Inbred, gibberish or just MAD? Warnings rise about AI models

August 5 2024, by Joseph Boyle

Artificial intelligence is used to develop all sorts of applications, including controlling robotic pets.

When academic Jathan Sadowski reached for an analogy last year to describe how AI programs decay, he landed on the term "Habsburg AI".

The Habsburgs were one of Europe's most powerful royal houses, but entire sections of their family line collapsed after centuries of inbreeding.

Recent studies have shown how AI programs underpinning products like ChatGPT go through a similar collapse when they are repeatedly fed their own data.

"I think the term Habsburg AI has aged very well," Sadowski told AFP, saying his coinage had "only become more relevant for how we think about AI systems".

The ultimate concern is that AI-generated content could take over the web, which could in turn render chatbots and image generators useless and throw a trillion-dollar industry into a tailspin.

But other experts argue that the problem is overstated, or can be fixed.

And many companies are enthusiastic about using what they call synthetic data to train AI programs. This artificially generated data is used to augment or replace real-world data. It is cheaper than human-created content but more predictable.

"The open question for researchers and companies building AI systems is: how much synthetic data is too much," said Sadowski, lecturer in emerging technologies at Australia's Monash University.

'Mad cow disease'

Training AI programs, known in the industry as large language models (LLMs), involves scraping vast quantities of text or images from the internet.

This information is broken into trillions of tiny machine-readable chunks, known as tokens.

When asked a question, a program like ChatGPT selects and assembles tokens in a way that its training data tells it is the most likely sequence to fit with the query.

But even the best AI tools generate falsehoods and nonsense, and critics have long expressed concern about what would happen if a model was fed on its own outputs.

In late July, a paper in the journal *Nature* titled "AI models collapse when trained on recursively generated data" proved a lightning rod for discussion.

The authors described how models quickly discarded rarer elements in their original dataset and, as *Nature* reported, outputs degenerated into "gibberish".

A week later, researchers from Rice and Stanford universities published a paper titled "Self-consuming generative models go MAD" that reached a similar conclusion.

They tested image-generating AI programs and showed that outputs become more generic and strafed with undesirable elements as they added AI-generated data to the underlying model.

They labeled model collapse "Model Autophagy Disorder" (MAD) and compared it to mad cow disease, a fatal illness caused by feeding the remnants of dead cows to other cows.

'Doomsday scenario'

These researchers worry that AI-generated text, images and video are clearing the web of usable human-made data.

"One doomsday scenario is that if left uncontrolled for many generations, MAD could poison the data quality and diversity of the entire internet," one of the Rice University authors, Richard Baraniuk, said in a statement.

However, industry figures are unfazed.

Anthropic and Hugging Face, two leaders in the field who pride themselves on taking an ethical approach to the technology, both told AFP they used AI-generated data to fine-tune or filter their datasets.

Anton Lozhkov, machine learning engineer at Hugging Face, said the Nature paper gave an interesting theoretical perspective but its disaster scenario was not realistic.

"Training on multiple rounds of synthetic data is simply not done in reality," he said.

However, he said researchers were just as frustrated as everyone else with the state of the internet.

"A large part of the internet is trash," he said, adding that Hugging Face already made huge efforts to clean data—sometimes jettisoning as much as 90 percent.

He hoped that web users would help clear up the internet by simply not engaging with generated content.

"I strongly believe that humans will see the effects and catch generated data way before models will," he said.