Center for AI Policy

Report on Misinformation from AI

Misinformation from AI[1] is a growing concern among the American public, as it holds the potential to empower authoritarian regimes and make it harder to distinguish truth from falsehood, genuine from artificial, and human from bot. Experts are concerned too, as the World Economic Forum has gone so far as to say that misinformation is society's biggest risk over the next two years. As such, this report aims to orient you to this problem, first taking stock of some of the damage already done, and then looking at how we can address misinformation from AI moving forward.

The Harm

It's late one evening and you get a call from an unknown number, normally letting it go to voicemail, but answering this time because your daughter is out of town on a ski trip. When you answer, you hear "Mom!" and your daughter sobbing. The only other thing you can make out amidst the distraught sobbing is her saying "I messed up."

Then you hear a man take control of the call. He tells you that he has your daughter, demanding $1 million and threatening to hurt her if you call the cops. Meanwhile, in the background you hear your daughter, sobbing and repeating "help me Mom, please".

This was Jennifer DeStefano's reality recently, a situation thankfully quickly resolved as she was surrounded by friends who were able to quickly get in contact with her daughter. Calls like these are happening increasingly frequently now, largely thanks to improvements in AI voice cloning technology, which now requires as little as three seconds of someone's speech to create a convincing replica.

The harms and damages are real, and below we'll walk through more examples, what they mean, and how we might better prevent them moving forward.

[1] In this report, AI misinformation refers to two possible uses of artificial intelligence (AI). One: creating content that conveys a false or misleading message, regardless of intent. Two: using bots to manufacture an appearance of support or opposition (e.g. Russia's use of bots to tweet support for the idea that the murder of the leader of the opposition party was really an inside job by that party). Note that academic researchers make more nuanced distinctions when defining "misinformation," "disinformation," and so on.

Creation of Misinformation

Text Examples

- In 2023, NewsGuard, as part of an ongoing analysis, found over 580 unreliable AI-generated news sites making claims like "Netanyahu's psychiatrist has committed suicide and the suicide note implicated Netanyahu."

Image Examples

- In 2023, an image of an explosion outside the Pentagon circulated widely, even making it to the White House press office, causing the stock market to drop even when the image still had noticeable flaws.

- In 2023, synthetic nude images were shared of female students at multiple (one, two) US high schools and of underage girls in a Spanish town.

- In 2023, synthetic images of the Turkish earthquake were used to scam people for donations.

Audio Examples

- As of 2020, at least three cases of using voice-mimicking for major theft have occurred, often involving impersonations of senior management and making away with $240,000, $10,000,000, and a staggering $35,000,000.

- In 2023, the altered audio of a 15-year-old was used to give the appearance of a kidnapping to try to extort the family.

- In 2023, synthetic audio was used to gain access to multi-factor authentication codes as part of a phishing scheme which successfully breached an IT company.

- In 2023, days before an influential Slovakian election, synthetic audio was released of the leader of the Liberal Progressive party talking about how to rig the election with a newspaper executive. This likely played a role in the pro-Kremlin party going on to win the election.

Video (Deepfakes) Examples

- In 2021, a ring of extortioners in India were exposed who used threats of DMing family members and friends videos of deepfakes porn to target women.

- In 2022, a deepfake circulated of Volodymyr Zelensky calling for surrender and his troops to put down their weapons.

- In 2023, a study uncovered a network of TikTok accounts combining synthetic and real video to push conspiracy theories like Jamie Foxx becoming paralyzed and blind as a result of the COVID-19 vaccine, garnering over 336 million views and 14.5 million likes.

- In 2023, former Pakistan Prime Minister Imran Khan shared a deepfake depicting a woman "fearlessly facing riot police" in an effort to give the impression that Pakistani women stood with him.

- In 2024, scammers swindled a multinational corporation out of $25,000,000, using deepfake video conference calls.

Why worry?

We'll likely see present harms carried into the future, where text will be used to scale phishing, audio will be used to run scams via impersonation, and deepfake porn will be used to degrade and extort. As these technologies become more accessible and more effective, we'll likely see increases in both the quality and quantity of the above types of harm.

But perhaps most worrying is the potential these technologies carry to erode our sense of truth. Sight and sound online can no longer serve as reliable indicators of truth, given our lack of ability to consistently discern synthetic from normal media (e.g. video, audio, text[2]). And this is all the more concerning when you consider that synthetic media is likely only to increase in presence, in an environment where people already struggle to discern what's real from what's fake. And as the technology gets further and further sophisticated, we'll increasingly find expert disagreement[3] on whether something is synthetic or not, creating further space for different "truths" to develop. The existence of deepfakes has already been used in the courtroom to call real evidence into question, and we can expect this lack of trust to only further drive distrust across American society, not just in what we see but between one another as well.

