision was driven by the companies' interest in controlling their models to derive profits, but regardless, the decision proves that it's a fallacy that innovation will be stifled without highly capable and dangerous open-source models available in the market.

Innovation can take many forms, including competing for funding and talent by demonstrating high levels of ethics and social responsibility, a tactic that led a group of "impact investors" to purchase shares in the company earlier this year. By setting rules that become the gold standard for ethical AI, including by following the recommendations above, the political leaders of early-adopting jurisdictions could also distinguish themselves and their polities as forward-thinking actors who understand the long-term ethical impacts of these technologies. Regulation also serves the purpose of rebalancing the playing field in favour of ethically focused companies. As I argue in the third recommendation in the "Government Action" section above, government funding for innovative start-ups working on AI governance, auditing, fairness and detection will position jurisdictions that are first to regulate as leaders in these fields. I hope that we'll see a future in which open-source AI systems flourish, but on the condition we can build the resilience in our distribution channels and other security systems to contain the significant risks that they pose.

> Innovation can take many forms….By setting rules that become the gold standard for ethical AI…the political leaders of early-adopting jurisdictions could also distinguish themselves and their polities as forward-thinking actors who understand the long-term ethical impacts of these technologies.

One useful analogy is the move toward organic food labelling. California was the first state in the United States to pass a true organic certification law in 1979. This meant that California organic farmers actually had it harder than other states for awhile, because they had a rigorous certification process to go through before they could label their food as organic. When national organic standards arrived in 1990, California organic farmers had an advantage, given their experience. Today, California produces more organic products than any other state in absolute terms, and is ranked fourth out of 50 states in relative acreage of organic farms.

Another useful example is seat belts. An op-ed by four former prominent US public servants draws the analogy well: "It took years for federal investigations and finally regulation to require the installation of seat belts, and eventually, new technologies emerged like airbag and automatic brakes. Those technological safeguards have saved countless lives. In their current form, AI technologies are dangerous at any speed."

The "first-mover advantage" is a common business concept, but it can also apply to the advancement of regulatory landscapes. The European Union is already being lauded for its DSA and Digital Markets Act, which are positioned to become de facto global standards. Pending the resolution of issues related to the regulation of foundation models, the European Union appears likely to be the first democracy in the world to enact major AI legislation with the EU AI Act. A strong version of this legislation will position the region's AI marketplace to be a model for the world and, via the "Brussels effect," have a strong influence on how companies behave around the world. If regulation spurs researchers to make innovations that reckon with open-source safety concerns early on, such as self-destructing model weights that prevent harmful fine-tuning, these regulatory changes could mean far more democratic access to AI in the future.

Conclusion

"I think how we regulate open-source AI is THE most important unresolved issue in the immediate term," Gary Marcus, a cognitive scientist, entrepreneur and professor emeritus at New York University, told me in a recent email exchange.

I agree. These recommendations are only a start at trying to resolve it. As one of my reviewers of an early draft of this essay noted, "These are hard, but maybe that's the point." Many of the proposed regulations here are "hard" from both a technical and a political perspective. They will be initially costly, at least transactionally, to implement, and they may require that some regulators make decisions that could leave certain powerful lobbyists and developers unhappy.

Unfortunately, given the misaligned incentives in the current AI and information ecosystems, and the vulnerability of our democratic institutions, as well as heightened geopolitical tensions, it is unlikely that industry itself will take the necessary actions quickly enough unless forced to do so. But unless such actions are taken, companies producing unsecured AI will bring in billions of dollars in investments and profits, while pushing the risks onto all of us.

Acknowledgements

The views expressed here are exclusively those of the author, but I owe a debt of gratitude to the following people and organizations for their support, feedback and conversations that made this article possible: Lea Shanley and Michael Mahoney of the International Computer Science Institute; Paul Samson, Aaron Shull, Dianna English and Emma Monteiro of the Centre for International Governance Innovation; Camille Crittenden and Brandie Nonnecke of CITRIS and the Banatao Institute; Jessica Newman, Hany Farid and Stuart Russell of the University of California, Berkeley; Samidh Chakrabarti of Stanford University; Alexis Crews, Tom Cunningham, Eric Davis and Theodora Skeadas of the Integrity Institute; Arushi Saxena of DynamoFL and the Integrity Institute; Camille Carlton of the Center for Humane Technology;

