FOUNDATION FOR AMERICAN INNOVATION

June 13, 2023

Geoffrey Cain
Senior Fellow, Foundation for American Innovation
Written Testimony for U.S. Senate Committee on the Judiciary
Subcommittee on Human Rights and the Law

"America, the Vanguard of Democracy, Must Stand Up to China's AI Totalitarianism"

Chairman Ossoff, Ranking Member Blackburn, and members of the Subcommittee:

Thank you for the opportunity to testify today. My testimony has two purposes:

1. First, to outline how China has created the world's most sophisticated and terrifying surveillance state using novel artificial intelligence (AI) technologies, and how American business elites helped make this happen.

2. Second, to suggest ways that the US can defend the use of AI with respect to democracy and human rights, to ensure that the Chinese Communist Party (CCP) cannot advance its malign global agenda with AI tools.

With AI, America's elites have learned little about the perils of engaging with China's one-party authoritarian state

On Friday, OpenAI CEO Sam Altman dialed into the annual conference at the Beijing Academy of Artificial Intelligence, three weeks after he testified before another subcommittee here at the Senate Judiciary Committee. He called on the People's Republic of China—a one-party authoritarian state that has used AI to carry out genocide against an ethnic minority—to help shape global AI safety guardrails. "With the emergence of increasingly powerful AI systems," he said, "the stakes for global cooperation have never been higher."[1]

To anyone who's lived in China, this was a curious and mind-boggling call to action. The Chinese Communist Party (CCP) has engineered a vast AI-powered surveillance system literally called "Sky Net." It runs AI-powered "alarms" that notify the police and intelligence services when someone unfurls a banner,[2] when a foreign journalist is traveling to certain parts of the country,[3] and when someone from an ethnic minority is

---

[1] Sarah Zheng, "OpenAI's CEO Calls on China to Help Shape AI Safety Guidelines," Bloomberg Technology, June 9, 2023, https://www.bloomberg.com/news/articles/2023-06-10/openai-s-ceo-altman-calls-on-china-to-help-shape-ai-safety-guidelines.
[2] Gulchehra Hoja, "In China, AI Cameras alert police when a banner is unfurled," *Radio Free Asia*, June 5, 2023, https://www.rfa.org/english/news/china/surveillance-06052023142155.html.
[3] Jimmy Quinn, "'Total Security State': Shanghai Intensifies Surveillance of Foreign Journalists Who Go to Xinjiang," *National Review*, May 2, 2023, https://www.nationalreview.com/corner/total-security-state-shanghai-intensifies-surveillance-of-foreign-journalists-who-go-to-xinjiang/.

present.[4] The government accuses entire groups, such as Muslim Uyghurs, of posing a terrorist threat, and relentlessly persecutes them with the use of AI tools.

It sounds like a dystopian science fiction story—think *1984* or *Minority Report*—but the CCP's AI totalitarianism has become a fact of daily life for the more than 1.4 billion people in China. In fact, the Chinese technologists who spoke at the same conference as Mr. Altman were some of the very people who built this monstrosity. They were executives at iFlyTek and Huawei, two AI giants and are heavily sanctioned by the US government for their involvement in human rights abuses.[5] If Mr. Altman plans on cooperating with China's AI developers, he better figure out who he's working with.

I've witnessed the results of their work firsthand. As an investigative journalist formerly in China, I was among the first people to document and expose the horrific surveillance state that oppressed the Uyghur population in the far western region of Xinjiang. Since 2017, the atrocity has morphed into the largest internment of ethnic minorities since the Holocaust, which the US State Department calls a genocide.[6]

Chinese authorities have hauled away 1.8 million people to concentration camps—about one-tenth of the ethnic minority population in Xinjiang—and have forced many of them into slave labor.[7] Because they have read too many books or have been caught praying, they have been declared enemies of the state, despite not being formally charged with any crime. This was all with the help of the AI surveillance system that scooped up data from facial recognition, voice recognition, and a network of police cameras covering every possible square inch of the region. Party authorities told Uyghurs they wanted to "cleanse" their minds of what they called "ideological viruses."

In December 2017, I was kicked out of China while researching my book, *The Perfect Police State: An Undercover Odyssey into China's Terrifying Surveillance Dystopia of the Future.* Ever since then, the AI-fueled police state has expanded to alarming levels. In 2018, I moved to Turkey and, for three years, tracked down former intelligence officers from China's Ministry of State Security, the powerful and secretive intelligence body. They had helped set up the AI surveillance systems in Xinjiang, were targeted by those same systems because they were Uyghurs, and then defected to safety.

---

[4] Drew Harwell and Eva Dou, "Huawei tested AI software that could recognize Uighur minorities and alert police, report says," *Washington Post*, December 8, 2020, https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/.

[5] Karen Hao, "Open AI CEO Calls for Collaboration with China to Counter AI Risks," *The Wall Street Journal*, June 10, 2023, https://www.wsj.com/articles/openai-ceo-calls-for-collaboration-with-china-to-counter-ai-risks-eda903fe.

[6] Secretary of State Michael R. Pompeo, "Determination of the Secretary of State on Atrocities in Xinjiang," U.S. Department of State, January 19, 2021, https://2017-2021.state.gov/determination-of-the-secretary-of-state-on-atrocities-in-xinjiang/index.html.

[7] Adrian Zenz, "China's Own Documents Show Potentially Genocidal Sterilization Plans in Xinjiang," *Foreign Policy*, July 1, 2020, https://www.wsj.com/articles/openai-ceo-calls-for-collaboration-with-china-to-counter-ai-risks-eda903fe.

These intelligence officers drew detailed diagrams in my possession that showed the workings of these surveillance systems and how facial recognition and voice recognition technologies helped fuel them. What they revealed was alarming, but not surprising. The highest echelons of CCP leadership held centralized control over many AI surveillance systems, as well as direct lines of influence over Chinese mega-companies such as Huawei and ByteDance. With the help of these companies, China's government had been making a concerted, malicious effort to expand these surveillance capabilities all over the world.

The development of AI is at the heart of China's global ambitions

The surveillance state that began in Xinjiang was a taste of the horrific power of AI when placed in the wrong hands. "Advanced technology is the sharp weapon of the modern state," China's President Xi Jinping said in a 2013 speech.[8] In July 2017, China unveiled its National AI Development Plan, calling AI a "historic opportunity" and pledging to align developments in AI with the government's authoritarian values. China has declared its goal as becoming the world leader in AI by 2030.[9] The goal reflects the totalitarian ambitions of President Xi, who has led the efforts to clamp down Uyghurs, Tibetans, Mongolians, and religious and political dissidents of all stripes.

Since then, we've seen the expansion of China's technology companies, using AI and other novel developments, all over the world. Huawei, the heavily sanctioned telecommunications firm, has led efforts to establish global surveillance systems, usually under the guise of AI-powered "smart cities" designed to fight crime and regulate traffic, but that in reality have been used to equip governments with the tools to spy on political dissidents. In October 2022, the FBI arrested two Chinese nationals who stood accused of bribing an undercover FBI officer to obtain inside intelligence about an investigation into Huawei.[10]

Meanwhile, ByteDance, the $220 billion mega-firm that owns TikTok, stands accused by a whistleblower of running an in-house CCP Committee that had access to all the app's data, including data stored in the US, according to a court filing.[11] Other sanctioned, lesser-known firms, such as AI facial and voice recognition companies iFlyTek, SenseTime, and Megvii, have emerged as global billion-dollar unicorns with the backing of the Chinese state and the involvement of US venture capital funds.

---

[8] Chris Buckley and Paul Mozur, "What Keeps Xi Jinping Awake at Night," *New York Times*, May 11, 2018, https://www.nytimes.com/2018/05/11/world/asia/xi-jinping-china-national-security.html.
[9] Graham Webster, Roger Creemers, Elsa Kania, and Paul Triolo, "Full Translation: China's 'New Generation Artificial Intelligence Plan,'" August 1, 2017, https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/
[10] Glenn Thrush and David McCabe, "Justice Dept. Charges 2 Chinese Citizens With Spying for Huawei," *New York Times*, October 24, 2022, https://www.nytimes.com/2022/10/24/us/politics/justice-dept-huawei.html.
[11] Thomas Fuller and Sapna Maheshwari, "Ex-ByteDance Executive Accuses Company of 'Lawlessness,'" *New York Times*, May 12, 2023, https://www.nytimes.com/2023/05/12/technology/tiktok-bytedance-lawsuit-china.html.

This situation is proving hard to continue in the age of technological decoupling. This month, Sequoia Capital, the preeminent venture capital firm that originally invested in Apple and Facebook, announced that it was splitting off its Chinese arm into a separate company.[12] Sequoia's China business was core to helping build China's AI industry, with a reported $22 billion stake in ByteDance, to name one of many examples.[13] Sequoia's spin-off suggests that American business executives are waking up to the unavoidable risks of doing business in China—of inadvertently helping build China's AI systems that damage human rights and the public good.

Generative AI is a threat to CCP censorship

In April 2023, the Cyberspace Administration of China announced draft regulations for generative AI, setting down potential rules that chatbot-produced content follow "socialist core values" and avoid information that undermines "state unity."[14] The CCP's goal is a continuation of its past strategy to align new technologies and censor information in line with its political values. ChatGPT has not made its service available in China, but there is already significant demand. The black market is already flourishing with offerings of overseas ChatGPT access to people in China, but these days could be numbered.[15]

Generative AI, however, is a departure from the surveillance technologies that have defined the evolution of China's political censorship. Generative AI services have the potential to empower regular people who want to produce large amounts of content that challenge government propaganda and narratives. The question is whether China's "Great Firewall"—the harsh internet censorship system—can stand up to the potential of generative AI. Will China one day see an information renaissance, with stories of the Tiananmen Square massacre and Hong Kong protestors spread across the internet through uncontrollable chatbots?

