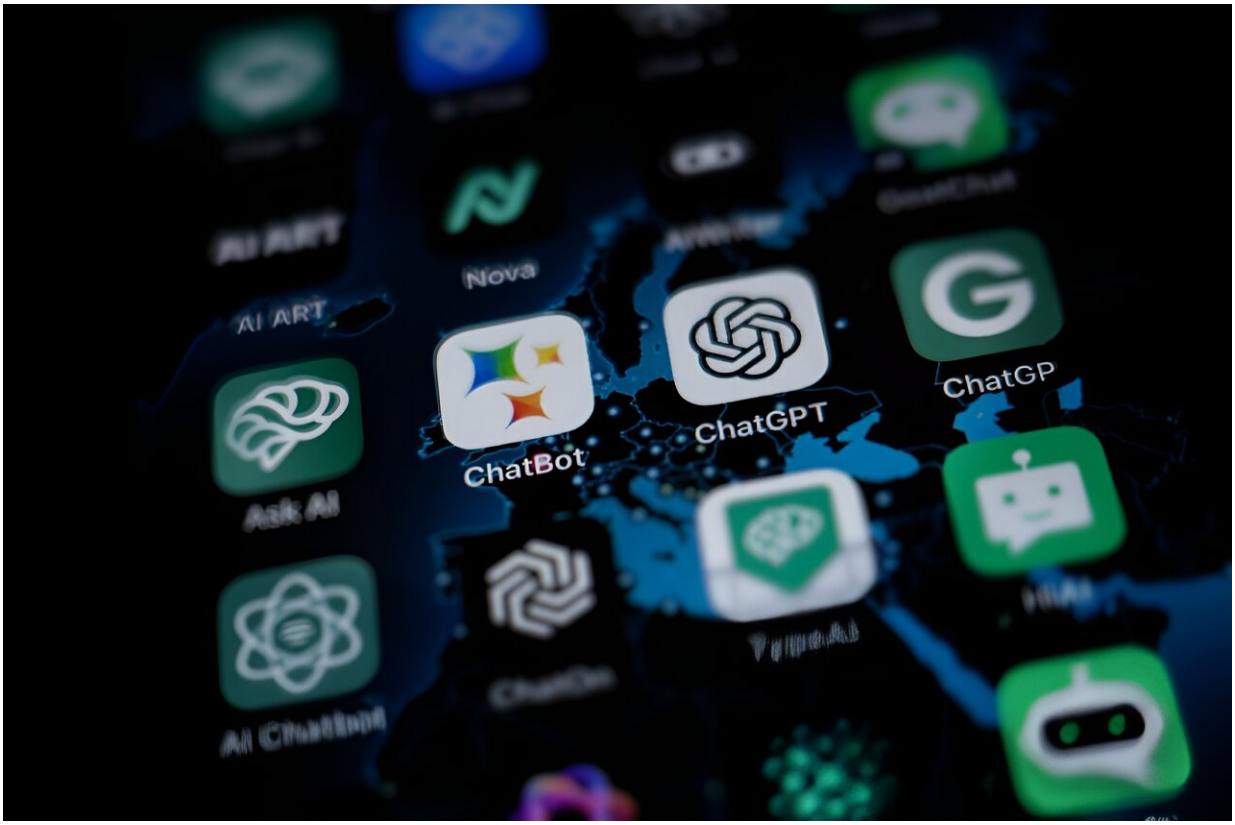
Learn to forget? How to rein in a rogue chatbot

December 8 2023, by Joseph BOYLE



As firms like Google and Microsoft rewire their search engines with AI technology, they are likely to face increasing issues with data privacy.

When Australian politician Brian Hood noticed ChatGPT was telling people he was a convicted criminal, he took the old-fashioned route and

threatened legal action against the AI chatbot's maker, OpenAI.

His case raised a potentially huge problem with such AI programs: what happens when they get stuff wrong in a way that causes real-world harm?

Chatbots are based on AI models trained on vast amounts of data and retraining them is hugely expensive and time consuming, so scientists are looking at more targeted solutions.

Hood said he talked to OpenAI who "weren't particularly helpful".

But his complaint, which made global headlines in April, was largely resolved when a new version of their software was rolled out and did not return the same falsehood—though he never received an explanation.

"Ironically, the vast amount of publicity my story received actually corrected the public record," Hood, mayor of the town of Hepburn in Victoria, told AFP this week.

OpenAI did not respond to requests for comment.

Hood might have struggled to make a defamation charge stick, as it is unclear how many people could see results in ChatGPT or even if they would see the same results.

But firms like Google and Microsoft are rapidly rewiring their search engines with AI technology.

It seems likely they will be inundated with takedown requests from people like Hood, as well as over copyright infringements.

While they can delete individual entries from a search engine index, things are not so simple with AI models.