Sam Gregory of WITNESS; Jeffrey Ladish of Palisade Research; Aviv Ovadya of the Centre for the Governance of AI; Yolanda Lannquist of The Future Society; Chris Riley of the Annenberg Public Policy Center at the University of Pennsylvania; David Morar of New America's Open Technology Institute; independent members of my research team, Owen Doyle, Zoë Brammer, Diane Chang and Jackson Eilers; my student research team — Milad Brown, Ashley Chan, Ruiyi Chen, Maddy Cooper, Anish Ganga, Daniel Jang, Parth Shinde, Amanda Simeon and Vinaya Sivakumar — at the Haas School of Business at the University of California, Berkeley; and the John S. and James L. Knight Foundation.

[1] In an earlier version of this essay, I fashioned the acronym DUMWAM as a shorthand for *d*ual-*u*se foundation *m*odels with *w*idely *a*vailable *m*odel weights. I suggest to readers that it can be remembered by imagining the feeling and sound of banging one's head on a keyboard while thinking about what a bad idea it is to offer unfettered access to dangerous AI systems to anyone in the world.

[2] It is likely that techniques for training new models will become more efficient over time and that today's regulatory thresholds will not necessarily hold. That points to the weakness of training compute thresholds as a proxy metric for model riskiness, and, at least in the European Union, this weakness is mitigated by the newly created AI Office's ability to designate models as posing "systemic risk" and thereby subject to greater regulatory burdens based on other qualitative assessments of model capabilities, such that the thresholds for applicability of the AI Act to general-purpose AI models can be adjusted in the future. It will be critically important for the European Union's AI Office to closely monitor developments in model technologies so that thresholds can be adjusted before highly efficient high-risk models are released in an unsecured manner.

The opinions expressed in this article/multimedia are those of the author(s) and do not necessarily reflect the views of CIGI or its Board of Directors.

CIGI

ARTIFICIAL INTELLIGENCE (247)
COMPETITION (41)
CYBERSECURITY (6)
DATA GOVERNANCE (286)
DEMOCRACY (442)
DIGITAL ECONOMY (93)
DIGITAL GOVERNANCE (8)
DIGITAL RIGHTS (4)
FINANCIAL GOVERNANCE (637)
FOREIGN INTERFERENCE (4)
FREEDOM OF THOUGHT (20)
G20/G7 (315)
GENDER (119)
GEOPOLITICS (442)
GLOBAL COOPERATION (10)
HUMAN RIGHTS (134)
INTELLECTUAL PROPERTY (191)
MULTILATERAL INSTITUTIONS (184)
NATIONAL SECURITY (473)
PLATFORM GOVERNANCE (674)
PRIVACY (181)
QUANTUM TECHNOLOGY (2)
SPACE GOVERNANCE (58)
SURVEILLANCE (174)
TRADE (746)
TRANSFORMATIVE TECHNOLOGIES (656)

Contact Careers Directions Privacy Notice
Media Relations

Safeguards for Using Artificial Intelligence in Election Administration | Brennan Center for Justice 16/09/2024, 16:53

BRENNAN CENTER FOR JUSTICE

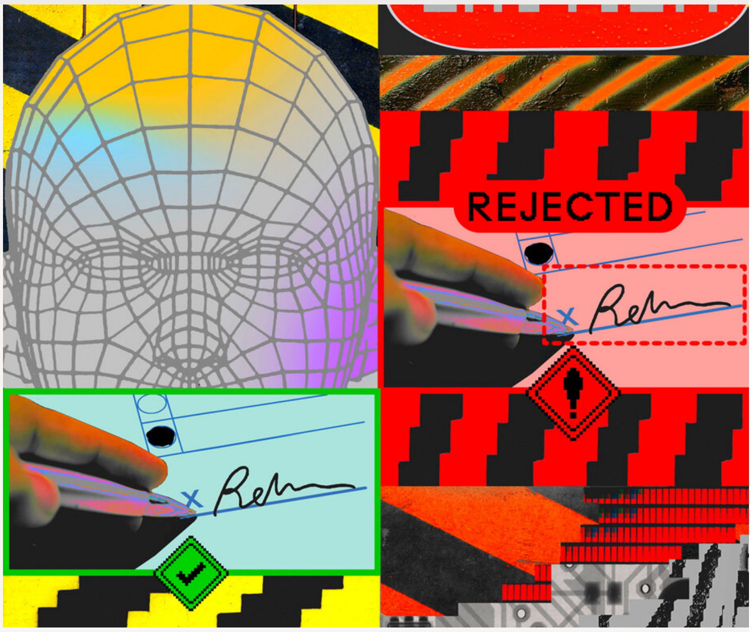
Issues Our Work Election 2024 Experts Get Involved About