Given the CCP's enormous success at censorship so far, I believe that it will once again succeed in coercing and coopting Chinese technology firms and transforming generative AI into a tool of state oppression. American technologists will unwittingly assist CCP goals if they cooperate too eagerly with state-connected Chinese companies, institutes, and people. As we have learned over the last decade, this is the sad truth of being a technologist in China.

---

[12] Shawn Johnson, "Neil Shen goes it alone in China after Sequoia split," *Financial Times*, June 9, 2023, https://www.ft.com/content/179eb51a-70eb-4c79-9e74-befd0f5a02b7.
[13] Alex Kondrad, "For Top VCs, ByteDance's Historic Windfall Remains a $220 Billion Mirage," *Forbes*, May 4, 2023, https://www.forbes.com/sites/alexkonrad/2023/05/04/bytedance-scrutiny-leaves-midas-investors-waiting-billions/.
[14] Chang Che, "China Says Chatbots Must Toe the Party Line," *New York Times*, April 24, 2023, https://www.nytimes.com/2023/04/24/world/asia/china-chatbots-ai.html.
[15] Caiwei Chen, "China's ChatGPT Black Market Is Thriving," *Wired*, March 7, 2023, https://www.wired.com/story/chinas-chatgpt-black-market-baidu/.

The US must use its global technological leadership to protect democracy and human rights from China's AI threats

The CCP is the greatest threat to human rights and democracy around the world. Although China is quickly catching up to US innovation, the US remains the leader in AI development. We must abandon the misguided idealism of working with Chinese companies and government bodies with the hope that AI will change the political system, allow for the opening of democratic discourse, and create safer global AI regulations. Rather than helping advance innovation, we will be doing the world a disservice by handing the keys to the CCP. Under Chinese law, these advanced AI applications will inevitably be used to oppress human rights and expand China's authoritarian footprint.

Rather, we should use our position of strength and our democratic values to carry out a two-fold strategy. First, AI talent and innovation must flow towards the direction of America and its allies. We must influence global AI standards, attract global AI talent away from China, and secure our software and hardware ecosystems from China's malign influences. Second, the most advanced American technologies and investments must not be allowed to flow in the direction of China. We must work against China's ambitions to develop advanced AI systems, influence global standards, and oppress dissidents around the world. The specific policy steps are as follows:

1. **The US must take the lead in developing global AI standards that uphold human rights and democratic values.**

   The CCP has loudly used multilateral membership bodies—the United Nations, the World Health Organization, and so forth—to shape global technology and science standards in its interests and to make countries all over the world dependent on Chinese technological innovation. The US must not shirk its global leadership, which would mean ceding ground to China and abandoning our allies in a moment of global struggle.

   In November 2021, 193 countries adopted the first-ever global agreement of AI ethics under the United Nations Educational, Scientific, and Cultural Organization (UNESCO), calling for a "do no harm" principle, personal data protection, and measures to prevent fairness and non-discrimination.[16] The US should leverage other United Nations bodies and the International Organization for Standardization (ISO) to build democratic AI principles and ensure that China's authoritarian goals do not crush the principles of human rights.

2. **American companies that help build China's oppressive AI ecosystem must be held accountable.**

[16] UNESCO, "Recommendations on the ethics of artificial intelligence," November 2021, https://www.unesco.org/en/articles/unesco-adopts-first-global-standard-ethics-artificial-intelligence.

China built its AI surveillance apparatus with the connivance and complacency of major American technology firms. The science corporation Thermo Fisher, for example, was caught selling DNA collection equipment directly to Xinjiang police authorities who used them for mass gathering of genetic data on the minority Uyghur population.[17] Since the late 1990s, Microsoft has established itself as the training ground for China's AI elites through its Beijing-based laboratory, Microsoft Research Asia. The laboratory has trained many of the AI leaders and developers who went on to found or join the executive leadership of rights-abusing firms such as Sensetime, Megvii, and iFlyTek. Beginning in 2019, the US government has sanctioned these individuals and their companies.[18]

So far, American technology giants have faced no punishment for their involvement in China's surveillance state. This subcommittee may consider drafting a bill that requires public corporations to publish their due diligence reports on their activities in China and the risks they have encountered with regards to human rights there. The subcommittee may also consider drafting a bill that criminalizes specific American business activities in China that are likely to support, directly or indirectly, human rights abuses by the CCP. This would include prison time for American business executives involved helping develop any form of AI in partnership with a Chinese entity, if the CCP will likely use that technology for the oppression of human rights and democratic values.

3. **Because Chinese software companies are required to partake in Chinese state intelligence operations, they should be compelled to separate their American businesses.**

Over the past decade, China has enacted a raft of draconian laws, such as the National Security Law and the National Intelligence Law, that require people in China to assist the government in intelligence-gathering when called upon, among other requirements.[19] While we in America have a system of due process and checks and balances that can guard against data overreach, in China no such rights exist. The private and personal data of Americans is not safe in the hands of Chinese-owned apps such as TikTok and Temu, whose owners and employees in China are required to hand over data to the state if it's requested.

Apps like TikTok are beginning to form the core of the US information environment, with sophisticated algorithms that recommend highly addictive

[17] Human Rights Watch, "China: Minority Region Collects DNA from Millions," December 13, 2017, https://www.hrw.org/news/2017/12/13/china-minority-region-collects-dna-millions.
[18] Kate Kaye, "Microsoft helped build AI in China. Chinese AI helped build Microsoft," *Protocol*, November 2, 2022, https://www.protocol.com/enterprise/us-china-ai-microsoft-research.
[19] Bonnie Girard, "The Real Danger of China's National Intelligence Law," February 23, 2019, https://thediplomat.com/2019/02/the-real-danger-of-chinas-national-intelligence-law/.

content, while being used to spy on US citizens.[20] This is a gaping breach of our ability to protect democratic values and human rights here in the US. In the event of conflict with China—an increasing likelihood with China's aggressive military posture—these apps have the potential to become misinformation machines designed to manipulate Americans with sophisticated and algorithmic propaganda. The solution is to force these firms to spin off their American operations into separate companies, ensuring their safety from CPP meddling.

4. **America and its allies must secure and coordinate global supply chains for advanced AI logic chips.**

The US has made remarkable progress in legislating and implementing export controls that prevent American firms from selling advanced chips and their components to China. In October 2022, the Biden administration implemented the most recent round of sanctions, restricting the export of certain services and equipment to China, effectively placing China generations behind American chip technologies for the latest AI applications.[21] Four months later, in February 2023, the Department of Commerce opened the first round of company grants under the CHIPS and Science Act, hoping to reshore semiconductor manufacturing capabilities and make the US more self-sufficient.[22]

The CHIPS and Science Act, however, is the starting point and not the last step. Advanced semiconductors are the most complex devices that humankind has ever made—and they cannot simply be manufactured end-to-end in the US. Chip supply chains depend on thousands of suppliers all over the world. The US needs to better coordinate with its key chip-producing and component-producing partners—South Korea, Taiwan, Japan, and the Netherlands—by upgrading the "Chip 4" talks into a formal consortium for coordinating R&D innovations.

The upgrade will enhance the implementation of the CHIPS and Science Act and the future of AI technologies by adding an element of multilateralism. Our technological partners will have better reason to believe their contributions to the US manufacturing ecosystem are profitable and worthwhile, a hedge against CCP aggression. If we can form a true semiconductor alliance, China will be unable to bully individual countries into supplying critical chip technologies for its AI systems.

---

[20] Emily Baker-White, "TikTok Spied on Forbes Journalists," *Forbes*, December 2, 2022, https://www.forbes.com/sites/emilybaker-white/2022/12/22/tiktok-tracks-forbes-journalists-bytedance/?sh=3b7173b97da5.
[21] Demetri Sevastopulo and Kathrin Hille, "US hits China with sweeping tech export controls," *Financial Times*, October 7, 2022, https://www.ft.com/content/6825bee4-52a7-4c86-b1aa-31c100708c3e.
[22] U.S. Department of Commerce, "Biden-Harris Administration Launches First CHIPS for America Funding Opportunity," February 28, 2023, https://www.commerce.gov/news/press-releases/2023/02/biden-harris-administration-launches-first-chips-america-funding.

As we enter the unprecedented age of generative AI, we must not allow China, a one-party authoritarian state, to infect the global AI ecosystem where it will oppress human dignity, civil liberties, and rule of law. We have seen the CCP's willingness to carry out genocide against its people with the help of AI surveillance systems. Now we must find ways to ensure that the words "never again" hold true. Thank you, Senators, for having me here today. I look forward to answering your questions.

United States Senate
Written Statement of Jennifer DeStefano
Abuses of Artificial Intelligence
June 13, 2023

Good Afternoon Senators, it is my great honor to speak with you today and to share my experience of how artificial intelligence is being weaponized to not only invoke fear and terror in the American public, but in the global community at large as it capitalizes on and redefines what we have known to be as “familiar”. I would like to take this moment to thank Senator Ossoff for inviting me to be here today. I would also like to thank Senator Blackburn for your concern on this ever evolving topic and community threat. AI is revolutionizing and unraveling the very foundation of our social fabric by creating doubt and fear in what was once never questioned, the sound of a loved one’s voice.

What is “familiar”? How many times have you received a phone call from your child and asked them to verify who is calling? How many times has a loved one reached out to you in despair and you stopped them to validate their identity? Did you hang up on them? Did you require to call them back to make sure you are speaking to the correct person? The answer is more than likely, never. Never have you stopped your loved one and questioned if the voice you are speaking with is really them. The sound of a loved one’s voice is often never questioned. It is designed by nature, it is designed by God, as a unique identity, as unique as a fingerprint. This familiar identity is how a mother knows if it’s her child crying in a room and it is how a newborn child instantly recognizes their mother.

