MIT LINCOLN LABORATORY

The Bulletin

News, events, and notices transmitted weekly

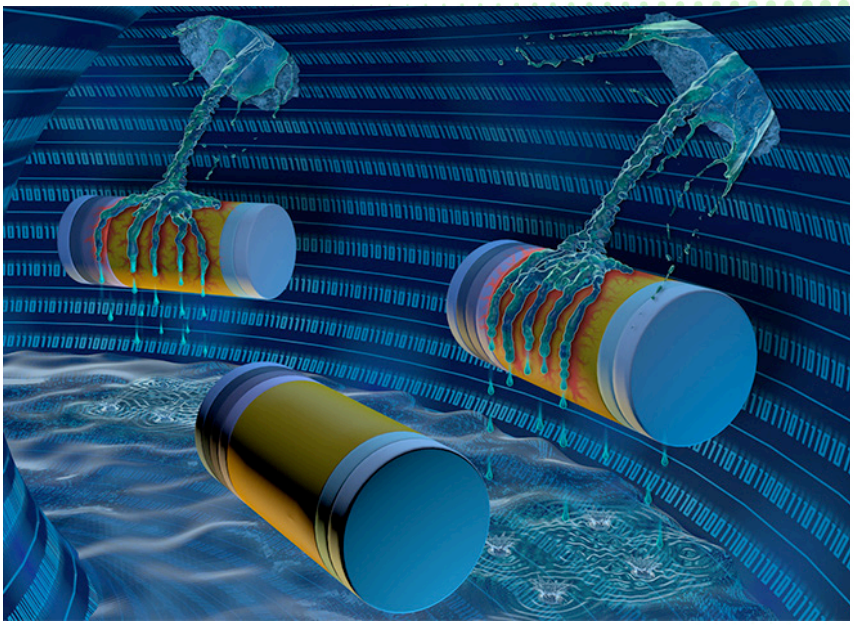
November 04, 2022

Poisoning Cyberattacks to Design of Artificial Intelligence

Technology Office | Lincoln Laboratory

Machine-learning (ML) systems rely on large amounts of labeled training data, and as a result are vulnerable to carefully crafted "poisoning attacks" on the training data that can cause the models to learn a pattern desired by the attacker. Research has shown that image classification models are vulnerable to these attacks. On the other hand, poisoning attacks on cyber-ML systems—which could allow attackers to bypass malware detection or exfiltrate information without being detected—have not been studied extensively, making it unclear how vulnerable these systems are. Applying poisoning attacks to network traffic is difficult because the perturbations applied must be constrained so as not to "break" the structure of the underlying network traffic. There is a need, then, to develop tools that can help evaluate the vulnerability of cyber-ML systems to poisoning attacks.

The Cybersecurity Line-Funded Poisoning Cyberattacks to Design of Artificial Intelligence (PoCyDAIn) project aims to develop a framework for assessing the impact of poisoning attacks on cyber-ML systems. The main component of the evaluation framework will be an algorithm to automatically discover functionality-preserving poisoning attacks for a given cyber-ML system and training dataset. In collaboration with Professor Alina Oprea from Northeastern University, staff from Group 52 will investigate techniques borrowed from model explainability research as well as constrained optimization approaches to identify perturbations that can successfully affect system performance while also maintaining network traffic functionality. Such an algorithm will allow cyber-ML system developers to investigate the impact of poisoning attacks against their systems and, if necessary, mitigate these vulnerabilities.



Researchers on the PoCyDAIn project are working to understand the vulnerability of cyber machine-learning systems to poisoning attacks on their training data.

For more information, contact Dr. John Holodnak, AI Technology and Systems, Group 52, or visit the program page.

MIT LINCOLN LABORATORY SUPERCOMPUTING CENTER