EXPLORE Our Work

Home // Our Work // Research & Reports // Safeguards for Using Artificial Intelligence in Election Administration

BC Issues Our Work Election 2024 Experts Get Involved About

Safeguards for Using Artificial Intelligence in Election Administration



Adequate transparency and oversight can ensure AI tools in election offices are helpful and not harmful.

Edgardo Cortés

Lawrence Norden

Heather Frase

Mia Hoffmann

PUBLISHED: December 12, 2023

As artificial intelligence tools become cheaper and more widely available, government agencies and private companies are rapidly deploying them to perform basic functions and increase productivity. Indeed, by one

View
and D
series

estimate, global spending on artificial intelligence, including software, hardware, and services, will reach $154 billion this year, and more than double that by 2026. As in other **government** and **private-sector** offices, election officials around the country already use AI to perform important but limited functions effectively. Most election offices, facing budget and staff constraints, will undoubtedly face substantial pressure to expand the use of AI to improve efficiency and service to voters, particularly as the rest of the world adopts this technology more widely.

In the course of writing this resource, we spoke with several election officials who are currently using or considering how to integrate AI into their work. While a number of election officials were excited about the ways in which new AI capabilities could improve the functioning of their offices, most expressed concern that they didn't have the proper tools to determine whether and how to incorporate these new technologies safely. They have good reason to worry. Countless examples of faulty AI deployment in recent years illustrate how AI systems can **exacerbate bias**, **"hallucinate" false information**, and otherwise **make mistakes** that human supervisors fail to notice.

Any office that works with AI should ensure that it does so with appropriate attention to quality, transparency, and consistency. These standards are especially vital for election offices, where accuracy and public trust are essential to preserving the health of our democracy and protecting the right to vote. In this resource, we examine how AI is already being used in election offices and how that use could evolve as the technology advances and becomes more widely available. We also offer election officials a set of preliminary recommendations for implementing safeguards for any deployed or planned AI systems ahead of the 2024 vote. A checklist summarizing these recommendations appears at the end of this resource.

As AI adoption expands across the election administration space, federal and state governments must develop certification standards and monitoring regimes for its use both in election offices and by vendors. President Joe Biden's October 2023 **Executive Order** on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence marks a pivotal first step, as it requires federal regulators to develop guidelines for AI use by critical infrastructure owners and operators (a designation that has included owners of **election infrastructure** since 2017) by late spring 2024.

Under its recently announced **artificial intelligence roadmap**, CISA will provide guidance for secure and resilient AI development and deployment, alongside recommendations for mitigating AI-enabled threats to critical infrastructure. But this is only a start. It remains unclear how far the development of these guidelines will go and what election systems they will cover. The recommendations in this resource are meant to assist election officials as they determine whether and how to integrate and use AI in election administration, whether before or after new federal guidelines are published next year.

Current and Potential Future Uses of AI in Election Administration

Artificial intelligence is an umbrella term for computer systems that use data, algorithms, and computing power to perform a range of tasks that historically required human intelligence, such as recognizing speech, identifying patterns in data, and making predictions. Today, AI tools **make movie recommendations**, **power facial recognition**, and even **drive cars**. Generative models — a subset of AI capable of producing realistic text, images, video, and audio in response to user prompts — have garnered widespread public attention since the release of ChatGPT in 2022. Although both generative and non-generative AI can behave unpredictably in new situations, understanding (and predicting) the former's at times unexpected behavior is often more difficult, mainly because generative AI is typically built using more parameters and vastly more data. Election officials need to have safeguards in place when using both generative and non-generative AI.