It was a typical Friday afternoon for our family kicking off a weekend of races and rehearsals that often divide our family across the state. As the parents of four children close in age, we tend to have to “divide and conquer”. My husband was with our older daughter Brie and our youngest son in Northern Arizona training for ski races. I was with our older son and youngest daughter Aubrey in the valley as she had rehearsal. Ski racing is a high risk sport and Brie had not raced in years. At age 15, she promised me she would take it easy and not hurt herself by pushing to hard. When I first received a call from an “unknown” number upon exiting my car, I was going to ignore it. On the final ring I chose to answer as “unknown” calls can often be a doctor or a hospital. I answered the phone “ Hello”, on the other end was our daughter Briana sobbing and crying saying “mom”. At first I thought nothing of it, she had run into race gates and bruised herself before, not to worry. I casually asked her what happened as I had her on speaker walking through the parking lot to meet her sister. Briana continued with “mom, I messed up” with more crying and sobbing. Not thinking twice, I asked her again, “ok what happened?” Suddenly a man’s voice barked at her to “lay down and put your head back”. At that moment I started to panic. My concern escalated and I demanded to know what was going on, but nothing could have prepared me for her response. “MOM THESE BAD MEN HAVE ME, HELP ME, HELP ME!!” She begged and pleaded as the phone was taken from her. A threatening and vulgar man took over the call “Listen here, I have your daughter, you tell anyone, you call the cops, I am going to pump her stomach so full of drugs, I am going to have my way with her, drop

her in mexico and you'll never see her again!" all the while Briana was in the background desperately pleading "mom help me!!!"

With my shaking hand on the door handle to the studio, I put the man on mute, flung open the door and started screaming for help. The next few minutes were a parent's worst nightmare. I was fortunate to have a few moms at the studio who surrounded me, hearing all of the vulgar threats the man was making. One mom ran outside and called 911. Our 13 year old daughter Aubrey stood paralyzed in fear. I needed her help, her sister was in trouble and we had to find her. Another mom ran to her to aid as they started making calls to her dad, her brothers, anyone that could help us figure out what happened to Brie. The kidnapper demanded a million dollars. That was not possible and so the kidnapper decided on $50,000, in cash. At this moment, the mom who called 911 came inside and shared with me that 911 was familiar with an AI scam where they can replicate your loved one's voice. I didn't believe this was a scam. It wasn't just Brie's voice, it was her cries, it was her sobs that were unique to her. It wasn't possible to fake that I protested. She told me that AI can also replicate inflection and emotion. That gave me a little hope but still was not enough. I proceeded with the negotiations. I asked for wiring instructions and routing numbers for the $50,000 but was refused. "Oh no" the man demanded, "that's traceable, that's not how this is going to go down. We are going to come pick you up!" "What?" I shouted, "You will agree to being picked up in a white van, with a bag over your head so you don't know where we are taking you. You better have all $50k in cash otherwise both you and your daughter are dead! If you don't agree to this, you will never see your daughter again!" he screamed. I had to stall, I asked the mom on the call with 911 to send police, I needed

to stall until I had police with me. Then the mom who was making calls with Aubrey was able to get my husband on the phone. He frantically located Brie resting safely in bed. Brie had no idea what was happening. As I was negotiating the arrangements of the abduction of myself to save my daughter, the mom came to me and told me she found Brie and that she was safe. I didn't believe her. How could she be safe with her father and yet be in the possession of kidnappers? It was not making any sense. I had to speak to Brie. I could not believe she was safe until I heard her voice say she was. I asked her over and over again if it was really her, if she was really safe, again, is this really Brie, are you sure you are really safe?! My mind was whirling. I do not remember how many times I needed reassurance, but when I finally took hold of the fact she was safe, I was furious. I lashed at the men for such a horrible attempt to scam and extort money. To go so far as to fake my daughter's kidnapping was beyond the lowest of the low for money. They continued to threaten to kill Brie. I made a promise that I was going to stop them, that not only were they never going to hurt my daughter, but that they were not going to continue to harm others with their scheme. After I hung up, I collapsed to the floor in tears of relief. When I called the police to pursue the matter, unfortunately I was met with this is a prank call. That it happens often and that I am probably not in harm's way (although not a guarantee). I was offered to have a police officer call me from another "unknown" number if it would make me feel better as law enforcement numbers are also blocked. That certainly did not make me feel better. Bottom line was no actual crime had been committed, no one was physically kidnapped, and no money was transferred, period, the end.

But that wasn't the end, it couldn't be the end. If it was the end, then this nightmare would never stop. I stayed up all night paralyzed in fear. Do they know where I am? Do they know where my daughter is? How did they get her voice? How did they get her crying, her sobs that are unique to her. She is not a very public person. Are we being cyber stalked? Targeted? So many questions that I could not leave unanswered, so I turned to our community and the response was overwhelming! Friends and neighbors came out of the woodwork with their stories. Kidnapping phone calls coming from their children's phones, bags of money being driven halfway to Mexico, even voices of young children nowhere to be found on social media and who do not have phones, the stories kept pouring in. Even my own mother received a call with my brother's voice claiming to be in an accident and needing money for the hospital bill! My mother is hard of hearing and quite spunky. After having the caller repeat the request multiple times, she realized the language used was not something my brother would say. She told the caller to call their real mother and hung up. The common response the victims received from authorities was that nothing could be done. In fact, one mother I know personally shared with me how she was even mocked by her son's school and security officer. She called his school frantically trying to locate her son when she received a call from him that he had been kidnapped. He even used his unique nickname during the call to self identify. Fortunately he was safe in class and she was told "this happens all the time" as her fear was dismissed. "It's the most frustrating, maddening, scary and invaded I've felt…my fear is that it is only a matter of time until someone actually follows through with the threat", she told me as she has been living in fear and concern for her son's safety ever sense.

Money scams have been around for thousands of years. We have all heard of "snake oil" and remember the days of "swap land" sold as paradise in Florida. This is entirely

different. This is terrorizing with lasting post traumatic stress. Even months later, sharing the story shakes me to my core. It was my daughter's voice. It was her cries, her sobs. It was the way she spoke. I will never be able to shake that voice out of mind. It's every parents' worst nightmare to hear your child pleading with fear and pain, knowing that they are being harmed and you are helpless and desperate. The longer this form of terror remains unpunishable, the farther and more egregious it will become. The thought crossed my mind before I hung on the "kidnappers" to follow through with the physical abduction of me. Was that what would it take to bring an end to this? Was that what it would take in order to have a pursuable criminal offense?

As our world moves at a lightning fast pace, the human element of familiarity that lays foundation to our social fabric of what is "known" and what is "truth", is being revolutionized with Artificial Intelligence. Some for good, and some for evil. No longer can we trust "seeing is believing", "I heard it with my own ears" nor even the sound of our own child's voice. This concept redefines and rewrites what the very meaning of "familiarity" means. Familiarity is defined as "the quality of being well known or knowledge of something" and further is defined as "relaxed friendliness or intimacy between people." Familiar and family share the root word "Famil" which establishes strength of a relationship between one person and another. I ask you, when your mother calls, are you going to hang up and call her back to make sure it is really her? When your child calls you in need of help, will you disconnect the call and say I don't believe its really you? Is this our new norm? Is this the future we are creating by enabling this abuse of Artificial Intelligence without consequence?

I want to thank you for your time and attention today. Congress has a large and looming task ahead. How do we move forward as a community with this haunting reality that is plaguing us? If left uncontrolled, unguarded and without consequence, it will rewrite our understanding and perception what is and what is not truth. It will erode our sense of “familiar” as it corrodes our confidence in what is real and what is not. This is a non-partisan matter and I have seen the hands reach across the aisle in unified concern. That gives me great hope. How to contain the ever evolving Artificial Intelligence and its unknowns, is not an easy task. My sincere thanks and humble appreciation for your time and attention today. I thank all of you, and especially Senator Ossoff and Congress at large, for tirelessly taking action to keep our community and world safe from the hands of evil. I am one person, one story, but I am not the only one and I certainly will not be the last one unless action is taken. I wish you God’s speed.

Testimony of Alexandra Reeve Givens
President & CEO, Center for Democracy & Technology

For the U.S. Senate Committee on the Judiciary Subcommittee on Human Rights and the Law
Hearing Entitled "Artificial Intelligence and Human Rights"

June 13, 2023

Chair Ossoff, Ranking Member Blackburn and other members of the Subcommittee, thank you for inviting me to testify today on the important issue of AI and human rights. The world's attention is rightly focused on the possibilities and the risks of AI systems. As policymakers look to address potential harms and promote responsible innovation, it is essential that they do so with a focus on human rights – and in particular, with the conviction that fundamental rights and freedoms belong inalienably to all people, including the rights to liberty, privacy, freedom of expression and opinion, peaceful assembly, and equal treatment before the law.[1]

AI systems are already being used in ways that threaten these rights, and rapid advancements in generative AI and text and image analysis will exacerbate the risks. Today I will focus on two distinct areas where AI harms are already being felt: the use of face recognition and biometric surveillance capabilities by law enforcement, and the impact of generative AI on elections and democratic discourse. For reasons I will explain in my testimony, these applications of AI are vastly different from one another, with different considerations at stake as Congress considers appropriate policy interventions.

Of course, these areas are not the only ways in which AI is impacting human rights. In previous testimony before the U.S. Senate Committee on Homeland Security and Government Affairs, I described risks posed by AI systems to people's civil rights and access to economic opportunities – for example when people are applying for jobs, housing, or credit – and potential policy responses.[2] I also described how AI is being used in ways that jeopardize the fair administration of public benefits programs, and steps the government should take to protect people's access to basic services and due process rights. Those issues are ripe and important priorities for government intervention.

At a time when many are discussing the long term existential risks of AI systems, there are concrete issues on which Congress and the U.S. government can act *today* — and, in doing so, demonstrate what it means to ensure AI is developed in a manner that centers democratic values and human rights.

