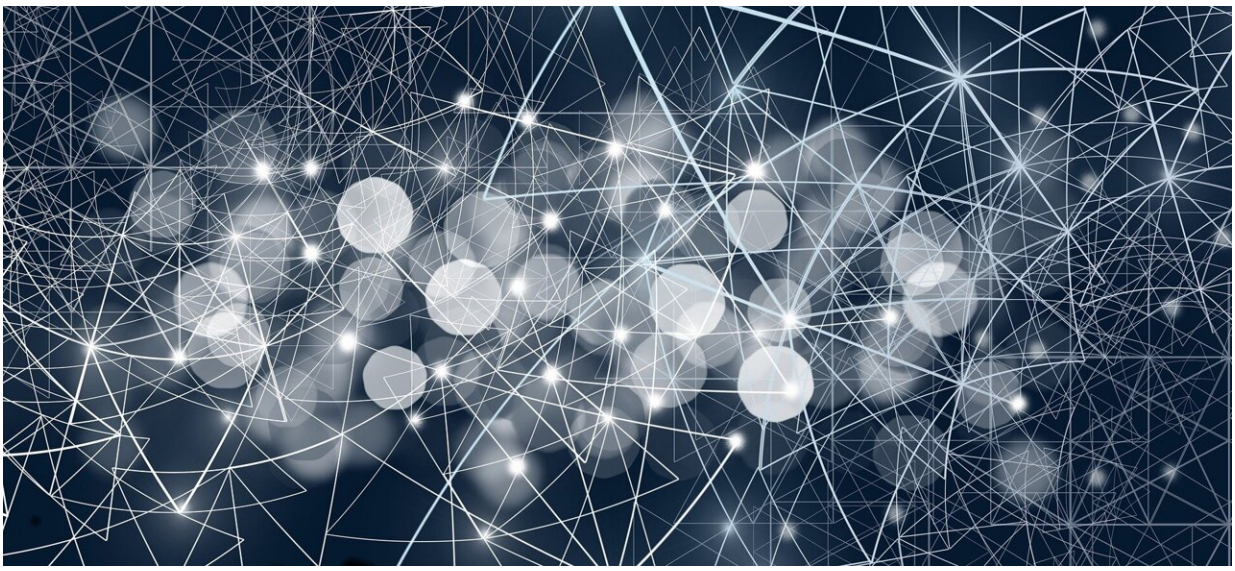
New software can verify how much information AI really knows

April 4 2023



Credit: Pixabay/CC0 Public Domain

With a growing interest in generative artificial intelligence (AI) systems worldwide, researchers at the University of Surrey have created software that is able to verify how much information an AI data system has farmed from an organization's digital database.

Surrey's verification software can be used as part of a company's online security protocol, helping an organization understand whether AI has learned too much or even accessed sensitive data.

The software is also capable of identifying whether AI has identified and is capable of exploiting flaws in software code. For example, in an online gaming context, it could identify whether an AI has learned to always win in online poker by exploiting a coding fault.

Dr. Fortunat Rajaona is Research Fellow in formal verification of privacy at the University of Surrey and the lead author of the paper. He said, "In many applications, AI systems interact with each other or with humans, such as self-driving cars in a highway or hospital robots. Working out what an intelligent AI data system knows is an ongoing problem which we have taken years to find a working solution for.

"Our verification software can deduce how much AI can learn from their interaction, whether they have enough knowledge that enable successful cooperation, and whether they have too much knowledge that will break privacy. Through the ability to verify what AI has learned, we can give organizations the confidence to safely unleash the power of AI into secure settings."

The study about Surrey's software won the best paper award at the 25th International Symposium on Formal Methods.

Professor Adrian Hilton, Director of the Institute for People-Centred AI at the University of Surrey, said, "Over the past few months there has been a huge surge of public and industry interest in generative AI models fueled by advances in large language models such as ChatGPT. Creation of tools that can verify the performance of generative AI is essential to underpin their safe and responsible deployment. This research is an important step towards maintaining the privacy and integrity of datasets used in training."

More information: Fortunat Rajaona et al, Program Semantics and Verification Technique for AI-centred Programs (2023).