Organizations in the private and public sectors **frequently use** AI for data management functions such as identifying duplicate records, and election offices are no exception. The **Electronic Registration Information Center** (ERIC), a multistate voter list maintenance effort, is one example of non-generative AI use in election administration. ERIC's **software** employs AI to support voter roll management by searching for duplicate entries across many data sets. ERIC validates possible matches by conducting a human review prior to sending matching data to member states. Once states receive data from ERIC about possible matches, they process them according to their respective state list maintenance rules per requirements established by the National Voter Registration Act.

ERIC's ability to identify potential matches using various data sets is considerably more advanced than earlier systems, such as the Interstate Voter Registration Crosscheck program, which utilized rudimentary data matching with limited date fields and led to high numbers of **false positive identifications**. In a typical example, a false positive would incorrectly identify two distinct voters on different voter rolls (or other databases, like the Social Security Administration's Limited Access Death Master File) as being the same person. False positives increase the workload and cost associated with list maintenance processes. More importantly, they can harm eligible voters and lead to disenfranchisement.

Election offices also use non-generative AI to match mail-in ballot signatures, historically a time- and labor-intensive task. Although the specific technology varies by vendor, ballots are generally fed through a scanner that captures an image of the signature and compares it with a signature already on file. Signatures that the software can match are processed for counting; signatures that cannot immediately be verified are set aside for **human review** and **further analysis**. This automation allows election offices to focus on researching and validating a smaller set of signatures before processing ballots for counting, thereby saving time and resources. Election offices should account for potential bias by incorporating examples during training of signatures the matching software is more likely to not validate, such as first-time or elderly voters, and include suggestions for appropriate human review of those ballots.

An increasing number of election offices are using **AI chatbots** to answer basic voter questions as well. (Chatbots can be either generative or non-generative, though we are not aware of any election offices yet exploring generative AI for this purpose.) Chatbots like those used by the **New York State Department of Motor Vehicles** and the **California secretary of state** can provide information outside of normal office hours and free up staff to deal with more complex issues. This technology also helps voters navigate election websites, providing important information like polling place times and locations and answers to frequently asked questions.

Non-generative AI chatbots typically produce pre-vetted responses or use a form of natural language processing similar to Amazon's Alexa virtual assistant technology. While both non-generative and generative AI chatbots risk unreliable or biased results, the latter would entail the added burden of verifying the accuracy of synthesized content. Any generative AI chatbot that an election office seeks to employ would require sufficient development and testing time to ensure, for example, that it accurately answers voter questions but appropriately redirects questions complicated or consequential enough to need staff attention.

In addition, some election offices are considering using AI to help create and translate voting materials. Here too, election officials would need to mitigate AI's potential to provide incorrect or unreliable information. At the very least, they would need to implement strong internal controls to ensure that all products are reviewed by the appropriate staff before release and corrected where needed — particularly when AI tools assist in translating election materials. This level of scrutiny is crucial given the nuance related to voter registration and voting requirements.

As the technology evolves, election officials will likely find myriad new ways for AI to assist in election administration. AI could serve as an extra proofreader for election materials and an extra set of "eyes" to ensure that ballots and other materials comply with the law and best design practices, or that materials are correctly translated. AI could also be used to identify new polling place locations based on traffic patterns, travel time for voters assigned to a polling place, public transportation routes, parking availability, and other factors. AI systems could even be used to analyze postelection data to improve future elections, identifying patterns in provisional voting, voter registration application rejections, and reasons for rejecting absentee or mail-in ballots. The possibilities are innumerable. Yet all of them will require similar quality-control standards and safeguards.

Risks Associated with AI Use in Elections

The risks associated with integrating AI into election processes are considerable. Some of them are inherent to AI technology, while others arise from human-machine interactions. A particular risk lies in the inevitable differences between an AI system's training data and the data it uses when deployed. As a result of this data disparity, AI typically performs worse in operation than on the benchmarks or performance results obtained during testing and presented by vendors.

As other resources in this series have explored, AI trained using past data and past decisions also risks perpetuating **biases** inculcated in those decisions. This **all-too-common phenomenon** can systematically disenfranchise groups of voters if the historical bias is not mitigated during an AI tool's development and implementation. Furthermore, generative AI chatbots can suffer from "**hallucinations**" — delivering incorrect information presented as fact — which risks providing voters with wrong information. Spotting incorrect or hallucinated information is difficult in contexts where election office staff cannot oversee the chatbot's responses, such as real-time interactions with voters. As such, using generative AI for election administration functions is often high risk.