[1] United Nations General Assembly. The Universal Declaration of Human Rights (UDHR). New York: United Nations General Assembly, 1948.
[2] Alexandra Reeve Givens, "Press Release: In Senate Testimony, CDT CEO Alexandra Givens Calls For Cross-Society Effort in Addressing Risks of AI", Center for Democracy & Technology, March 8, 2023, https://cdt.org/insights/press-release-in-senate-testimony-cdt-ceo-alexandra-givens-calls-for-cross-society-effort-in-addressing-risks-of-ai/.

AI & Government Surveillance

Last fall, many of us were inspired by the images of brave Iranian women protesting the death of 22-year-old Mahsa Amini after she was arrested for allegedly improperly wearing the hijab. But we were not the only ones watching those protests. In Iran today, face recognition technology allows the government to identify protestors and take action against them. Demonstrators have received text messages from local police stating that they were observed at a protest and should not join further demonstrations.[3] Iranian officials also announced that they would use face recognition in public spaces to detect and identify women who were not "correctly" wearing a hijab.[4] A member of parliament explained that women who dress improperly would receive text message warnings, followed by penalties such as their bank accounts being blocked. In Iran, citizens must use biometric national identity cards to receive pensions and food rations, open bank accounts and access the domestic internet – making these threats of automated punishments all too real. In this context, AI systems are enabling a repressive regime to identify dissenters, subject them to pervasive surveillance, and then automate their punishment.

Face recognition technology has been used in similar ways by the Chinese government, to promote social control through mass enforcement and public shaming of minor offenses such as jaywalking,[5] as well as for its notorious treatment of China's Uyghur minority.[6] Face recognition has also been used to identify protestors in Russia, Hong Kong and Uganda, among other countries.[7]

Such examples may feel far from the United States, but the technical capabilities exist here, and we do not have adequate legal frameworks to address them. In the U.S. there have already been abuses: In 2020, police in multiple Florida cities used facial recognition to identify and catalog activists engaging in peaceful civil rights protests supporting the Black Lives Matter movement.[8] In Baltimore, face recognition technology was used in real time to target people who were protesting after the death of Freddie Gray, with law enforcement scanning the crowd to identify individuals with outstanding warrants for unrelated offenses, and arresting them on site.[9] When

---

[3] Sam Biddle and Murtaza Hussain, "Hacked Documents: How Iran Can Track And Control Protesters' Phones", *The Intercept*, Oct. 28, 2022, https://theintercept.com/2022/10/28/iran-protests-phone-surveillance/.

[4] Khari Johnson, "Iran to use facial recognition to identify women without hijabs", *Ars Technica*, Jan. 11, 2023, https://arstechnica.com/tech-policy/2023/01/iran-to-use-facial-recognition-to-identify-women-without-hijabs/.

[5] Alfred Ng, "How China uses facial recognition to control human behavior", *CNET*, Aug. 11, 2020, https://www.cnet.com/news/politics/in-china-facial-recognition-public-shaming-and-control-go-hand-in-hand/ ("The punishing of these minor offenses is by design, surveillance experts said. The threat of public humiliation through facial recognition helps Chinese officials direct over a billion people toward what it considers acceptable behavior, from what you wear to how you cross the street").

[6] Paul Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority", *The New York Times*, Apr. 14, 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

[7] Paul Mozur, "In Hong Kong Protests, Faces Become Weapons", *The New York Times*, July 26, 2019, https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html; Lena Masri, "Facial recognition is helping Putin curb dissent with the aid of U.S. tech", *Reuters*, Mar. 28, 2023, https://www.reuters.com/investigates/special-report/ukraine-crisis-russia-detentions/; Stephen Kafeero, "Uganda is using Huawei's facial recognition tech to crack down on dissent after anti-government protests", *Quartz*, Nov. 27, 2020, https://qz.com/africa/1938976/uganda-uses-chinas-huawei-facial-recognition-to-snare-protesters.

[8] Joanne Cavanaugh Simpson and Marc Freeman, "South Florida police quietly ran facial recognition scans to identify peaceful protestors. Is that legal?", *Sun Sentinel*, June 26, 2021, https://www.sun-sentinel.com/2021/06/26/south-florida-police-quietly-ran-facial-recognition-scans-to-identify-peaceful-protestors-is-that-legal/.

[9] Kevin Rector and Alison Knezevich, "Social media companies rescind access to Geofeedia, which fed information to police during 2015 unrest", *The Baltimore Sun*, Oct. 11, 2016, https://www.baltimoresun.com/news/crime/bs-md-geofeedia-update-20161011-story.html.

face recognition is used in this way, it violates people's rights to freedom of expression and peaceful assembly. Congress must act to rein it in.

Facial recognition technology is becoming more widely available and cheaper to use. A study by Georgetown's Center on Privacy and Technology published in 2016 showed that at least one in four state and local law enforcement agencies had access to facial recognition – and that was seven years ago.[10] Research suggests that the FBI conducts thousands of scans per month, matched against reference databases of hundreds of millions of photos.[11] Several years ago, Americans were shocked to learn about the practices of the private company Clearview AI, which claims to have scraped over 20 *billion* photographs from the internet to power its face recognition systems.[12] Clearview has now been used by over 3000 federal, state and local law enforcement agencies in the United States to provide facial recognition services.[13]

Policymakers should treat facial recognition as a priority because it is a double-edged sword: Facial recognition is dangerous when it works poorly, and dangerous in an entirely different way when it works well. States have begun to respond to this threat, with over a dozen enacting meaningful limits and some jurisdictions banning the technology.[14] It is critical that Congress act as well. As our nation considers its approach to governing AI, this is an area where Congress could draw a clear contrast to autocratic regimes, demonstrating America's commitment to human rights.

The urgent need for regulation of facial recognition technology is clear. Facial recognition misidentifications have already caused numerous innocent people to be wrongfully arrested and jailed. Most recently, Randel Reid was held for six days in a Georgia jail because a facial recognition system misidentified him,[15] the latest in a series of known cases.[16] Because of police overreliance on AI, these individuals faced indignity, deprivation of liberty, and lasting harms such as loss of employment, steep legal fees, and mental trauma.[17] And since police use of facial recognition is often hidden,[18] these incidents likely represent just the tip of the iceberg.[19]

---

[10] The Perpetual Line-Up: Unregulated Police Face Recognition in America, Georgetown Law Center on Privacy and Technology, Oct. 18, 2016, https://www.perpetuallineup.org/.
[11] *Id.*; *see also* Charlie Osborne, "FBI, ICE plunder DMV driver database 'gold mine' for facial recognition scans", *ZDNET*, July 8, 2019, https://www.zdnet.com/article/fbi-and-ice-are-using-dmv-gold-mine-for-facial-recognition-scans/.
[12] Kashmir Hill, "Your Face is Not Your Own", *The New York Times Magazine*, Mar. 18, 2021, https://www.nytimes.com/interactive/2021/03/18/magazine/facial-recognition-clearview-ai.html.
[13] *Id.*
[14] Jake Laperruque, "Limiting Face Recognition Surveillance: Progress and Paths Forward", Center for Democracy & Technology, Aug. 23, 2022, https://cdt.org/insights/limiting-face-recognition-surveillance-progress-and-paths-forward/.
[15] Kashmir Hill and Ryan Mac, "'Thousands of Dollars for Something I Didn't Do'", *The New York Times*, Mar. 31, 2023, https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html.
[16] Khari Johnson, "How Wrongful Arrests Based on AI Derailed 3 Men's Lives", *WIRED*, Mar. 7, 2022, https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/.
[17] *Id.*; *see also* Elaisha Stokes, "Wrongful arrest exposes racial bias in facial recognition technology", *CBS News*, Nov. 19, 2020, https://www.cbsnews.com/news/detroit-facial-recognition-surveillance-camera-racial-bias-crime/; Kashmir Hill, "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match", *The New York Times*, Dec. 29, 2020, https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.
[18] Khari Johnson, "The Hidden Role of Facial Recognition Tech in Many Arrests", *WIRED*, Mar. 7, 2022, https://www.wired.com/story/hidden-role-facial-recognition-tech-arrests/; Jennifer Valentino-DeVries, "How the Police Use Facial Recognition, and Where It Falls Short", *The New York Times*, Jan. 12, 2020, https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html.
[19] Disturbingly, some facial recognition misidentifications likely have resulted in prison time for innocent persons, either wrongfully convicted or pressured to accept a plea bargain out of fear of long sentences or extended time in pretrial detention.

Misidentification stems from a range of causes. Most facial recognition systems display algorithmic bias; studies have repeatedly shown propensity to misidentify people of color and women at higher rates than white people and men.[20] Software settings and nature of use impact accuracy as well. Many law enforcement agencies, including the FBI, set their systems to return several potential matches for *every* facial recognition scan even if the "confidence threshold"—meaning the required level of certainty to list an individual as a possible match—is unreliably low.[21] Law enforcement also regularly uses dubious methods to alter or replace images before scanning, from using CGI to artificially fill in uncaptured portions of a face, to replacing photos entirely with a composite sketch or celebrity look alike.[22] Finally, accuracy can vary significantly based on image quality: Lighting, photo resolution, distance, camera angle, and facial obstructions can all have a major impact on whether facial recognition returns accurate matches.[23] This is critical because even if algorithmic bias were solved, and responsible settings and use parameters were employed, varying image quality will always cause misidentification risk.

Just as serious as misidentifications are the dangers of accurate facial recognition being used for surveillance. The examples I shared previously from Iran, China, Russia, Uganda – and at least three U.S. cities – shows how easily face recognition technology can impinge on people's rights to express themselves through protest and to peacefully assemble. Facial recognition could be employed to monitor, catalog, and engage in disparate targeting of individuals participating in a variety of sensitive or constitutionally protected activities, such as attending a political rally, going to a house of worship, purchasing a firearm from a licensed shop, or visiting a medical clinic. Absent strong limits, law enforcement authorities could misuse AI technology to track and catalog individuals' most sensitive activities with little effort, and on an unprecedented scale. The U.S. must show leadership by curtailing such a direct assault on civil liberties.

