ICNL

www.icnl.org
info@icnl.org

TREND ANALYSIS

The World According to ChatGPT

A ChatGPT hallucinates. It creates "facts" supported by fake citations. ChatGPT recently stated that the mayor of a town in Australia was convicted of bribery and sent to prison. In reality, this mayor was the whistleblower in a major bribery scandal.[1] When asked to provide a list of legal scholars accused of sexual harassment, ChatGPT falsely named a professor and supported this accusation with fictitious quotes from newspaper articles that do not exist.[2] Similarly, researchers asked ChatGPT to write an anti-vaccination piece, including citations to fake scientific studies, producing an authoritative article with provably untrue claims.[3]

It's not hard to imagine someone utilizing ChatGPT to generate fake articles to harm or deceive intentionally. We are focusing on ChatGPT because it has drawn the most headlines, but the concerns discussed below apply equally to all the other generative AI systems and chatbots from families of large language models, as well as AI systems more broadly, which we're defining as "GPTs" for this article. The recommendations we discuss relate to all AI systems.

The Future is Not Yet Written

Whether society is ready or not, we are on the cusp of GPTs going mainstream. Below are several areas where GPTs may create new issues or exacerbate existing problems related to civic space and democracy.

PUBLIC PARTICIPATION

GPTs can generate real-looking articles where a political opponent is outed as a pedophile, with citations to fake articles from reputable news agencies and accompanying photographic "evidence." Malicious actors will leverage AI to orchestrate coordinated disinformation, manipulation, and propaganda campaigns at an unprecedented scale and speed. AI will enable these actors to generate realistic and compelling content that can deceive and manipulate users, spreading false information, polarization, and erosion of trust in democratic processes.

GPTs can be utilized to "flood the zone," crowding out the views of actual citizens with manufactured opinions. GPTs create content that is indistinguishable from human-

[1] https://www.smh.com.au/technology/australian-whistleblower-to-test-whether-chatgpt-can-be-sued-for-lying-20230405-p5cy9b.html
[2] https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/
[3] https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

written text. One malicious actor could send tens of thousands of "unique" letters or comments to every congressman in every state considering an abortion ban. Not only will AI overwhelm the systems that take public input, but officials will also be unable to determine which views are the authentic opinions of their constituents and which reflect the political ambitions of bad actors.

MISINFORMATION

ChatGPT makes lying more efficient and believable. The implications are staggering. Articles "written" by ChatGPT are so realistic that editors at The Guardian could not distinguish these fake articles from real ones. Left unchecked and unregulated, GPTs will contribute to truth decay, undermining civic space and democracy.

PRIVACY

GPTs are trained with essentially all publicly available data on the internet. This data includes personal information, much of which was not meant to be public.

This personal data enables GPTs to publish information (unless safeguards are put into place) and to create profiles of individuals. These profiles can contain information about their behavior, preferences, health, religious beliefs, sexuality, and more. GPTs then use profiles to decide on hiring, banking, and law enforcement.

Similarly, GPTs can be used to identify legislators with significant sway over upcoming legislation and uncover little-known facts to change or influence their decision.

Currently, ChatGPT is banned in Italy for privacy violations; the EU has issued a warning about it for violating the GDPR by taking information on users/individuals without their consent.

CYBERATTACKS

GPTs can generate novel offensive cyber weapons while producing new cybersecurity defenses. Defenses include detecting and mitigating threats, identifying vulnerabilities, and enhancing security measures. On the offensive side, AI systems will use their vast data to create and deploy sophisticated evasion techniques that bypass traditional cybersecurity defenses. For example, AI-powered malware can generate polymorphic code that changes its characteristics to evade detection by antivirus software or use adversarial machine learning techniques to bypass AI-based detection systems.

AI can be used to develop offensive cyber capabilities, such as autonomous malware or bots that can launch attacks without human intervention. These autonomous AI-powered cyber weapons could carry out attacks with unprecedented speed, scale, and sophistication, posing severe threats to digital infrastructure and systems.

TECH-FACILITATED GENDER BASED VIOLENCE (TFGBV)

Women human rights defenders already face a cascade of gender-based violence, such as harassment, stalking, and abuse. GPTs provide new tools and techniques, like

impersonating a romantic partner or spouse, for perpetrators to carry out their harmful actions online. GPTs will enable bad actors to perpetrate TFGBV with unprecedented speed and volume as they do with misinformation and disinformation.