Election office staff should review AI tools' decisions and outputs whether those systems use generative or non-generative AI. Such reviews will require training to mitigate automation bias (the tendency to over-rely on automated decisions because they appear objective and accurate) and confirmation bias (the predisposition to favor information when it confirms existing beliefs or values). Insufficient transparency about where and how AI tools are used in processes that affect voters' ability to cast their ballots compounds these risks by preventing external scrutiny. Internal system evaluations are generally protected information, and independent external analyses are often impossible because election offices cannot share data with third parties.

Two other concerns are worth mentioning, although their discussion and mitigation are beyond the scope of this resource. First, as an **earlier installment in this series** discusses in more detail, election officials will need to take steps to prevent attacks against AI systems integrated into election administration. Second, the use of AI in certain contexts has implications for constituents' privacy. Using voters' data — usually without their knowledge or consent — to train or fine-tune AI tools provided by third parties raises serious concerns around data protection, ownership, and control, especially when records contain sensitive information like names, birth dates, addresses, and signatures. Moreover, some AI uses touch on the principle of anonymity in elections. The use of biometric data for voter registration or identity verification creates a log of identified voter behavior that could threaten voter anonymity if paired with local ballot counts or similar data.

The absence of regulations or governmental guidance on safe AI implementation amplifies these risks. As resource-constrained election officials look to the benefits of AI, they must also assess the risks and potential downsides of adopting these new technologies. In doing so, they must recognize that incorporating AI tools in election administration without appropriate risk mitigation measures and transparency could compromise voter confidence heading into the 2024 election cycle and beyond.

Recommendation for Election Offices: AI CPR (Choose, Plan, and Review)

In deciding whether to employ AI, election officials should implement and follow a transparent selection process to **choose** the specific AI tool for any given election administration task. If and when they do choose a particular AI system, officials need to carefully **plan** that system's integration into their workflows and processes. Part of that planning must include identifying and preparing for problems that may surface as the system is incorporated. They must also be able to shift resources as needed. Finally, they must establish thorough **review** procedures to ensure that the output of any AI tool deployed in an election office is assessed by staff for accuracy, quality, transparency, and consistency. Below, we describe important considerations at each of these three stages.

Choose AI Systems with Caution

Opt for the Simplest Choice

In choosing any system (AI-based or not) for use in election administration, all else being equal, we recommend that election officials choose the simplest tool possible. When it comes to AI, though simpler AI algorithms may be less refined than more complex ones, they are also easier to understand and explain, and they allow for greater transparency. Should questions or anomalies arise, determining answers and solutions will be easier with a simple AI model than an elaborate one. The most complicated AI systems currently available belong to the latest class of generative AI, followed by non-generative **neural networks** and other **deep learning** algorithms. Basic machine learning models like **clustering algorithms** or **decision trees** are among the simplest AI tools available today.

A useful practice to facilitate choosing the simplest possible system is for election officials to narrowly define the tasks that the AI will perform and identify a list of requirements. Requirements can range from price considerations or necessary IT and data infrastructure to the need for additional functionalities or minimum performance levels reflecting the risk level that election officials are willing to accept for a given task. Establishing these parameters ahead of the selection process will help both to ensure transparency around the criteria used for assessing proposals and to prevent “scope creep” when vendors demonstrate capabilities of more advanced systems.

Plan for Human Involvement

If an AI tool could result in someone being removed from the voter rolls, being denied the ability to cast a ballot, or not having their vote counted, then election officials should choose a system that requires human involvement in making final decisions. Human involvement helps to safeguard against AI performance irregularities and bias. Most jurisdictions have processes that require additional review before rejecting vote-by-mail or absentee ballots for a non-signature match. Generally, this review involves **bipartisan teams** that must reach a consensus before rejecting a ballot. Twenty-four states currently have **processes** in place that require election offices to notify voters should questions arise about their signature and to provide them the opportunity to respond and cure the issue. Such processes are vital to ensure that AI systems do not inadvertently prevent voters from having their votes counted. The planning and review stages outlined below will need to factor in this human involvement.

