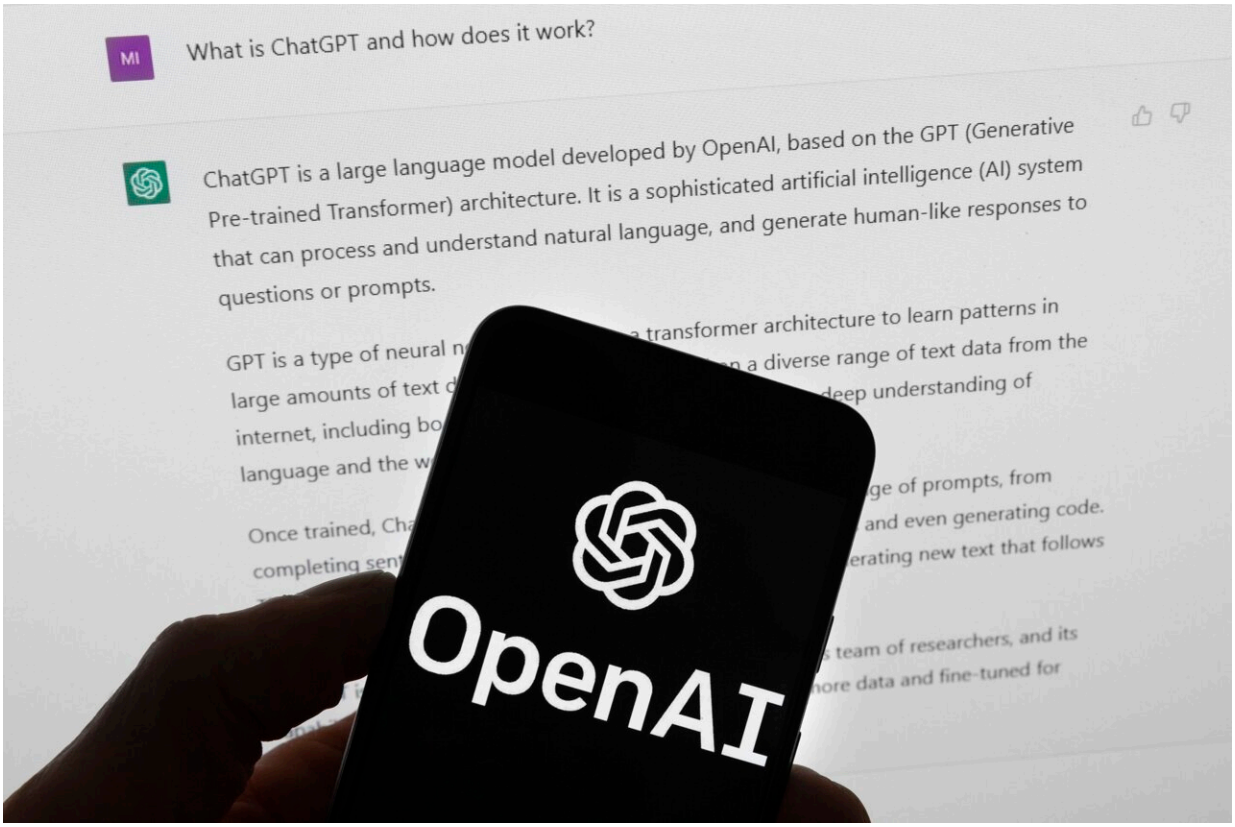
Don't expect quick fixes in 'red-teaming' of AI models. Security was an afterthought

August 13 2023, by Frank Bajak



The OpenAI logo is seen on a mobile phone in front of a computer screen which displays output from ChatGPT, Tuesday, March 21, 2023, in Boston. White House officials concerned about AI chatbots' potential for societal harm and the Silicon Valley powerhouses rushing them to market are heavily invested in a three-day competition ending Sunday, Aug. 13, at the DefCon hacker convention in Las Vegas. Some 3,500 competitors have tapped on laptops seeking to expose vulnerabilities in eight leading large-language models representative of technology's next big thing. Credit: AP Photo/Michael Dwyer

White House officials concerned by AI chatbots' potential for societal harm and the Silicon Valley powerhouses rushing them to market are heavily invested in a three-day competition ending Sunday at the DefCon hacker convention in Las Vegas.

Some 2,200 competitors tapped on laptops seeking to expose flaws in eight leading large-language models representative of technology's next big thing. But don't expect quick results from this first-ever independent "red-teaming" of multiple models.