[2] Worth noting, it's not just lay people that struggle to identify synthetic text: expert linguists struggle too.
[3] In 2023, after a member of the ruling Ethiopian Prosperity Party was killed, state-affiliated media released audio linking a militia group to the killing. An organization promptly identified it as synthetic, only to have a fact-checking group accuse that organization of being fake.

Misinformation via Bots

Examples

- In 2011, an estimated 25,860 bots tweeted 440,793 times in an effort to drown out anti-Kremlin hashtags following the disputed 2011 Russian parliamentary elections.

- In 2011, bots likely operated by the Syrian state were used to drown out content about the Syrian protests and government crackdown, mostly by tweeting innocuous material continuously with #Syria.

- In 2015, the Russian Internet Research Agency (RIRA) conducted a misinformation campaign to suggest that Boris Nemtsov's (the Russian opposition leader at the time) murder was an inside job by the opposition.

- In 2014, bots created buzz around Cynk, a penny stock, causing a 200x increase in market value, and then a subsequent crash which lead to large losses.

- In 2016, Russian cyborgs[4] undertook a mission to influence the US election, degrading both Democrat and Republican candidates.

- In 2017, bots were used to smear and intimidate Mexican journalists, tweeting things like "@albertoescorcia the boss has seen your article where you say bad things about him, say goodbye to your family, you're a dead man."

- In 2017, the FCC was flooded with fake comments, millions of impersonations achieved by just one college student, part of a broader trend across comments on federal regulation.

- In 2020, Russian and Iranian cyborgs were deployed to spread misinformation about COVID-19 by claiming e.g. that the idea that COVID was a cover story to cover up the harmful effects of 5G.

- In 2021, Twitter identified and removed 2,048 CCP connected accounts tweeting support for CCP narratives about the quality of life for Uyghurs in Xinjiang.

- In 2022, Meta and Twitter removed multiple anti-Ukraine bots tied to the Russian state.

[4] Cyborg is used to refer to accounts that blend bot activity but still some management by humans

- In 2023, numerous bots were deployed with content favoring current president Erdoğan and his AKP party in an attempt to influence the 2023 Turkish presidential election, a close race which Erdoğan went on to win.

Why worry?

Though each risk is concerning, the most harmful effects may come from use by authoritarian governments. We've documented 16 such governments using bots and synthetic media to distort information on social or political issues domestically, but such bot activity has been detected in over 57 countries, both as a product of increasing domestic use and as foreign interference. Such tactics have been used by governments from Russia to Syria to China to suppress dissent in times of turmoil and craft public opinion to their advantage. They have also contributed to democratic backsliding in countries like Turkey where President Recep Erdoğan has used bots to help maintain his grip over a country that came close to voting him out.

If these trends are allowed to continue, and if we stay on our current trajectory of development, then we're unlikely to develop technical safeguards in time to be able to effectively address the bots we are likely to face. We already struggle to tell the difference between a bot and a human, a problem which will become even worse as bots become more effective. Moreover, we should expect more, not fewer, of these bots in the future as they become cheaper to produce and deploy, bringing us year by year closer to platforms full of bots capable of dynamic back-and-forth conversations that you're unlikely to identify as synthetic. Combine this ability with bots' potential future ability to deploy AI-generated content, and it's not hard to imagine a truly disorienting social media environment. Despite the veneer of a genuine seeming exchange, a subtle attempt to influence you could lie behind every post, every comment, every message; this environment of suspicion wouldn't degrade just the experience itself, but likely our trust in one another too.

Public Perception

- For the public, the biggest AI-related concern is the spread of misleading synthetic content.

- 76% are concerned about misinformation from frontier AI models (like ChatGPT), and only 54% think they can tell the difference between real and synthetic media.

- 76% of the public is concerned with deepfakes driving misinformation, which they perceive to be the most "real" threat from AI.

Policies Directed at the Issue

Existing

The non-comprehensive list below outlines some of the current proposed legislation targeting misinformation from AI. Generally, these acts shine when they focus on further investigation into how we can best address the harms of synthetic media, and when they introduce liability for malicious uses, helping establish a proper incentive structure away from such actions. Among current legislation, the DEEPFAKES Accountability Act especially stands out for combining both of these strengths into one bill.

