FTC probe of OpenAI: Consumer protection is the opening salvo of U.S. AI regulation

July 19 2023, by Anjana Susarla

Credit: CC0 Public Domain

The Federal Trade Commission has launched an investigation of ChatGPT maker OpenAI for potential violations of consumer protection laws. The FTC sent the company a 20-page demand for information in

the week of July 10, 2023. The move comes as European regulators have begun to take action, and Congress is working on legislation to regulate the artificial intelligence industry.

The FTC has asked OpenAI to provide details of all complaints the company has received from users regarding "false, misleading, disparaging, or harmful" statements put out by OpenAI, and whether OpenAI engaged in unfair or deceptive practices relating to risks of harm to consumers, including reputational harm. The agency has asked detailed questions about how OpenAI obtains its data, how it trains its models, the processes it uses for human feedback, risk assessment and mitigation, and its mechanisms for privacy protection.

As a researcher of social media and AI, I recognize the immensely transformative potential of generative AI models, but I believe that these systems pose risks. In particular, in the context of consumer protection, these models can produce errors, exhibit biases and violate personal data privacy.

Hidden power

At the heart of chatbots such as ChatGPT and image generation tools such as DALL-E lies the power of generative AI models that can create realistic content from text, images, audio and video inputs. These tools can be accessed through a browser or a smartphone app.

Since these AI models have no predefined use, they can be fine-tuned for a wide range of applications in a variety of domains ranging from finance to biology. The models, trained on vast quantities of data, can be adapted for different tasks with little to no coding and sometimes as easily as by describing a task in simple language.

Given that AI models such as GPT-3 and GPT-4 were developed by

private organizations using proprietary data sets, the public doesn't know the nature of the data used to train them. The opacity of training data and the complexity of the model architecture—GPT-3 was trained on over 175 billion variables or "parameters"—make it difficult for anyone to audit these models. Consequently, it's difficult to prove that the way they are built or trained causes harm.

Hallucinations

In language model AIs, a hallucination is a confident response that is inaccurate and seemingly not justified by a model's training data. Even some generative AI models that were designed to be less prone to hallucinations have amplified them.

There is a danger that generative AI models can produce incorrect or misleading information that can end up being damaging to users. A study investigating ChatGPT's ability to generate factually correct scientific writing in the medical field found that ChatGPT ended up either generating citations to nonexistent papers or reporting nonexistent results . My collaborators and I found similar patterns in our investigations.

Such hallucinations can cause real damage when the models are used without adequate supervision. For example, ChatGPT falsely claimed that a professor it named had been accused of sexual harassment. And a radio host has filed a defamation lawsuit against OpenAI regarding ChatGPT falsely claiming that there was a legal complaint against him for embezzlement.

Bias and discrimination

Without adequate safeguards or protections, generative AI models trained on vast quantities of data collected from the internet can end up

replicating existing societal biases. For example, organizations that use generative AI models to design recruiting campaigns could end up unintentionally discriminating against some groups of people.

When a journalist asked DALL-E 2 to generate images of "a technology journalist writing an article about a new AI system that can create remarkable and strange images," it generated only pictures of men. An AI portrait app exhibited several sociocultural biases, for example by lightening the skin color of an actress.

Data privacy

Another major concern, especially pertinent to the FTC investigation, is the risk of privacy breaches where the AI may end up revealing sensitive or confidential information. A hacker could gain access to sensitive information about people whose data was used to train an AI model.

Researchers have cautioned about risks from manipulations called prompt injection attacks, which can trick generative AI into giving out information that it shouldn't. "Indirect prompt injection" attacks could trick AI models with steps such as sending someone a calendar invitation with instructions for their digital assistant to export the recipient's data and send it to the hacker.

Some solutions

The European Commission has published ethical guidelines for trustworthy AI that include an assessment checklist for six different aspects of AI systems: human agency and oversight; technical robustness and safety; privacy and data governance; transparency, diversity, nondiscrimination and fairness; societal and environmental well-being; and accountability.

Better documentation of AI developers' processes can help in highlighting potential harms. For example, researchers of algorithmic fairness have proposed model cards, which are similar to nutritional labels for food. Data statements and datasheets, which characterize data sets used to train AI models, would serve a similar role.

Amazon Web Services, for instance, introduced AI service cards that describe the uses and limitations of some models it provides. The cards describe the models' capabilities, training data and intended uses.

The FTC's inquiry hints that this type of disclosure may be a direction that U.S. regulators take. Also, if the FTC finds OpenAI has violated consumer protection laws, it could fine the company or put it under a consent decree.

Provided by The Conversation