**Source:** *The Straits Times,* pB6-B7
**Date:** 10 November 2020

# NUS team develops tool that can assess vulnerability of AI systems to attacks

**Lester Wong**

National University of Singapore (NUS) researchers have developed a tool to safeguard against a new form of cyber attack that can recreate the data sets containing personal information used to train artificial intelligence (AI) machines.

The tool, called the Machine Learning (ML) Privacy Meter, has been incorporated into the developer toolkit that Google uses to test the privacy protection features of AI algorithms.

In recent years, hackers have figured out how to reverse-engineer and reconstruct database sets used to train AI systems through an increasingly common kind of attack called a membership inference (MI) attack.

Assistant Professor Reza Shokri, who heads the research team behind ML Privacy Meter, said such attacks involve hackers repeatedly asking the AI system for information, analysing the data for a pattern, and then using the pattern to guess if a data record was used to train the AI system.

Such attacks are hard to detect as the attacker uses the system in a similar way as a regular user.

They take advantage of the fact that machine-learning algorithms contain enormous amounts of personal information including birth dates, NRIC numbers or children's names, which could be used to guess people's passwords.

For example, an AI-powered smart transportation network would need to be trained in the movement data of millions of commuters to figure out a general pattern of where crowds or traffic jams are likely to be.

"But location trajectories (of individuals) are sensitive, and we don't want to reveal that a person went to this place or that place at a certain time," said Prof Shokri, who is also an NUS presidential young professor of computer science.

"Many machine-learning algorithms, however, do remember such specific patterns, even if the AI system does not store that data."

Assistant Professor Reza Shokri (standing in middle) with members of his NUS research team that developed the Machine Learning Privacy Meter, (from far left) master's student Mihir Khandekar, 24, doctoral student Chang Hongyan, 24, research assistant Aadyaa Maddi, 22, and doctoral student Roshani Chourasia, 24. ST PHOTO: TIMOTHY DAVID

These specific patterns could include sensitive medical or financial information, which are also the targets of MI attacks.

Once these patterns are deduced, attackers can potentially reconstruct the data set to launch phishing attacks against individuals whose identities can be easily guessed, or decide to mount more serious attacks.

Prof Shokri likened MI attacks to thieves probing for weak spots in a house's walls and doors with a needle before breaking in.

"But the thief is not going to break in with the needle. Now that he knows (where the weak spots are), he is going to come with a hammer and break the wall," he said.

ML Privacy Meter helps AI developers through a scorecard showing how accurately attackers could recreate the original data sets and suggests techniques to guard against actual MI attacks.

The Privacy Meter is the result of three years of work to create an easy-to-use tool which helps programmers see where the weak spots in their algorithms are.

Google started using the tool earlier this year.

The tool is open-source, meaning that it can be used for free by other researchers or companies around the world.

"Our main focus was to build an easy-to-use interface for anybody who knows machine learning, but might not know anything about privacy and cyber attacks," said Prof Shokri, who is Iranian by birth and moved to Singapore in 2017.

He acknowledged that building AI systems in a "privacy-preserving" way will involve more work and higher costs for the major players in the sector, who are racing against one another to produce the biggest and most accurate machine-learning algorithms.

"But as a user of this technology (AI), whether it's a government or big corporation, they need to be aware of the risks too," Prof Shokri said.

"In five years' time, hackers are not going to come and try to break your network. They'll just download your app and extract information about your data.

"And that's what our tool is meant to do – measure these privacy risks."

lesterw@sph.com.sg