Given the range of risks facial recognition poses to civil rights and civil liberties, there is not a silver bullet policy solution: lawmakers need to enact a broad set of safeguards to prevent harm,

---

[20] Joy Buolamwini and Timnit Gebru (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, *Fairness, Accountability and Transparency, Proceedings of Machine Learning Research* 81:77-91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf; Patrick Grother, Mei Ngan, and Kayee Hanaoka (Dec. 2019). *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, National Institute Of Science and Technology. https://doi.org/10.6028/NIST.IR.8280.

[21] Kimberly J. Del Greco, "Facial Recognition Technology: Ensuring Transparency in Government Use", Federal Bureau of Investigation, June 4, 2019, https://www.fbi.gov/news/testimony/facial-recognition-technology-ensuring-transparency-in-government-use; Drew Harwell, "Oregon became a testing ground for Amazon's facial-recognition policing. But what if Rekognition gets it wrong?", *The Washington Post*, April 30, 2019, https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/ ("But deputies here are not shown that search-confidence measurement when they use the tool. Instead, they are given five possible matches for every search, even if the system's certainty in a match is far lower").

[22] James O'Neill, "How Facial Recognition Makes You Safer", *The New York Times*, June 9, 2019, https://www.nytimes.com/2019/06/09/opinion/facial-recognition-police-new-york-city.html; Clare Garvie, "Garbage In, Garbage Out | Face Recognition on Flawed Data", Georgetown Law Center on Privacy & Technology, May 16, 2019, https://www.flawedfacedata.com/ ("One detective from the Facial Identification Section (FIS), responsible for conducting face recognition searches for the NYPD, noted that the suspect looked like the actor Woody Harrelson …. A Google image search for the actor predictably returned high-quality images, which detectives then submitted to the face recognition algorithm in place of the suspect's photo.")

[23] The Constitution Project's Task Force on Facial Recognition Surveillance and Jake Laperruque, "Facing the Future of Surveillance", Project on Government Oversight, Mar. 4, 2019, https://www.pogo.org/report/2019/03/facing-the-future-of-surveillance.

both from misidentifications and misuse.[24] The Center for Democracy & Technology views the following measures as key to effectively regulating law enforcement use of facial recognition:

1) **A warrant rule:** Law enforcement use of facial recognition should require obtaining a warrant from a judge, based on probable cause that the individual to be scanned committed a crime.[25] Warrants are a fundamental privacy safeguard and key to preventing abuse, notably using facial recognition to identify, catalog, and target individuals engaged in lawful and sensitive activities, such as protests.
2) **A serious crime limit:** Face recognition technology should be restricted for use only in investigating serious offenses.[26] Such limitations would prevent selective targeting and prosecution, as well as prevent misidentifications in scenarios least likely to receive due scrutiny: the investigation and prosecution of low-level crimes.
3) **Notification for arrested individuals:** Law enforcement should not be allowed to routinely hide their use of facial recognition from defendants and the broader public.[27] This common practice undermines defendants' due process rights, and prevents examination of errors and other meaningful oversight.
4) **Prohibiting overreliance on matches:** Police should not be permitted to use facial recognition as the sole basis for arrests or other police actions. Given that the technology's accuracy varies significantly based on a range of factors, independent investigative work is essential.
5) **Prohibiting untargeted scans:** Facial recognition technology may soon focus on untargeted scans—whereby every individual passing through a video feed is identified with facial recognition—but this method is far too unreliable for law enforcement use. Pilot programs have produced false positives of 81 to 96 percent.[28] Even if these extreme error rates were to improve, such a use of face recognition technology would constitute unacceptable dragnet surveillance that should not be deployed.
6) **Testing and accuracy standards:** Any law enforcement use of facial recognition should require that software be subject to independent testing and meet accuracy standards. Testing should focus on live field conditions that replicate investigative use, and accuracy standards should limit use to algorithms with highest overall accuracy and that display no variance based on demographic traits.

---

[24] While our recommendations focus on safeguards and limits for law enforcement use of facial recognition, it is important to acknowledge that many privacy, civil rights, and civil liberties groups—including CDT—have called for a moratorium on facial recognition, or for its use by law enforcement to be banned entirely. Some local face recognition laws have taken this approach. CDT supports enacting a moratorium while evaluating proper restrictions and safeguards as providing the strongest protections for civil rights and civil liberties. *See, e.g.*, LDF Letter re: July 13, 2021 Subcommittee on Crime, Terrorism, and Homeland Security Hearing on Law Enforcement Use of Facial Recognition Technology, https://www.naacpldf.org/wp-content/uploads/2021.07.20-LDF-Statement-on-Law-Enforcement-U_Emily-Fisher-1.pdf.

[25] This should include sensible limited exceptions, such as identifying victims and incapacitated persons.

[26] A serious crime limit has been used for over 50 years to prevent wiretap surveillance from becoming pervasive. *See* 18 U.S.C. § 2516.

[27] Khari Johnson, "The Hidden Role of Facial Recognition Tech in Many Arrests", *WIRED*, Mar. 7, 2022, https://www.wired.com/story/hidden-role-facial-recognition-tech-arrests/; Jennifer Valentino-DeVries, "How the Police Use Facial Recognition, and Where It Falls Short", *The New York Times*, Jan. 12, 2020, https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html.

[28] Lizzie Dearden, "Facial recognition wrongly identifies public as potential criminals 96% of time, figures reveal", *The Independent*, May 7, 2019, https://www.independent.co.uk/news/uk/home-news/facial-recognition-london-inaccurate-met-police-trials-a8898946.html; Rachel England, "UK police's facial recognition system has an 81 percent error rate", *Engadget*, July 4, 2019, https://www.engadget.com/2019-07-04-uk-met-facial-recognition-failure-rate.html.

The adoption of face recognition laws by over a dozen states[29] demonstrates an emerging consensus for regulating this surveillance. Unfortunately, thus far Congress has placed no limits on facial recognition, leaving this powerful technology unrestricted. Last year a bill was introduced in the House, H.R. 9061, The Facial Recognition Act, that included many of the recommendations listed above, and that the Center for Democracy & Technology endorsed.[30] We encourage Congress to act with urgency to place safeguards on this form of AI surveillance, and focus on the policies described above.

Generative AI, Elections & Democratic Discourse

Turning to my second area of focus, rapid advances in generative AI are spurring creativity and innovation, but also raise significant threats for human rights. Already there have been instances showing the professional, reputational and potential physical harms that may arise when people rely on generated results as accurate, not accounting for the likelihood of "hallucinations", or mistaken results.[31] Generative AI tools are likely to exacerbate fraud, as tools make it easier to quickly generate massive amounts of convincing text, as well as personalized scams, or to trick people by impersonating a familiar voice.[32] Deepfakes – videos or images that have been digitally manipulated to misrepresent the voice and likeness of another person – can misrepresent public figures or events in a way that threatens elections, national security, and general public order.[33] Deepfakes can also be used to defraud, harass, and extort people.[34] None of these harms is new, but they are made cheaper, faster, and more effective by the ease, speed and widespread accessibility of generative AI tools.

The threats to elections and democratic discourse are particularly worth highlighting. In previous elections, operatives used robocalls to spread incorrect information about mail-in voting in an effort to suppress Black voter turnout,[35] and deceptive text messages to spread intentionally misleading voting instructions for a Kansas ballot initiative in 2022.[36] It is easy to imagine bad actors using AI to exponentially grow and personalize voter suppression or other targeting efforts, increasing their harmful impact. Today, consumers can often spot a scam email, text or robocall because it uses non-personalized language and there may be grammatical

---

[29] Jake Laperruque, "Limiting Face Recognition Surveillance: Progress and Paths Forward", Center for Democracy & Technology, Aug. 23, 2022, https://cdt.org/insights/limiting-face-recognition-surveillance-progress-and-paths-forward/.
[30] Jake Laperruque, "The Facial Recognition Act: A Promising Path to Put Guardrails on a Dangerously Unregulated Surveillance Technology", *Lawfare*, Nov. 1, 2022, https://www.lawfareblog.com/facial-recognition-act-promising-path-put-guardrails-dangerously-unregulated-surveillance-technology.
[31] Karen Weise and Cade Metz, "When A.I. Chatbots Hallucinate", *The New York Times*, May 1, 2023, https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html.
[32] Steve Mollman, "Scammers are using voice-cloning A.I. tools to sound like victims' relatives in desperate need of financial help. It's working", *Fortune*, Mar. 5, 2023, https://fortune.com/2023/03/05/scammers-ai-voice-cloning-tricking-victims-sound-like-relatives-needing-money/.
[33] Shannon Bond, "Fake viral images of an explosion at the Pentagon were probably created by AI", *NPR*, May 22, 2023, https://www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-probably-created-by-ai; David Klepper and Ali Swenson, "AI presents political peril for 2024 with threat to mislead voters", *AP News*, May 14, 2023, https://apnews.com/article/artificial-intelligence-misinformation-deepfakes-2024-election-trump-59fb51002661ac5290089060b3ae39a0.
[34] *See e.g.,* Henry Ajder, Giorgio Patrini and Francesco Cavalli, "Automating Image Abuse: Deepfake bots on Telegram", *Sensity*, Oct. 2020 (deepfake bots on Telegram digitally "undress" more than 100,000 women on the platform); Thomas Brewster, "Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find", *Forbes*, Oct. 14, 2021, https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=7d29a3f87559 (audio deepfake of executives' voices used to steal millions of dollars from companies).
[35] Christine Chung, "They Used Robocalls to Suppress Black Votes. Now They Have to Register Voters.", *The New York Times*, Dec. 1, 2022, https://www.nytimes.com/2022/12/01/us/politics/wohl-burkman-voter-suppression-ohio.html.
[36] Isaac Stanley-Becker, "Misleading Kansas abortion texts linked to Republican-aligned firm", *The Washington Post*, Aug. 2, 2022, https://www.washingtonpost.com/politics/2022/08/02/kansas-abortion-texts/.

or language errors (or, in the case of robocalls, a notably automated voice). Generative AI tools will make it easier to create tailored, accurate, realistic messages that draw victims in.