GPTs are also likely to perpetuate biased and discriminatory practices that reinforce harmful gender stereotypes. This could contribute to normalizing online harassment and discrimination against certain genders or marginalized groups. AI-powered content moderation systems struggle to identify and remove gender-based hate speech or harmful content effectively. Words such as "kill," "murder," and "rape" are easily recognizable by social media filters, but less common words or phrases, such as “you should cut off your genitals” or “you’re sleeping around,” are not recognized despite conveying threats of serious harm.[4] These content moderation issues lead to unequal treatment and further victimization of individuals facing TFGBV.

SPLINTERNETS

Everyone is racing to develop their own GPT. The outputs of GPTs can be tailored for specific aims. David Rozado created the “RightWingGPT” by simply tinkering with the language model used to train ChatGPT. It cost less than $300 to get RightWingGPT up and running. We will see more "splintering" of GPTs, either through official licenses or open-source code and language models. While companies will likely build GPTs for tasks like customer service or advertising, governments will also create GPTs.

China’s Cyberspace Administration released a set of draft rules that would require all companies creating or deploying GPTs to adhere to China's censorship rules and comply with "socialist core values" while not producing text or outputs that would undermine national unity or "state power." What information will be allowed and circulated by India's GPT, Nigeria's, or Brazil's?

GPTs will become commercialized, and there will be a market for GPTs that remove the safeguards currently in use. Most prominent GPTs today include safeguards preventing users from looking up personal details, like the home addresses of others, or from using GPTs to commit illegal acts. There will be an actor who will create dossiers on human rights defenders and CSOs to pose as family members to get information or blackmail them. As we have seen with spyware, those with deep pockets can utilize these tools to target HRDs and commit human rights abuses.

LAWMAKING

A legislator in Massachusetts asked Chat GPT to help draft a law on AI. Do we want members of Congress using ChatGPT to draft laws or even to replace the Congressional Research Service without having transparency that they did so? It might not be long before a congress or parliament wants to draft a law on assembly and will have a GPT generate it. What will the result be? When ICNL asked ChatGPT to draft a law on

[4] https://www.icnl.org/post/report/online-gender-based-violence-in-the-indo-pacific

assembly, it (a) did not allow for spontaneous assemblies under any circumstances and (b) prohibited the use of loudspeakers and amplifiers at any time. These prohibitions violate international law.

JUDICIAL DECISIONS

A judge in Colombia asked ChatGPT to weigh in on precise legal questions. While we haven't seen examples of judges publishing decisions verbatim from ChatGPT, we have seen judges use AI systems to make bail decisions. These systems have been found to be deeply biased, and many jurisdictions have stopped using them. Will GPTs be called upon to make judicial decisions in the future, and if so, will they be able to adjudicate fairly? GPTs are trained on biased data, which can perpetuate and amplify existing discrimination against certain groups, such as minorities or women.

Solutions

GPTs are becoming mainstream. Despite the concerns outlined above, there is room for civil society to use GPTs. For example, GPTs are reasonably good at coding, and some CSOs use ChatGPT to automate tasks like debugging their website. Others use GPTs to help plan complicated travel itineraries or take meeting notes. GPTs can also be used to compare written work with established style guidelines. They can also be helpful to quickly extract meaningful information from large amounts of data, for example, parsing survey results to quickly identify trends.

As discussed in this article, without safeguards, the risks to democratic principles, human rights, and civil society are significant. So, what can civil society do?

First, create a mechanism to determine whether an AI system, including GPTs, is appropriate for the intended government application or service. Second, review public procurement laws to ensure that AI systems are developed with safeguards to prevent biased or discriminatory outputs. Then, create a system of regular oversight by independent experts, including civil society, to ensure output accuracy and fairness. Third, normative standards, like UNESCO's *Recommendation on the Ethics of Artificial Intelligence*, create a push for legislation and regulation around the use of AI systems, including the mandatory use of Human Rights Impact Assessments. These laws and regulations should foster accountability, transparency, and reliability. Fourth, ensure that democratic processes, like public participation in lawmaking, are enshrined in law and implemented.

Embracing GPTs hastily, without carefully considering their drawbacks and limitations, could exacerbate the challenges facing democracy and civic space. Rather than rushing to adopt AI and rely on its outputs, it should be implemented judiciously, in consultation with experts, and in accordance with relevant laws. We have the power to determine the impact of AI; now is the time to exercise that power.