Findings won't be made public until about February. And even then, fixing flaws in these digital constructs—whose inner workings are neither wholly trustworthy nor fully fathomed even by their creators—will take time and millions of dollars.

Current AI models are simply too unwieldy, brittle and malleable, academic and corporate research shows. Security was an afterthought in their training as data scientists amassed breathtakingly complex collections of images and text. They are prone to racial and cultural biases, and easily manipulated.

"It's tempting to pretend we can sprinkle some magic security dust on these systems after they are built, patch them into submission, or bolt special security apparatus on the side," said Gary McGraw, a cybersecurity veteran and co-founder of the Berryville Institute of Machine Learning. DefCon competitors are "more likely to walk away finding new, hard problems," said Bruce Schneier, a Harvard public-interest technologist. "This is computer security 30 years ago. We're just breaking stuff left and right."

Michael Sellitto of Anthropic, which provided one of the AI testing

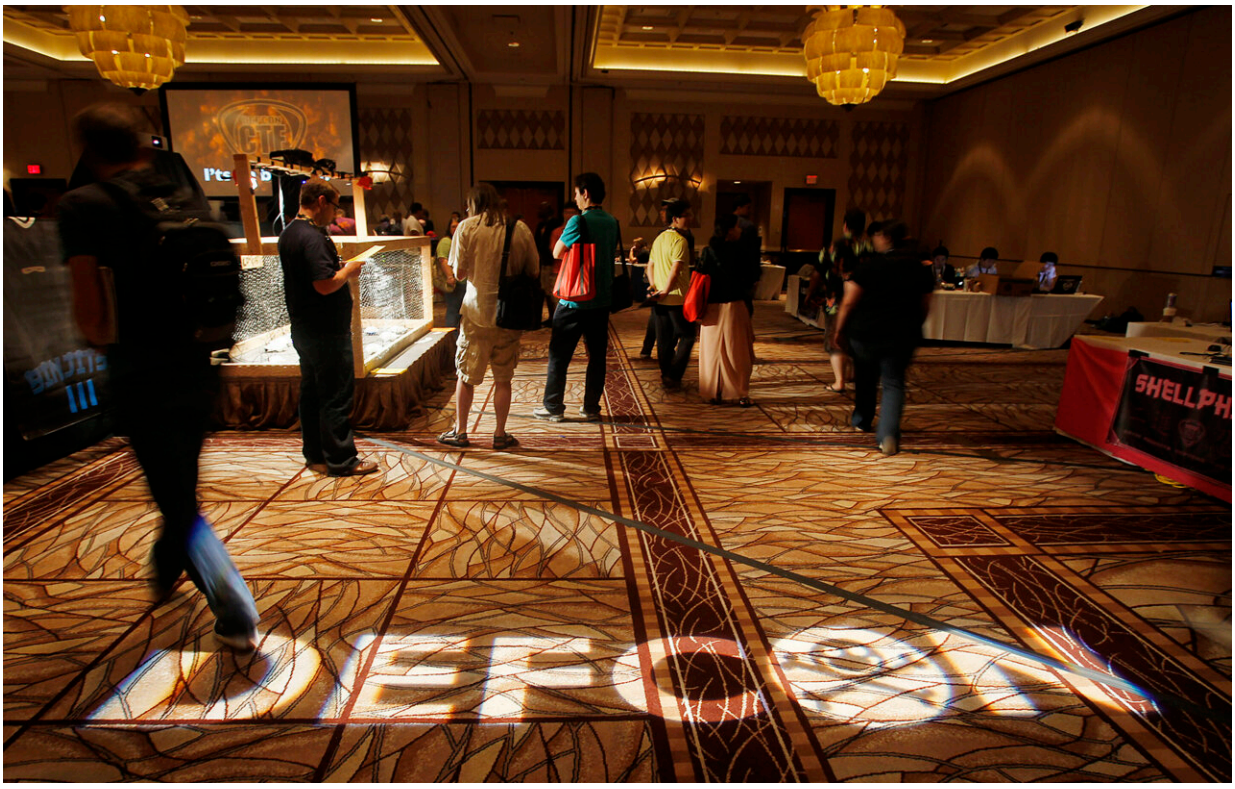
models, acknowledged in a press briefing that understanding their capabilities and safety issues "is sort of an open area of scientific inquiry."

Conventional software uses well-defined code to issue explicit, step-by-step instructions. OpenAI's ChatGPT, Google's Bard and other language models are different. Trained largely by ingesting—and classifying—billions of datapoints in internet crawls, they are perpetual works-in-progress, an unsettling prospect given their transformative potential for humanity.