Generated images can also twist public understanding of political figures and events. Recordings of public figures' voices have been manipulated to trick senior government officials into thinking they are speaking with government leaders.[37] Videos and images have been digitally altered to make public officials appear incompetent, compromised, or to misrepresent their policy positions.[38] Experts have warned how deepfakes, which are difficult to authenticate or rebut, could impact an election in the closing days of voting, when there is little time to set the record straight, or before a debate.[39] More generally, the growth of inauthentic content makes it harder for people to know what news and content they can trust, such that even authentic content is undermined. Journalists, whistleblowers, and human rights defenders are experiencing these effects already, facing higher hurdles than ever before to establish and defend their credibility.[40]

While the rise of affordable generated content poses new threats to public discourse, policy interventions must be approached with care. This is because there are many legitimate reasons why people use software to generate and alter content: from laypeople and artists using AI to make creative works; to people engaging in parody; actors being de-aged in a movie; voices being sampled for a music track; or researchers altering images of North American and European cities to show what they would look like if they faced the same bombardment as the cities attacked in the Syrian war.[41] Barring or heavily restricting such activities would harm free expression, creativity and innovation, and quickly run afoul of the First Amendment.

Efforts to restrict or condition the distribution of generative images may also suppress protected expressive activities. To give one example, in recent years a number of companies and stakeholders have come together in the Content Authenticity Initiative, an impressive undertaking that allows photographers and other content creators to attach immutable provenance signals showing the authenticity of their work (such as details of the image's creator, date/time/location, tracked edits and more).[42] This is a creative solution to help newspapers, human rights watchdogs and others reassure the public about the authenticity and provenance of images they create and display. But *mandating* the use of such an authenticity standard (or

---

[37] *See e.g.,* Bobby Allyn, "Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn", *NPR*, Mar. 16, 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia (the minute long deepfake video "shows a rendering of the Ukrainian president appearing to tell his soldiers to lay down their arms and surrender the fight against Russia"); Philip Oltermann, "European politicians duped into deepfake video calls with mayor of Kyiv", *The Guardian*, Jun. 25, 2022, https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko.

[38] *See e.g.,* Hannah Denham, "Another fake video of Pelosi goes viral on Facebook", *The Washington Post,* Aug. 3, 2020, https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/ (video depicts Pelosi slurring her speech and appearing intoxicated"); Alexandra Ulmer and Anna Tong, "Deepfaking it: America's 2024 election collides with AI boom", *Reuters*, May 30, 2023, https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/; Zeke Miller, "Rubio Campaign Fires Back at Cruz Over Photoshopped Image", *Time*, Feb. 18, 2016, https://time.com/4229092/marco-rubio-ted-cruz-photoshop/. While running for re-election in 2019, Houston's mayor said a critical ad ran by a fellow candidate broke a Texas law that bans certain misleading political deepfakes. Ivory Hecker, "Mayor Turner calls for criminal investigation of Tony Buzbee's attack ad", *Fox 26 Houston*, Oct. 17, 2019, https://www.fox26houston.com/news/mayor-turner-calls-for-criminal-investigation-of-tony-buzbees-attack-ad.

[39] James Bickerton, "Deepfakes Could Destroy the 2024 Election", *Newsweek*, Mar. 24, 2023, https://www.newsweek.com/deepfakes-could-destroy-2024-election-1790037.

[40] Sam Gregory, "Tracing trust: Why we must build authenticity infrastructure that works for all", *Witness*, May 2020, https://blog.witness.org/2020/05/authenticity-infrastructure/.

[41] Tiffany Hsu, "As Deepfakes Flourish, Countries Struggle With Response", *The New York Times*, Jan. 22, 2023, https://www.nytimes.com/2023/01/22/business/media/deepfake-regulation-difficulty.html.

[42] *See* Content Authenticity Initiative, https://contentauthenticity.org/.

prohibiting the distribution of materials without such standards) would be deeply problematic, because it would suppress the posting and sharing of lawful images whose creators lacked the resources or awareness to use a provenance tool, who face safety risks if their work can be traced back to them, or who simply do not want to do so.

The challenges of regulating deepfakes does not mean policymakers must sit idle. To the contrary, there are concrete steps Congress can take to increase transparency and accountability in the design, development and use of generative AI tools, as well as appropriations provisions, oversight of relevant federal agencies, and steps such as hearings, convenings, and/or the creation of a Commission to highlight best practices and novel innovations to address potential harms.

1) **Mandating transparency & disclosures of AI risks.** Several legislative proposals introduced last Congress seek to increase the accountable design and transparency of AI systems, including the Algorithmic Accountability Act, and the algorithmic impact assessment provision of the bipartisan American Data Privacy & Protection Act. These measures were drafted before the wide-scale public release of generative AI systems, but their principles lay an important foundation for future work.

   As a starting point, Congress could require the developers of AI systems that can be used in high-risk settings to disclose how their tools are developed and designed, to test them using frameworks based on principles such as those set out in the Blueprint for an AI Bill of Rights and the NIST AI Risk Management Framework, and to share the analysis of those tests with an outside regulator (with some version made available for the public and for independent researchers, balancing concerns about the potential privacy and safety aspects of such disclosures). Such steps would increase transparency and support meaningful public dialogue about how tools are developed and governed. They would also normalize the principle that companies designing and deploying AI tools *must* analyze and document how they work, identify potential risks, and disclose the steps they have taken to mitigate those risks. Such legislation would establish an essential baseline, and need not foreclose potential legislation on minimum design and safety standards, the specific regulation of highly capable foundation models, or further steps to address other high-risk AI uses.

2) **Examining how existing criminal and civil laws map onto harms created by new tools, and filling gaps.** In some instances, the appropriate framework to address harms created by generative AI (and other AI systems) may be litigation under existing laws. For example, people who use AI to perpetrate scams could be prosecuted for fraud, extortion, or harassment; face investigation by the Federal Trade Commission for unfair and deceptive trade practices or the Federal Elections Commission for violating campaign laws; or face civil litigation for claims such as fraud, intentional infliction of emotional distress, harassment, defamation and intellectual property violations. Congress should monitor whether these existing legal frameworks adequately address emerging harms.[43]

[43] Four federal agencies recently announced their efforts to enforce existing laws to protect the American public from AI-related harms. Other agencies should take similar steps, and Congressional committees of relevant jurisdiction can support these efforts to understand how existing

In assessing liability, courts will have to tackle the complex question of whether and when developers of generative AI tools bear legal liability for the content those tools produce. Courts will have to consider whether the content generated by an AI tool is properly considered to be the speech of the user who prompted its creation, or something partially or wholly created or developed by the AI tool itself. This will likely differ depending on the fact pattern: for example, whether a user inputted specific prompts aiming to generate the content that gave rise to litigation, such as soliciting a list of crimes committed by a private individual and publishing that list with reckless disregard for whether the information was true, or whether the AI tool was the source of the content giving rise to litigation, , such as making up dangerously incorrect medical advice in response to a query. In addition to statutes and case law regarding intermediary liability protections, courts will need to consider a range of common law principles from across civil and criminal law, including standards for aiding and abetting liability, and questions of knowledge and intent for both the user and the developer (and, if different, the deployer) of the tool. Companies will need to point to content policies and technical safeguards they have in place to mitigate foreseeable misuses and other harms.

As courts grapple with these and other complex issues, Congress can shine a light and drive public discourse — and then act as appropriate to fill in the gaps. This could include hearings and reports by Congressional committees in their areas of jurisdiction, commissioning reports by the GAO or federal agencies, or, more formally, the creation of an expert Commission to advance such work.[44]

3) **Advancing best practices for responsible design and governance of generative AI systems**. There is an urgent need for companies developing generative AI systems to develop robust safety processes and other governance measures, as many of their CEOs have themselves publicly declared.[45] This can include steps ranging from well-developed content policies and technical safeguards that limit the creation of certain high-risk content or uses of the technology;[46] robust pre- and post-release testing to identify and address bias and potential harms; improved interfaces, labeling and product descriptions to better educate

---

laws map onto novel fact patterns. *See* Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems, Apr. 25, 2023, https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf.

[44] *See, e.g.,* Deepfake Task Force Act, S.2559, 117th Cong. (2021-2022); American Data Privacy & Protection Act of 2022 (ADPPA), H.R. 8152, 117th Cong. (2021-2022). These proposals both focus on creating a task force (or in the case of the ADPPA, mandating annual reporting by the Commerce Department) on the uses and harms of deepfakes and advancements in deepfake detection technology. But a Commission could also be charged with reporting on and assessing existing legal frameworks for addressing and seeking redress for other harms.

[45] *See e.g.,* Sam Altman, Oversight of A.I.: Rules for Artificial Intelligence Hearing before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law, 118th Cong. (2023), https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence; Sundar Pichai, "Why Google thinks we need to regulate AI", *Financial Times*, Jan. 20, 2020, https://www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04 (CEO of Google stating that "there is no question in [his] mind that artificial intelligence needs to be regulated"); Brad Smith, "Meeting the AI moment: advancing the future through responsible AI", Microsoft, Feb. 2, 2023, https://blogs.microsoft.com/on-the-issues/2023/02/02/responsible-ai-chatgpt-artificial-intelligence/ (Vice Chair & President of Microsoft calling for effective AI regulations that "center on the highest risk applications and be outcomes-focused and durable").