Anticipate Performance Disparities, Reliability Issues, and Output Variability

When selecting an AI tool, election officials should assume that the system will not perform as effectively as vendor metrics claim. In developing and training AI models, vendors inevitably use a training data set that is different than the data the AI is fed during actual use. This scenario frequently leads to degraded performance in real-world applications. Additionally, because of data eccentricities, data collection processes, and population differences, the same AI tool’s performance can vary substantially between districts and between population groups within the same district. As a result, AI tools are likely to perform less effectively on actual constituents’ data compared with benchmarks or results presented by vendors.

In particular, name-matching algorithm performance **has been shown** to vary across racial groups, with the lowest accuracy found among Asian names. A **study of voter list maintenance errors** in Wisconsin also revealed that members of minority groups, especially Hispanic and Black people, were more than twice as likely to be inaccurately flagged as potentially ineligible to vote than white people. Similarly, AI-powered signature-matching achieves between **74 and 96 percent** accuracy in controlled conditions, whereas in practice, ballots from young and first-time mail-in voters, elderly voters, voters with disabilities, and nonwhite voters **are more likely to be rejected**. Unrepresentative training data coupled with low-quality signature images, often captured using DMV signature pads, to match against lowers the effectiveness of signature-matching software.

Implementing this technology for voter roll management thus raises major concerns. One mitigation strategy that election officials can utilize in choosing AI systems is to require vendors to use a data set provided by the election office for any demonstrations during the request for proposal and contracting process. This approach can provide further insight into system performance. Importantly, election officials should ensure that only publicly available data is used or that potential vendors are required to destroy the data after the selection process has concluded and not retain or share the data for other purposes.

Although a general strength of generative AI is its ability to respond to unanticipated or unusual requests or questions, election officials must bear in mind that current generative AI tools often suffer from reliability issues. Generative AI chatbots may produce different responses to the same request, and they regularly produce incorrect or hallucinated replies. In addition, the underlying language models are **frequently fine-tuned** and updated, which in turn affects the behavior of systems built on them.

Finally, when deciding whether to use a generative tool, election officials must consider whether variations in content and quality are acceptable. For most election-related tasks, variability that could result in an office propagating misinformation is not a tolerable outcome. As such, election offices should not adopt generative AI systems for critical functions without national or **state standards** in place to guide appropriate uses and provide baseline assurances of system reliability and safety.

Plan for AI Use — and for Potential Problems

Election offices should devise both internally and externally focused implementation plans for any AI system they seek to incorporate. Internally, election officials should consider staffing and training needs, prepare process and workflow reorganizations, and assign oversight responsibilities. Externally, they should inform constituents about the AI's purpose and functionality and connect with other offices employing the same tool. Most importantly, officials should develop contingency plans to handle potential failures in deployed systems.

Develop Staff Training

Before deploying an AI tool, election officials must consider the training needs of their staff. While the following list is not all-inclusive, training should impart a high-level grasp of the AI system. Staff must understand the exact tasks the AI performs, its step-by-step processes, the underlying data utilized, and its expected performance. For instance, rather than thinking of a signature verification system simplistically as a time-saving bot that can verify mail-in ballots, staff should see it as a software tool that attempts to match the signature on a ballot to an image on record using a computer vision algorithm, and that it does so with an average accuracy rate of 85 percent.

At a minimum, staff training should cover

- familiarization with the user interface;
- common risks and issues associated with data and AI (such as those described above), how they could occur in the context of the office's constituency and its election administration work, the system's limitations, and how to address problems;
- internal processes for flagging issues with the AI and accountability guidelines in case of failure or errors; and
- requirements for — and the importance of — human involvement in decisions that directly implicate voter rolls, vote casting, and vote counting, including techniques for mitigating bias.

Prioritize Transparency

Constituents have a right to know about AI systems involved in election administration. Election officials must be transparent about when, for what, and how AI tools will be used. Before deployment, election offices should work with the AI developers to prepare and publish documentation in nontechnical language. These documents should describe the system's functionality and how it will be used, what is known about its performance, limitations, and issues, and any measures taken to mitigate risk for the particular election administration task for which it will be deployed. Constituents should have opportunities to discuss questions and concerns with officials to build trust in the technology and in election administrators' oversight capabilities. The need for transparency and documentation should be outlined in the request for proposal process and included in vendor contracts so that relevant information cannot be hidden from public view under the guise of proprietary information.