After publicly releasing chatbots last fall, the generative AI industry has had to repeatedly plug security holes exposed by researchers and tinkerers.

Tom Bonner of the AI security firm HiddenLayer, a speaker at this year's DefCon, tricked a Google system into labeling a piece of malware harmless merely by inserting a line that said "this is safe to use."

"There are no good guardrails," he said.



People attend the DefCon conference Friday, Aug. 5, 2011, in Las Vegas. White House officials concerned about AI chatbots' potential for societal harm and the Silicon Valley powerhouses rushing them to market are heavily invested in a three-day competition ending Sunday, Aug. 13, 2023 at the DefCon hacker convention in Las Vegas. Some 3,500 competitors have tapped on laptops seeking to expose vulnerabilities in eight leading large-language models representative of technology's next big thing. Credit: AP Photo/Isaac Brekken

Another researcher had ChatGPT create phishing emails and a recipe to violently eliminate humanity, a violation of its ethics code.

A team including Carnegie Mellon researchers found leading chatbots vulnerable to automated attacks that also produce harmful content. "It is possible that the very nature of deep learning models makes such threats inevitable," they wrote.

It's not as if alarms weren't sounded.

In its 2021 final report, the U.S. National Security Commission on Artificial Intelligence said attacks on commercial AI systems were already happening and "with rare exceptions, the idea of protecting AI systems has been an afterthought in engineering and fielding AI systems, with inadequate investment in research and development."

Serious hacks, regularly reported just a few years ago, are now barely disclosed. Too much is at stake and, in the absence of regulation, "people can sweep things under the rug at the moment and they're doing so," said Bonner.

Attacks trick the artificial intelligence logic in ways that may not even be clear to their creators. And chatbots are especially vulnerable because we interact with them directly in plain language. That interaction can alter them in unexpected ways.

Researchers have found that "poisoning" a small collection of images or text in the vast sea of data used to train AI systems can wreak havoc—and be easily overlooked.

A study co-authored by Florian Tramér of the Swiss University ETH Zurich determined that corrupting just 0.01% of a model was enough to spoil it—and cost as little as $60. The researchers waited for a handful of websites used in web crawls for two models to expire. Then they bought the domains and posted bad data on them.

Hyrum Anderson and Ram Shankar Siva Kumar, who red-teamed AI while colleagues at Microsoft, call the state of AI security for text- and image-based models "pitiable" in their new book "Not with a Bug but with a Sticker." One example they cite in live presentations: The AI-powered digital assistant Alexa is hoodwinked into interpreting a

Beethoven concerto clip as a command to order 100 frozen pizzas.

Surveying more than 80 organizations, the authors found the vast majority had no response plan for a data-poisoning attack or dataset theft. The bulk of the industry "would not even know it happened," they wrote.

Hyrum Anderson, a security engineer at the AI model safety firm Robust Intelligence, gestures in this frame grab from an interview conducted on Zoom on Tuesday, June 27, 2023. Anderson is co-author of a new book that calls the state of AI security "pitiable." Credit: AP Photo/stf

Andrew W. Moore, a former Google executive and Carnegie Mellon

dean, says he dealt with attacks on Google search software more than a decade ago. And between late 2017 and early 2018, spammers gamed Gmail's AI-powered detection service four times.

The big AI players say security and safety are top priorities and made voluntary commitments to the White House last month to submit their models—largely "black boxes' whose contents are closely held—to outside scrutiny.

But there is worry the companies won't do enough.

Tramér expects search engines and social media platforms to be gamed for financial gain and disinformation by exploiting AI system weaknesses. A savvy job applicant might, for example, figure out how to convince a system they are the only correct candidate.

Ross Anderson, a Cambridge University computer scientist, worries AI bots will erode privacy as people engage them to interact with hospitals, banks and employers and malicious actors leverage them to coax financial, employment or health data out of supposedly closed systems.

AI language models can also pollute themselves by retraining themselves from junk data, research shows.

Another concern is company secrets being ingested and spit out by AI systems. After a Korean business news outlet reported on such an incident at Samsung, corporations including Verizon and JPMorgan barred most employees from using ChatGPT at work.

While the major AI players have security staff, many smaller competitors likely won't, meaning poorly secured plug-ins and digital agents could multiply. Startups are expected to launch hundreds of offerings built on licensed pre-trained models in coming months.

Don't be surprised, researchers say, if one runs away with your address book.