[46] For example, OpenAI claims that its image generator DALL-E cannot create images of public figures, and that it restricts any "scaled" usage of its products for political purposes, such as the use of its AI to send out mass personalized emails to constituents. Reporters testing these claims have found significant exceptions and workarounds. Robust, well-tested and publicly disclosed content policies form an important aspect of safety testing. Alexandra Ulmer and Anna Tong, "Deepfaking it: America's 2024 election collides with AI boom", *Reuters*, May 30, 2023, https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/.

users about the systems' limitations and risks of inaccurate results;[47] safeguarding systems against security threats, and more.

Governments in different countries are pressing companies on what these steps should look like.[48] Whether or not these steps are ripe for legislation, Congress can play a role in driving forward these efforts – and, most critically, ensuring they are not taking place behind closed doors with only companies in attendance, but instead with meaningful participation from civil society and independent sources of expertise.

4) **Scaling agencies' capacity to address deepfakes and boost authentic sources of information.** It has long been said that the best remedy to combat undesirable speech is counterspeech[49] – but in our cacophonous information ecosystem, it takes work for counterspeech to be effective. There are steps policymakers can take to mature the systems that can help individuals better understand content authenticity and identify reliable sources of information. As one step, the government could increase funding and other efforts to support the development of technologies that assist in deepfake detection.[50] Policymakers could also support and foster awareness of voluntary efforts to authenticate content, funding research projects through the National Science Foundation and other programs, or raising awareness of private sector efforts to encourage the quick development of such work.[51]

   Critically, Congress and the Administration should significantly ramp up efforts to equip key institutions so they can identify and debunk manipulated content that threatens national security, financial markets, election administration, public health and similar priority areas. The bipartisan Deepfake Task Force Act proposed last Congress provides a good bipartisan foundation from which to start. That measure directed the creation of a task force comprised of government and non-government experts to "investigate the feasibility of, and obstacles to, developing and deploying standards and technologies for determining digital content provenance", and created "a formal mechanism for interagency coordination and information sharing to facilitate the creation and implementation of a national strategy to address the growing threats posed by digital content forgeries."[52]

---

[47] Michal Luria, "Your ChatGPT Relationship Status Shouldn't Be Complicated", *WIRED*, Apr. 11, 2023, https://www.wired.com/story/chatgpt-social-roles-psychology/.

[48] Ryan Browne, "With ChatGPT hype swirling, UK government urges regulators to come up with rules for A.I.", *CNBC*, Mar. 29, 2023, https://www.cnbc.com/2023/03/29/with-chatgpt-hype-swirling-uk-government-urges-regulators-to-come-up-with-rules-for-ai.html; Ryan Browne, "Europe takes aim at ChatGPT with what might soon be the West's first A.I. law. Here's what it means", *CNBC*, May 15, 2023, https://www.cnbc.com/2023/05/15/eu-ai-act-europe-takes-aim-at-chatgpt-with-landmark-regulation.html. In the U.S., the White House issued the AI Bill of Rights in October 2022 and the National Institute of Standards and Technology (NIST) followed in January 2023 with an AI Risk Management Framework, and officials have spoken about ways in which these map onto the risks posed by generative AI. *See* Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/ostp/ai-bill-of-rights/; National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[49] *Whitney v. Cal.*, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring) ("If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence.").

[50] *See, e.g.,* IOGAN Act, Pub. L. No. 116-258 (2020), directing the National Science Foundation and the National Institute of Standards and Technology (NIST) to support research on generative adversarial networks. The proposed American Data Privacy & Protection Act of 2022 would have required the Secretary of Commerce to publish an annual report on common sources of digital content forgeries, an assessment of the uses, applications and harms of digital content forgeries, and an analysis of the methods and standards available for detection and counter-measures such as labeling. American Data Privacy & Protection Act of 2022, H.R. 8152, Section 305, 117th Cong. (2021-2022).

[51] Shirin Ghaffary, "What will stop AI from flooding the internet with fake images?, *Vox*, Jun. 3, 2023, https://www.vox.com/technology/23746060/ai-generative-fake-images-photoshop-google-microsoft-adobe.

[52] Section 5709 of the National Defense Authorization Act of 2020 also took steps to improve government agency awareness and competency to address deepfakes. It directed the Director of National Intelligence to produce a report on the technological capabilities of foreign actors with

Capacity-building efforts could also include funding training, resources and using oversight pressure to ensure public institutions take steps to best earn public trust when they speak out. To give one simple example, research by my organization, the Center for Democracy & Technology, revealed that only in 1 in 4 official election websites uses the trusted ".gov" domain managed by DHS, while other election officials use ".com" domains that can be easily spoofed. The result is to undermine the role of such websites as a source for people to access trusted information about the administration of elections. Funding, education and oversight could help election officials address this simple vulnerability.

Conclusion

The examples of face recognition and misleading information about elections show two very different ways in which AI is already impacting Americans' human rights and the structure of our democracy. Critically, these examples show that there are concrete steps policymakers can take, today, to address the potential harms that can arise from certain uses of AI. As commentators around the world assess the existential threats posed by AI systems, it is important to remember that existential threats can also include threats to the fabric of society: undermining individual rights, equality and economic mobility, and an informed public discourse that is the bedrock of a functioning democracy. On many of these issues, there are steps that technology companies, regulatory agencies and Congress can take right now to address and reduce AI-driven harms. Thank you for the opportunity to share these thoughts today.

respect to "machine-manipulated media, machine generated text, generative adversarial networks, and related machine-learning technologies", and analysis of the counter-technologies that have been or could be developed and deployed to address such uses, among other factors. National Defense Authorization Act of 2020, Pub. L. No. 116-92 (2019).

Subcommittee on Human Rights and the Law

Hearing: "Artificial Intelligence and Human Rights"

Written Statement of Aleksander Mądry[1]

June 13th, 2023

Chairman Ossoff, Ranking Member Blackburn and Members of the Committee, thank you for inviting me to testify. Much has already been said and written about how AI may transform society, both about the opportunities and risks—from AI's potential to enhance our productivity, creativity, and overall quality of life to its ability to perpetuate discrimination, drive economic inequality, and pose an existential risk.

I will not reprise those conversations here. Instead, I will focus my testimony on one issue that I find particularly salient, time-sensitive and extremely worrisome: *how AI could erode central tenets that enable our society to function, including our ability to carry out democratic decision-making.*

Specifically, I will discuss how AI is poised to fundamentally transform mechanisms for the dissemination and understanding of information, and the unsettling implications of those changes. I will also sketch out what could be done to mitigate these emerging risks.

How will AI transform the information ecosystem?

Changes in information technologies—whether the invention of the printing press, the advent of e-mail, or the emergence of social media—do not just make information more accessible, they fundamentally change the dynamics of information sharing and acquisition. While we are still dealing with the transformations in this space brought to us by email and social media, there is already a new transformation afoot—a transformation fueled by recent developments in AI that is likely to be more consequential than anything we have been experiencing recently.

With the advent of AI—especially the newest wave of generative AI—anyone who can use a chatbot is in a position to become a "trusted source"—a *highly personalized* source, in fact. Indeed, as more of what we see becomes generated and disseminated by AI, the lines between humans and bots are becoming blurred. We need to start to be more wary than ever about how information reaches us, its trustworthiness and its ability to persuade us.

More precisely, AI is changing the information-delivery landscape in three key ways:

(a) It enables the creation of content—written text, photos and, soon, videos—that seems extremely realistic.

(b) The language produced by Large Language Models (LLMs) like chatGPT or Google Bard can seem natural and highly persuasive, in no small part since we are wired to believe that such speech can come only from humans.

(c) It makes the creation of such content cheap and broadly accessible—even to parties with little if any technical expertise—making it frighteningly easy to deploy it at scale.

[1] I have recently started a professional leave from MIT, which I am spending at OpenAI. I am providing this testimony solely in my personal capacity and as an MIT faculty. I am not in any way representing OpenAI.

We are already seeing early adoption of generative AI in our information sphere, from art [10], to copywriting [7], to political ads [13], but these are just a tip of the iceberg. We will see much, much more very soon. The onset of this technology brings with it a whole spectrum of risks and potential harms. I will highlight just a few of them below.

Enhancing "traditional" cybercrime. One immediate impact of the newest wave of generative AI is that "traditional" spam and phishing campaigns are even easier to conduct. What previously required careful photo editing and writing (as well as some non-trivial human involvement) now only requires a few clicks. The recent use of an AI-generated fake image of a fire near the Pentagon is just one illustration of that [9].

Also, the fact that generative AI can convincingly impersonate a human online poses a fundamental challenge to our existing mechanisms for protecting our information infrastructure, public discourse and governance. After all, the bot detection and moderation algorithms that our online discussion platforms use—whether they be Internet forums, newspaper comment sections, or Twitter—tend to rely on some kind of "prove that you are human" tests. How will these platforms cope with malicious parties that can field swarms of sophisticated, AI-driven bots that are able to breeze through such tests?

"Spear-phishing" and personalized blackmail. The enhancement of the "traditional" deception is, however, just the beginning. AI's unique ability to create content that is both convincing and personalized means that, for example, phishing will no longer need to involve generic emails sent out to thousands of recipients, hoping someone will get duped. Instead, we will have "spear-phishing," where both the message *and* the whole conversation that ensues are *fully automated* and *customized* to you.

In fact, there is a very real possibility that a new kind of blackmail scheme will emerge. In such a scheme, someone's photo from social media is edited to depict them in a compromising situation, and then they are threatened that the edited photo will be made public unless they pay up. How many of us would not pay to simply make the problem go away? Thanks to AI these kinds of schemes can now be executed (again) *fully automatically*, *cheaply* and *at scale*.

In addition—as one of the other witnesses has experienced herself [3]—the AI-fueled ability to impersonate the voice of just about any person enables a whole new array of scams [12]. As the ability to generate video with AI improves, other troubling possibilities such as targeted AI-generated explicit content [2] will become an even more acute problem too.