Prepare Contingency Plans

Election officials must have contingency plans in place before incorporating AI technology. AI contingency plans must include appropriate preparations to manage any potential failures in a deployed AI system. First and foremost, election offices must be able to disable an AI tool without impairing any election process — a fundamental best practice for using AI in a safe and trustworthy manner. AI tools should not be integrated into election processes in a way that makes it impossible to remove them if necessary.

Contingency plans must identify the conditions under which an AI tool will be turned off along with which staff members are authorized to make such a determination. Election offices must ensure that staff are aware of these conditions and are trained to identify them and to report issues, flaws, and problems to the responsible officials. Offices must also have a strategy in place for how to proceed if the use of AI is halted. This strategy should include identifying additional personnel or other resources that can be redirected to carry out certain tasks to ensure their timely completion.

Seek Other Users' Input

The experiences of other users can help inform election offices newly adopting AI tools. Election officials should ask potential vendors for lists of other offices currently using their systems during the request for proposal process and should reach out to those offices when evaluating bids. Many AI tools are relatively new, so users are often the ones who discover their strengths and weaknesses. Learning from other users' experiences in the elections space will be valuable for shaping effective training and implementation and for identifying resource needs and contingencies.

Review AI Processes and Performance

System reviews are an essential best practice when using AI tools. The extent and frequency of reviews will vary depending on the gravity of the election administration task at hand and the risk associated with it. Low-risk or low-impact applications (for example, an AI system used to check whether ballots comply with best design practices) may only need a process for getting user or voter feedback and a periodic review of the AI's performance. However, systems that help decide if someone gets to vote or if a vote is counted need more frequent and direct human oversight.

Institute Straightforward Review Processes

Election officials should establish clear processes for collecting, assessing, and resolving issues identified by both internal and external stakeholders and for reviewing AI system performance. These processes should include soliciting staff and constituent feedback, monitoring use and output logs, tracking issues, and surveying help desk tickets.

Audits of issues and performance should occur before and after elections. Pre-election reviews are paramount to safeguard voting rights and identify if an AI's contingency plan needs to be implemented. Postelection reviews will help improve future use and should assess all processes that AI touched, including evaluations of performance across demographic groups to reveal any potential biases. These reviews present an opportunity for election officials to work with federal partners on meaningful assessment tools for deployed AI systems, much as federal agency assessment tools exist for reviewing polling place accessibility and election office cybersecurity.

Ensure Human Involvement in Final Decisions That Affect Voters

People are the most critical factor in the successful deployment of AI systems in election offices. Decisions that directly affect an individual's right to vote and ability to cast a ballot cannot be left solely to AI — trained individuals must be involved in reviewing consequential decisions based on AI analysis and AI-produced information. Regarding AI-assisted translations of election materials, if staff are not fluent in all relevant languages, officials should consider partnering with trusted local community groups to ensure translation accuracy. When incorporating AI technology into election administration processes, officials should also consider that these additional trainings and reviews may add or shift costs to different times in the election calendar.

Establish Challenge and Redress Procedures

Election officials must provide a process for challenging and reviewing AI-assisted decisions. Voters harmed by decisions made based on AI should be able to appeal and request reviews of those decisions. How these processes should occur will vary from jurisdiction to jurisdiction; existing state and local procedures for review and remedy should be assessed for appropriateness in light of AI-assisted decision-making and amended where necessary. For instance, what if a voter **is directed to the wrong polling place** by an agency chatbot and forced to cast a provisional ballot as a result? That voter needs a way to make sure that their ballot is counted nonetheless, especially because the action was prompted by inaccurate information provided by the election office. This is to say nothing of errors generated by AI-based signature-matching software, for example, or any number of other conceivable AI errors.

Enacting clear and accessible processes for constituents to challenge AI-driven decisions — processes that initiate a swift human review and an appropriate resolution — is imperative both to provide an added layer of protection to voting rights and to continually evaluate the performance of AI systems employed in election administration.

