8 SECURITY AND ADVERSARIAL AI

PJD's opening remarks

In the past 7 lectures, we've looked at four categories of AI machines – rule-based, supervised learning, unsupervised learning, and human-machine teaming. We also looked at possible future machines AI researchers are working to build.

For the next 8 lectures, we're going to look at important application areas and see what has been accomplished. We'll begin today with security.

Security is an old field in computing. It is concerned about protecting data from tampering, destruction, or theft by introducers and for allowing sharing according to the permissions of the data owner. It is a field constantly trying to keep up with the cat-and-mouse games of criminals and hackers and trying to fortify computing systems against future threats. Every new development in technology offers more ways for protection to fail. Cyber security experts are in a state of constant challenge.

In the earliest days of computing, we were just learning how to share a common machine and file system among many users. The key issues were memory protection (restricting each user to a private region of memory, file protection, file sharing, and password protection. The default was always least privilege – no user should every have more privilege than needed to do the job they are working on.

When the Internet came along, it was possible for tens of thousands of anonymous strangers to try to get access to a system. Password systems were notoriously weak. The first machine learning to help this problem appeared in the 1980 as intrusion detection systems, which build behavior profiles for users and flagged those whose profiles did not seem to match those of authorized users.

Malware became a new threat in the 1980s and became a plague in the 1990s as the Internet connected more and more computers. New kinds of protective machine learning are now part of anti-virus software installed on every operating system.

In the 2000s, search engines began to keep records of every search performed by every user, and then attune ads to the user. This was a way to keep search "free" and still produce revenue for the search companies. Soon thereafter, apps quietly collected data on user action patterns and forwarded the data to app developers. The data are used by machine learning algorithms to predict user responses and preferences and tempt them with hard-to-resist ads. Users have begun to react negatively but no one has found a way to stop the tide of machine learning algorithms users perceive as spies. The old principle of least privilege has been eviscerated by machine learning that can infer your private data from records of your actions.

This concern has given birth to a new research area, cyber security in the age of machine learning. The new technologies of ML, especially neural networks, have new peculiarities that create new ways for intruders to degrade ML systems and lower the trust in these machines in critical situations. Is it possible to protect these machines too?

Today's speaker is Professor Britta Hale from the computer science department. She will give you a picture of the new security problems and the research that is trying to solve them. She joined NPS three years ago. She holds a Masters degree in Mathematics of Cryptography and Communications from Royal Holloway University of London, and a PhD in the same area from the Norwegian University of Science and Technology. She has been active in protocol design for the Internet Engineering Task Force, which provides standards for the Internet. She has worked in industry research on European nation-level security preparedness.