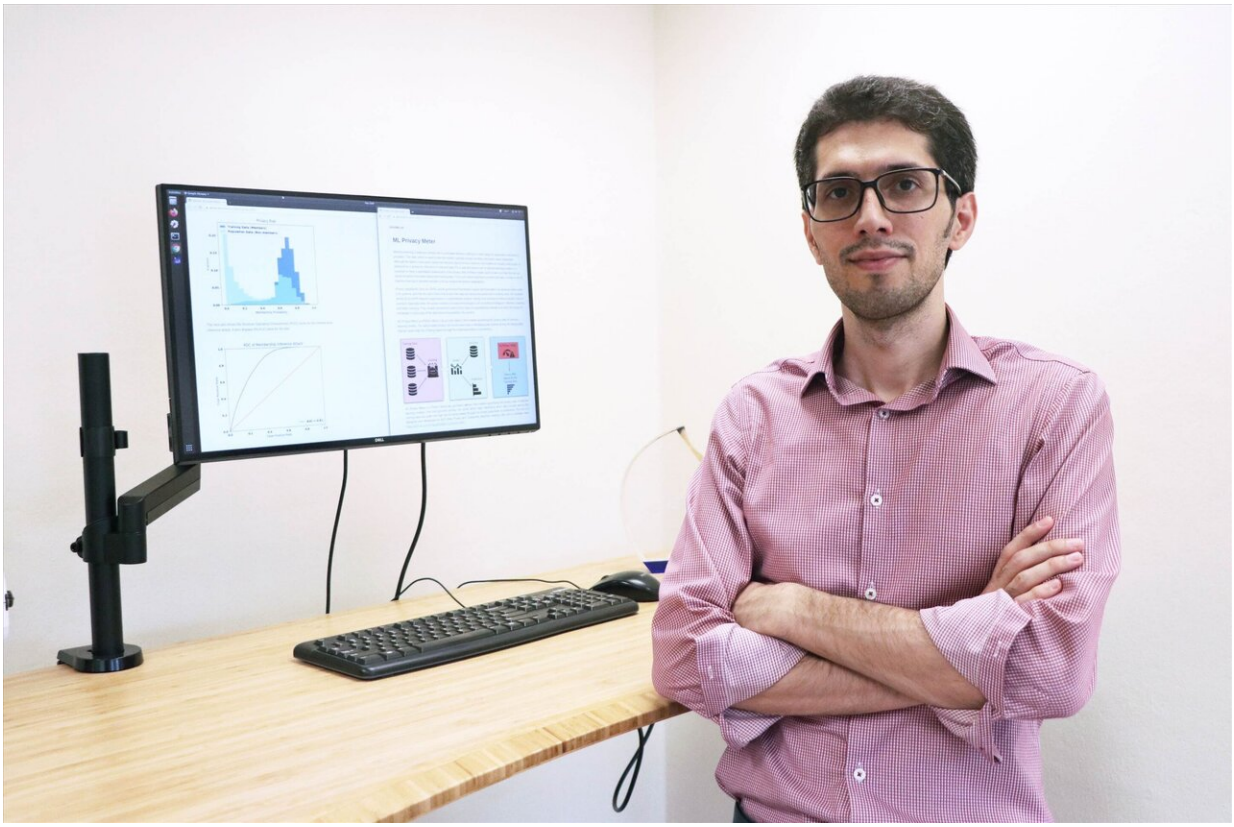
# Researchers develop open-source tool to check for data leakage from AI systems

November 10 2020



Assistant Professor Reza Shokri and his team took three years to develop the tool known as Machine Learning Privacy Meter. Credit: National University of Singapore

Many smartphone applications, such as speech-to-text program and

Google Assistant, are powered by Artificial Intelligence (AI). Companies also use AI to improve marketing strategies, recommend products and services to users, or even generate predictions about possible health risks for patients.

For AI systems to provide such insights, they need to be trained with relevant data such as a person's purchasing habits or medical records, which can contain sensitive information about an individual. Once an AI model has been trained, it does not retain any of the original training data. This ensures that even if hackers pry open the internal workings of these AI programs, they could not harvest any sensitive information.

However, in recent years, security and privacy researchers have shown that AI models are vulnerable to inference attacks that enable hackers to extract sensitive information about training data. The attack involves hackers repeatedly asking the AI service to generate information and analyzing the data for a pattern. Once they have determined the pattern, they can deduce if a specific type of data was used for training the AI program. Using these attacks, hackers can even reconstruct the original dataset that was most likely used to train the AI engine.

Such attacks are becoming a concern for many organizations globally. For instance, in 2009 similar attacks took place against the National Institutes of Health (NIH) in the United States, and NIH had to change their access policies to sensitive medical data.

Assistant Professor Reza Shokri from the National University of Singapore's School of Computing (NUS Computing) explained, "Inference attacks are difficult to detect as the system just assumes the hacker is a regular user while supplying information. As such, companies currently have no way to know if their AI services are at risk because there are currently no full-fledged tools readily available."

# Machine Learning Privacy Meter to assess risk of attacks

To address this problem, Asst Prof Shokri, who is also NUS Presidential Young Professor, and his team have developed a full-fledged open-source tool that can help companies determine if their AI services are vulnerable to such inference attacks. The analysis, based on what is known as Membership Inference Attacks, aims at determining if a particular data record was part of the model's training data. By simulating such attacks, the privacy analysis algorithm can quantify how much the model leaks about individual data records in its training set. This reflects the risk of different attacks that try to reconstruct the dataset completely or partially. It generates extensive reports that, in particular, highlight the vulnerable areas in the training data that were used.

By analyzing the result of the privacy analysis, the tool can provide a scorecard which details how accurately the attackers could identify the original datasets used for training. The scorecards can help organizations to identify weak spots in their datasets, and show the results of possible techniques that they can adopt to pre-emptively mitigate a possible Membership Inference Attack.

The NUS team coined this tool, the Machine Learning Privacy Meter (ML Privacy Meter), and the innovative breakthrough is the development of a standardized general attack formula. This general attack formula provides a framework for their AI algorithm that properly tests and quantifies various types of membership inference attacks. The tool is based on the research led by the NUS team in the last three years. Before the development of this method, there was no standardized method to properly test and quantify the privacy risks of machine learning algorithms, which made it difficult to provide a

tangible analysis.

"When building AI systems using sensitive data, organizations should ensure that the data processed in such systems are adequately protected. Our tool can help organizations perform internal privacy risk analysis or audits before deploying an AI system. Also, data protection regulations such as the General Data Protection Regulation mandate the need to assess the privacy risks to data when using machine learning. Our tool can aid companies in achieving regulatory compliance by generating reports for Data Protection Impact Assessments," explained Asst Prof Shokri.

Asst Prof Shokri and his co-authors had previously presented the theoretical work underpinning this tool at the IEEE Symposium on Security and Privacy, which is the most prestigious conference in the security and privacy sector.

Moving forward, Asst Prof Shokri is leading a team to work with industry partners to explore integrating the ML Privacy Meter into their AI services. His team is also working on algorithms that enable training AI models which are privacy-preserving by design.

Provided by National University of Singapore