Conclusion

Increasing AI integration in election administration presents many opportunities for improving the voter experience and increasing the efficiency of election offices, but it also introduces new risks to electoral integrity and to the fundamental democratic principle of free and fair elections. While the capabilities of AI products have grown rapidly over the past few years, many of their inherent problems remain unsolved. AI systems often behave in unreliable ways and perform more effectively for some demographics than for others. Many AI tools are inscrutable in their decision-making processes, preventing meaningful human oversight. These issues, left unchecked, can erode citizens' basic constitutional right to vote. The AI CPR recommendations laid out in this resource are intended to serve as a road map for mitigating these risks. Election officials seeking to use AI should follow this road map as they adopt this exciting but fraught technology.

Appendix: AI CPR Checklist

We cannot expect local election offices to create safeguards for the use of AI technology by themselves any more than we can expect them to defend themselves single-handedly against cyberattacks by nation-states. This overview of AI CPR (Choose, Plan, and Review) is a non-exhaustive list of considerations to help election officials determine whether and how to incorporate AI in election administration.

Choose with caution.

- Choose the simplest systems (including non-AI systems) that meet your needs.
- If you choose an AI tool that implicates the voter roll, vote casting, or vote counting, choose one that requires human involvement for decisions.
- Assume that the system will perform worse in operation relative to vendor metrics and decide if that is acceptable for the application at hand.
- Avoid adopting generative AI systems for critical tasks without national or state standards in place to guide appropriate uses.

Plan for use and problems.

- Establish processes for transparency (toward constituents and stakeholders alike) around when, for what, and how AI systems will be used.
    - Prepare documentation and publications in nontechnical language.
    - Provide information about the purpose, functionality, and oversight of AI tools.
    - Inform constituents about opportunities to raise issues and contest AI-supported decisions.

- Have a contingency plan in place before deploying AI technology.
- Identify the conditions that warrant suspending AI systems and what will happen in their absence.
- Ensure sufficient training of staff. At a minimum, training should cover
    - a high-level understanding of and familiarity with the AI system;
    - effective and safe use of the AI tool;

  - common risks and issues, along with how they may occur during use;
  - internal processes for flagging issues with the system and accountability guidelines in case of failure or errors; and
  - how staff will make final determinations if an AI tool is used to support decisions that directly implicate voter rolls, vote casting, or vote counting.

- Ask potential vendors for lists of other election offices currently using systems under consideration. Contact those offices for advice and lessons learned.

Review processes and performance.

- Implement review processes and create infrastructure for safe use of AI tools. At a minimum:
  - Collect the information needed for reviewing the performance and integration of AI systems. This information can include staff feedback, use and output logs, constituent feedback, issue tracking, and help desk tickets.
  - Ensure human review of decisions that directly implicate voter rolls, vote casting, and vote counting.
  - Provide means for constituents harmed by AI systems to request that decisions be reviewed and changed if necessary.
  - Gather and resolve issues found both internally and externally.

- Postelection:
  - Review and refine processes based on lessons learned and constituent feedback.
  - Conduct performance evaluations of processes that incorporated AI systems. In particular, assess whether acceptance or false rejection rates varied for demographic groups.

Heather Frase and Mia Hoffmann are, respectively, a senior fellow and a research fellow at Georgetown's Center for Security and Emerging Technology. Edgardo Cortés and Lawrence Norden are, respectively, an election security advisor and the senior director of the Brennan Center Elections and Government Program.

More from the *AI and Democracy* series ▶

EXPERT BRIEF

Artificial Intelligence, Participatory Democracy, and Responsive Government

Government must implement safeguards against malicious uses of AI that could misrepresent public opinion and distort policymaking.

Mekela Panditharatne, **Daniel I. Weiner**, **Douglas Kriner** // November 3, 2023

READ MORE

EXPERT BRIEF

Generative AI in Political Advertising

Political campaigns should be watchful of the risks and opportunities of using generative AI to engage with voters.

Christina LaChapelle, **Catherine Tucker** // November 28, 2023

READ MORE

EXPERT BRIEF

Deepfakes, Elections, and Shrinking the Liar's Dividend

Heightened public awareness of the power of generative AI could give politicians an incentive to lie about the authenticity of real content.

Josh A. Goldstein, **Andrew Lohn** // January 23, 2024

READ MORE