Personalized persuasion at scale. This expansion of the cybercrime toolkit is hardly the only worrisome consequence though. Indeed, AI is bound to transform how we think about any information campaign—be it ideological, political or commercial. Specifically, such campaigns will no longer need to rely solely on the promoted message to go viral. Instead, they can be fielded with generative AI and the promoted messaging might reach its intended audience *individually* and in a *highly personalized* manner. So, it will not be about some post that came across your social media timeline. Rather it will be about a Facebook "friend" that you met online. Friend who is actually an AI-driven agent impersonating a human. Friend that only subtly weaves in political commentary or product endorsements or any other messaging in between your engaging conversations about sports, movies or favorite food.

Similarly, instead of trying to corral a critical mass of people to campaign for a cause—whether on social media, via direct calling, or letter-writing—a single actor can field a campaign by themselves, using generative AI-driven bots in place of people. A campaign that is *equally effective* (thanks to the sophistication of these bots) but needs neither any buy-in from the broader population nor even comparable resources. As far as I know, as of now, this would all be legal too.

Automated creation of addictive content. AI doesn't just produce content that mimics reality and appears human-like and personalized—it can also make this content *personable*. There is a lot of information about our habits, preferences, hobbies and values that can be gleaned from sources such as our social media accounts. This could make interacting with AI not only attractive and persuasive but also addictive to us. After all, loneliness and an unmet need for some kind of intimacy with others are a growing problem in our society [8], and the kind of focus, "fit" and "care" such AI-driven "friends" would seem to exhibit could be extremely alluring.

This aspect of AI could (and, I hope, will) play a positive role too [1]. But imagine the power someone who is able to deploy such AI-powered agents could have over us, especially at scale. What if that power gets abused? What if these capabilities are harnessed to supercharge the "attention economy" that already drives much of our social media and online commerce? What would this mean for our productivity and long-term happiness? How do we feel about having our children being exposed to all of that?

Eroding trust in information and written (or audio-visual) records. Thanks to AI we are entering the era when *any* record could plausibly be faked. How does this affect our collective discourse as well as the legal and governance system? After all, we are a society whose foundations rely on the veracity and binding of such records—think contracts, deposition recording, or video evidence in criminal cases—and this reliance will only increase as more of our critical interactions occur in the digital sphere. How does our society adapt to such a tectonic shift?

What can we do?

The concerns I have outlined above may paint a rather bleak and, potentially, daunting landscape. But there is much we can (and should) do here. Specifically, we need a combination of technical solutions and policy actions that will reinforce each other. After all, policy can help drive the development and implementation of technical remedies, and technical innovations can, in turn, unlock new policy options. Let me describe some of these below.

Technical solutions

On the technical front, we need tools that can help humans judge the authenticity of content—to understand the extent to which it was generated by a human and/or AI. These tools can take a variety of forms (and for many of them we already have proof-of-concept prototypes):

Watermarking and deepfake detection tools. One promising idea for ensuring the authenticity of content is "watermarking"—that is, placing an imperceptible "signature" in generated content that makes clear AI was used. This watermark can then be detected by any content consumer. Researchers have developed prototypes of watermarking systems, both in the context of large language

models [4] and image generations models [14]. Much more work is needed, however, to make them sufficiently robust and then policies might be needed to drive their adoption too. Also, like all such technologies, there would likely be an "arms race"—tools will be developed to evade the watermark system and improved techniques will be needed to respond to that.

Watermarks need to be placed in documents directly by the AI providers, but there is also a line of work on detecting AI-generated content in the absence of cooperation from the developers of a given AI model [6]. Of course, this lack of cooperation makes it easier for malicious actors to thwart these detection techniques, causing the corresponding "arms race" to be much more challenging.

Protection against unauthorized AI-powered content editing. Another problem that technology can help address is unauthorized AI-powered content editing—that is, the ability to use AI-powered editing tools to manipulate content against the wishes of its creators or people depicted in it. (Think, for example, of the personalized blackmail scheme described earlier, which involved a malicious party manipulating photos the victim had published on social media.) Could we develop a way for users to protect the photos they put online, to make it impossible—or, at least, much harder—to modify using AI? It turns out that such an "immunization" capability is a possibility [11] but, again, much more work is needed.

Provenance certification techniques. Beyond detecting AI-generated content, tools may be needed to *prove* the authenticity of content. This could involve, for example, leveraging cryptographic tools to provide automatic certification of the authenticity or provenance of a given document by tracing it to the exact primary source that created it (e.g., the person who took a given photo). When such a technology is broadly available, content might be presumed to be fake unless verification proves it to be real.

However, just to reiterate: no matter how work on such tools proceeds, these tools will *not* be a panacea. They will be neither perfect nor foolproof, either—that is not technically possible. Nonetheless, these tools can provide the necessary "friction" that makes undesirable use of AI that much harder to execute and they will also create "footholds" for the policy action.

Policy solutions

As I noted above, technological approaches will need to work hand-in-hand with policy. Here are some possible policy approaches to pursue.

AI-generated content disclosure requirement. One relatively straightforward step would be to require that any consumer-facing AI-generated content be labeled as such. This kind of mandatory disclosure would, for example, likely hamper an AI-powered persuasion campaign we described above—at least, as long as this rule was abided by.

Of course, deciding the exact level of AI involvement that would trigger such a mandate—as well as the form it would need to take—would require careful deliberation. And the rules would have to be updated as the technology and the use of it evolved. In particular, it would be important to avoid the "user desensitization" effect, in which the users stop paying attention to the corresponding disclosures due to being bombarded with them at every occasion (and for trivial reasons). (Such desensitization seems to have occurred, for example, in the context of the web cookie usage disclosure

and consent requirements imposed in the European General Data Protection Regulation (GDPR) [5].)

Accelerating the use of content authenticity tools. As discussed earlier, content authenticity tools such as watermarking, deepfake detection, protection against unauthorized AI-powered editing, or provenance certification can be very useful but their effectiveness is hardly guaranteed. Even leaving aside technical questions, the efficacy of these solutions will critically depend on how broadly adopted they are. We need here a broad cooperation of the industry players that develop the relevant AI systems, so as to establish consistent expectations and standards. Policy can accelerate this process and broaden the use of such techniques, through incentives and/or mandates. After all, we don't know if market incentives will ever be sufficiently strong to drive the development and deployment of these technologies; they certainly are not enough at this point.

Client identification and suspicious activity reporting mandates. One possible approach to deterring rogue actors could be adapted from anti-money laundering laws. It would require providers of sufficiently capable AI services to implement adequate client identification mechanisms. These AI providers would then be expected to monitor the usage of the tools they supply to flag (and, potentially, block) suspicious activity as well as to report it to appropriate governmental agencies (such as FBI) or other organizations.

Advance "AI literacy" efforts. Of course, no technical solution or set of regulations will ever suffice to fully mitigate the risks AI now poses. It is thus crucial that, in addition to "email literacy" and "social media literacy," we think about promotion of "AI literacy." The public needs to understand how to judiciously interact with AI systems—and how to be on the lookout for when they are interacting with AI in the first place. This includes helping the public avoid the natural tendency to anthropomorphize AI systems. After all, AI does not reason; it merely mimics reasoning—at least as of now. We also must go from assuming that content is authentic until proven otherwise to assuming that content is fake until proven otherwise—or at the very least discounting the value of unverified content.

Overall, there is a need for a shift in the public mindset to accommodate how AI is changing the world. We thus need a decisive policy thinking on how to advance such AI literacy more intentionally, instead of relying on our society learning it the "hard way."

To conclude, let me reiterate that I am excited about the positive impacts that AI can have, but I also want to be clear about and mindful of the risks it gives rise to. Today, my aim is to highlight one family of such risks. I am optimistic that we can mitigate these risks, but this will require work. It cannot be left to chance. And we need to get started now.

Thank you and I am looking forward to your questions.

Acknowledgements

I am grateful for invaluable help from Sarah Cen, David Goldston, Andrew Ilyas, and Luis Videgaray.

References

[1] Sai Balasubramanian. AI offers promise and peril in tackling loneliness. *Forbes*. https://www.forbes.com/sites/saibala/2023/05/17/can-artificial-intelligence-solve-the-growing-mental-health-crisis/.

[2] Karen Hao. Deepfake porn is ruining women's lives. Now the law may finally ban it. *Technology Review*. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

[3] Faith Karimi. 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. *CNN.com*. https://www.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html.

[4] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *Arxiv preprint arXiv:2306.04634*, 2023.

[5] Oksana Kulyk, Nina Gerber, Annika Hilt, and Melanie Volkamer. Has the GDPR hype affected users' reaction to cookie disclaimers? *Journal of Cybersecurity*, 6(1), 12 2020.

[6] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes. In *Arxiv preprint arXiv:2004.11138*, 2020.

[7] Johan Moreno. Canva opens up access to docs in beta, adds "magic write" generative AI copywriting tools. *Forbes*. https://www.forbes.com/sites/johanmoreno/2022/12/07/canva-opens-up-access-to-docs-in-beta-adds-magic-write-generative-ai-copywriting-tools/.

[8] Vivek H. Murthy. Our epidemic of loneliness and isolation. *The U.S. Surgeon General's Advisory*. https://www.hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf.

[9] Donie O'Sullivan and Jon Passantino. 'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion. *CNN.com*. https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html.

[10] Kevin Roose. An A.I.-generated picture won an art prize. artists aren't happy. *New York Times*. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html.

[11] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Mądry. Raising the cost of malicious ai-powered image editing. In *Arxiv preprint arXiv:2302.06588*, 2023.

[12] Pranshu Verma. They thought loved ones were calling for help. It was an AI scam. *The Washington Post*. https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/.

[13] James Vincent. DeSantis attack ad uses fake AI images of Trump embracing Fauci. *The Verge*. https://www.theverge.com/2023/6/8/23753626/deepfake-political-attack-ad-ron-desantis-donald-trump-anthony-fauci.

[14] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. In *Arxiv preprint arXiv:2305.20030*, 2023.