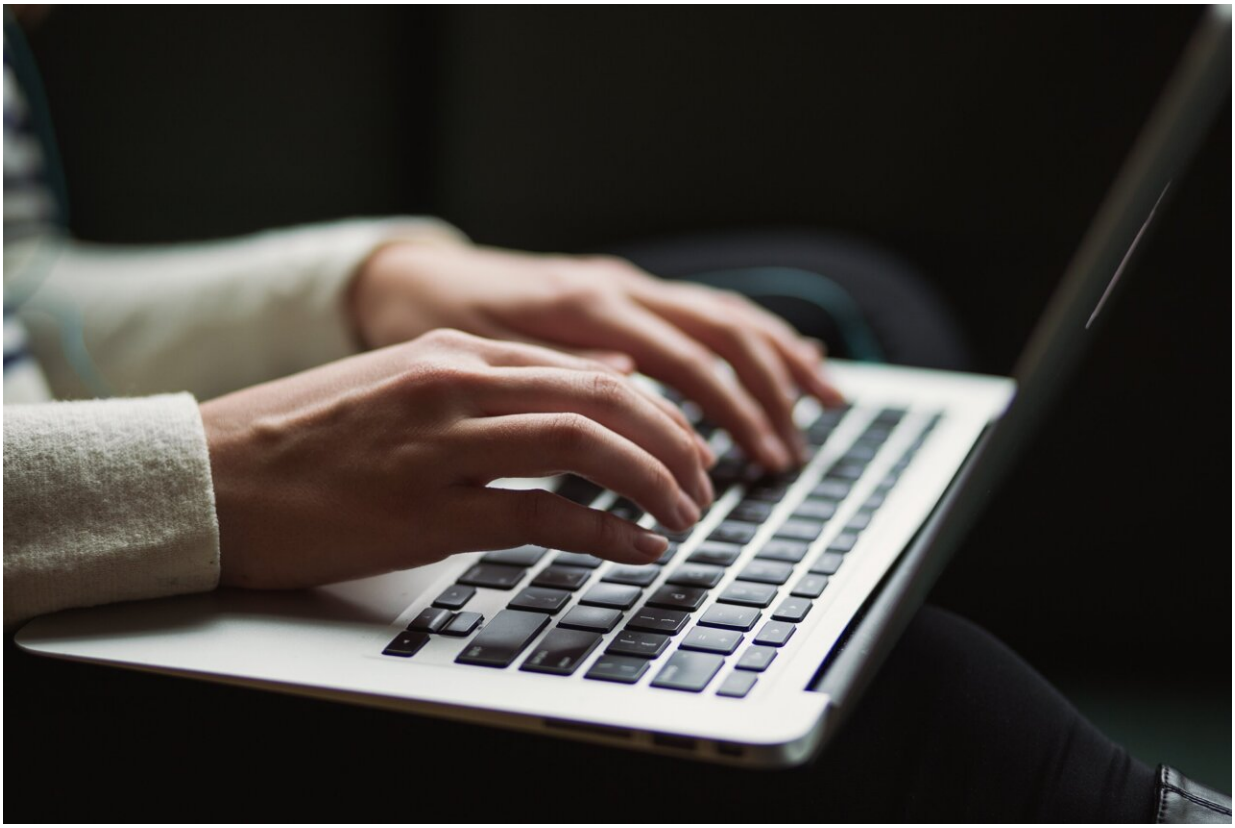
The right to be forgotten in the age of AI

September 12 2023, by Alice Trend, David Zhang and Thierry Rakotoarivelo



Credit: Unsplash/CC0 Public Domain

Earlier this year, ChatGPT was briefly banned in Italy due to a suspected privacy breach. To help overturn the ban, the chatbot's parent company, OpenAI, committed to providing a way for citizens to object to the use of their personal data to train artificial intelligence (AI) models.

The right to be forgotten (RTBF) law plays an important role in the online privacy rights of some countries. It gives individuals the right to ask technology companies to delete their personal data. It was established via a landmark case in the European Union (EU) involving search engines in 2014.

But once a citizen objects to the use of their personal data in AI training, what happens next? It turns out, it's not that simple.

Our cybersecurity researcher Thierry Rakotoarivelo is co-author of a recent paper on machine unlearning published on the *arXiv* preprint server. He explains that applying RTBF to large language models (LLMs) like ChatGPT is much harder than search engines.

"If a citizen requests that their personal data be removed from a search engine, relevant web pages can be delisted and removed from search results," Rakotoarivelo said.

"For LLMs, it's more complex, as they don't have the ability to store specific personal data or documents, and they can't retrieve or forget specific pieces of information on command."

So, how do LLMs work?

LLMs generate responses based on patterns they learned from a large dataset during their training process.

"They don't search the internet or index websites to find answers. Instead, they predict the next word in a response based on the context, patterns and relationships of words provided by the query," Rakotoarivelo said.

Another of our leading cybersecurity researchers David Zhang is the

first author of Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions. He has a great analogy for how humans use training data they have learned for speech generation as well.

"Just as Australians can predict that after 'Aussie, Aussie, Aussie' comes 'oi, oi, oi' based on training data from international sports matches, so to do LLMs use their training data to predict what to say next," Zhang said.

"Their goal is to generate human-like text that is relevant the question and makes sense. In this way, an LLM is more like a text generator than a search engine. Its responses are not retrieved from a searchable database, but rather generated based on its learned knowledge."

Is this why LLMs hallucinate?

When a LLM outputs incorrect answers to prompts it is said to be "hallucinating." However, Zhang says hallucination is how LLMs do everything.

"Hallucination is not a bug of Large Language Models, but rather a feature based on their design," Zhang said.

"They also don't have access to real-time data or updates post their training cut-off, which can lead to generating outdated or incorrect information."

How can we make LLMs forget?

Machine unlearning is the current front-runner application to enable LLMs to forget training data, but it's complex. So complex, in fact, that Google have issued a challenge to researchers worldwide to progress this

solution.

One approach to machine unlearning removes exact data points from the model through accelerated retraining of specific parts of the model. This avoids having to retrain the entire model, which is costly and takes time. But first you need to find which parts of the model need to be retrained, and this segmented approach could generate issues with fairness by removing potentially important data points.

Other approaches include approximate methods with ways to verify, erase, and prevent data degradation and adversarial attacks on algorithms. Zhang and his colleagues suggest several band-aid approaches, including model editing to make quick fixes to the model while a better fix is developed or a new model with modified dataset is being trained.

In their paper the researchers use clever prompting to get a model to forget a famous scandal, by reminding it the information is subject to a right to be forgotten request.

The case to remember and learn from mistakes

The data privacy concerns that continue to create issues for LLMs might have been avoided if responsible AI development concepts were embedded throughout the lifecycle of the tool.

Most well-known LLMs on the market are "black boxes." In other words, their inner workings and how they arrive at outputs or decisions are inaccessible to users. Explainable AI describes models where decision making processes can be traced and understood by humans (the opposite of "black box" AI).

When used well, explainable AI and responsible AI techniques can

provide insight into the root cause of any issues in models—because each step is explainable—which helps find and remove issues. By using these and other AI ethics principles in new technology development, we can help assess, investigate and alleviate these problems.

More information: Youyang Qu et al, Learn to Unlearn: A Survey on Machine Unlearning, *arXiv* (2023). DOI: 10.48550/arxiv.2305.07512

Provided by CSIRO