- H.R.5586, **DEEPFAKES Accountability Act**: mandates both content provenance (enabling technical identification of the origin of a given piece of media) and clear labeling of synthetic media (including identification of what was altered specifically), providing legal recourse against malicious, unlabeled impersonations.
    - The Attorney General will develop the technical specifications for digital provenance. Developers will be required to make their outputs digital provenance compliant, online platforms required to do the same for their platforms (and establish a method for detection of synthetic media).
    - It also instructs the Departments of Homeland Security to establish a "Deepfake Task Force" which will research and develop methods to detect ("or otherwise counter and combat") synthetic media.

- H.R.6943**, No AI FRAUD Act**: similar to the No FAKES Act, this grants individual's property right to their voice and visual likeness, making publication and distribution of all forms of harmful, non-consensual synthetic media (video, picture, audio) liable to pay $5,000 per violation in damages.
    - Going beyond the No FAKES Act, this bill also includes a penalty of $50,000 for distributing or making available "personalized cloning services" capable of creating such non-consensual synthetic media.

- H.R.5808 **Preventing Deep Fake Scams Act**: establishes a task force to study the impact of AI on fraud, and make proposals on what regulation or legislation can best protect consumers from AI fraud.

- H.R.3106 **Preventing Deepfakes of Intimate Images Act**: makes non-consensual deepfake porn illegal, punishable by up to two years in prison.

- S.2765 **Advisory for AI-Generated Content Act**: tasks the FTC with issuing regulations to establish standards for watermarking AI-generated material. These standards would then become a requirement for any entity that produces AI-generated content.

Potential

If we are to fully address the risk of misinformation from AI, we'll need to go further. The above acts provide a great starting point, but many are betting on a world where identification and origin of AI-generated content are easily discernible. However, years of development hasn't led to reliable methods of identification[5], a task that will likely grow harder as models become more capable.

Further, new bills are needed to address misinformation via bots. To fully neglect this category will mean unchecked proliferation of bots which can manipulate public opinion, erode trust in institutions, and undermine democratic processes. We focused above on how bots can be used by authoritarian governments to amplify government propaganda and suppress dissent, but the effects will be felt at home too, as bots can sway elections, influence public discourse, and erode trust in elected officials.

Below, we propose policies that would help America build up the systems needed to address this risk head on. These policies can be complementary to current efforts focused at reducing the risks of AI and creation of misinformation and more directly address the issue of misinformation via bots. This is an incomplete set of policies, and as such our first suggestion is aimed at growing knowledge of gaps and opportunities for properly addressing the risk.

- **Establish a task force to study ways to address misinformation via bots.** Given the current uncertain feasibility of many interventions to reduce AI-driven misinformation, a task force would be a prudent way for the government to take actionable steps to reduce this particular risk without betting on a strategy that hasn't yet been confirmed as an effective intervention. Such a task force could be especially effective if it's able to monitor the other sorts of interventions which could be developed and tested over a year, giving it the option to confirm effective solutions or narrow the space by identifying ineffective interventions.

- **Fund evaluations for misinformation capabilities**. The US Government should play an active role in the space of evaluations, which are experiments to test AI capabilities. In particular, government agencies like NIST and the DOE

[5] See OpenAI's attempt at detecting AI-written text, which failed about 75% of the time; or see DARPA's Medifor project, which concluded without generating an effective classifier algorithm for photos and videos.

should develop ways to probe powerful models for misinformation and deception capabilities before release. Congress should ensure these agencies have adequate funding for such efforts, some of which were ordered by the recent White House AI Executive Order.

- **Require fingerprints for model outputs**. Congress should require embedding of "digital fingerprints" into model outputs, like Google's SynthID, which are designed to leave imperceptible traces in synthetic content (image, audio, video, text[6]) that doesn't affect the quality of the generation but is still identifiable through technical means. Lawmakers should pass a bill like the Advisory for AI-Generated Content Act to accomplish this.
  Furthermore, lawmakers should encourage further innovation in research that strengthens these fingerprints against current methods of evasion, such as paraphrasing attacks for synthetic text. Congress should open up competitions which award the most enduring, tamper-proof, fingerprinting techniques. Sec. 1543 of the FY2024 NDAA, which establishes a prize competition for detecting and watermarking AI outputs, is an excellent example of the kinds of bills we need.

[6] SynthID currently just functions for images and audio, but there are other fingerprinting efforts targeted at text, and video could likely be identifiable by some combination of methods used for audio and images.