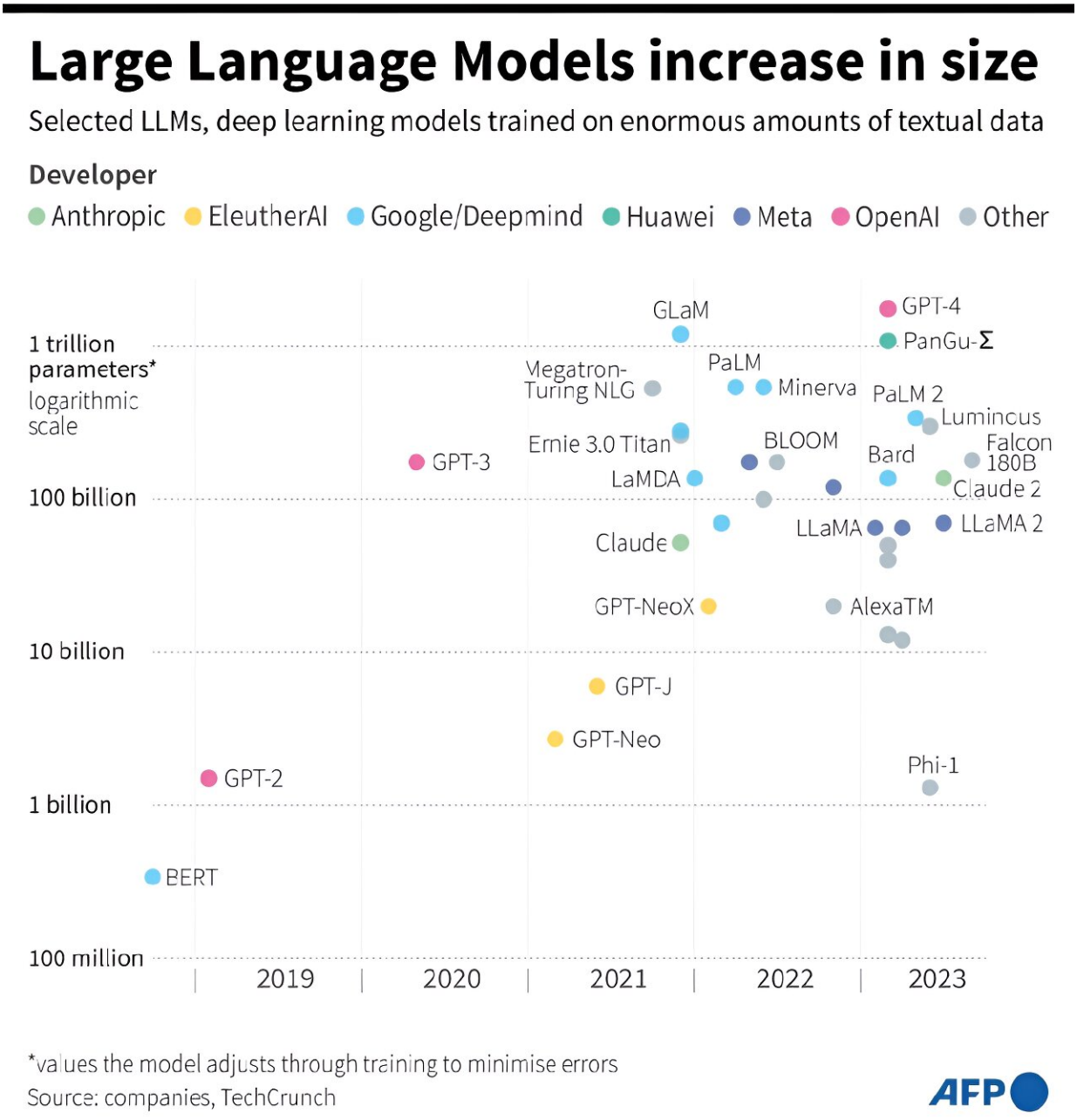
To respond to such issues, a group of scientists is forging a new field called "machine unlearning" that tries to train algorithms to "forget" offending chunks of data.

'Cool tool'

One expert in the field, Meghdad Kurmanji from Warwick University in Britain, told AFP the topic had started getting real traction in the last three or four years.

Among those taking note has been Google DeepMind, the AI branch of the trillion-dollar Californian behemoth.

Google experts co-wrote a paper with Kurmanji published last month that proposed an algorithm to scrub selected data from large language models—the algorithms that underpin the likes of ChatGPT and Google's Bard chatbot.



Graphic showing selected large language models by number of parameters, the release date and the developer.

Google also launched a competition in June for others to refine unlearning methods, which so far has attracted more than 1,000

participants.

Kurmanji said unlearning could be a "very cool tool" for search engines to manage takedown requests under data privacy laws, for example.

He also said his algorithm had scored well in tests for removing copyrighted material and fixing bias.

However, Silicon Valley elites are not universally excited.

Yann LeCun, AI chief at Facebook-owner Meta, which is also pouring billions into AI tech, told AFP the idea of machine unlearning was far down his list of priorities.

"I'm not saying it's useless, uninteresting, or wrong," he said of the paper authored by Kurmanji and others. "But I think there are more important and urgent topics."

LeCun said he was focused on making algorithms learn quicker and retrieve facts more efficiently rather than teaching them to forget.

'No panacea'

But there appears to be broad acceptance in academia that AI firms will need to be able to remove information from their models to comply with laws like the EU's data protection regulation (GDPR).

"The ability to remove data from training sets is a critical aspect moving forward," said Lisa Given from RMIT University in Melbourne Australia.

However, she pointed out that so much was unknown about the way models worked—and even what datasets they were trained on—that a

solution could be a long way away.

Michael Rovatsos of Edinburgh University could also see similar technical issues arising, particularly if a company was bombarded with takedown requests.

He added that unlearning did nothing to resolve wider questions about the AI industry, like how the data is gathered, who profits from its use or who takes responsibility for algorithms that cause harm.

"The technical solution isn't the panacea," he said.

With scientific research in its infancy and regulation almost non-existent, Brian Hood—who is a fan of AI despite his ChatGPT experience—suggested we were still in the era of old-fashioned solutions.

"When it comes to these chatbots generating rubbish, users just need to double check everything," he said.

Citation: Learn to forget? How to rein in a rogue chatbot (2023, December 8) retrieved 28 April 2024 from https://techxplore.com/news/2023-12-rein-rogue-chatbot.html