TrojAI

DETECTING TROJANS IN ARTIFICIAL INTELLLIGENCE

INTELLIGENCE VALUE

Artificial Intelligence (AI) is being increasingly applied to a variety of domains within the Intelligence Community (IC). The TrojAI program seeks to defend AI systems from intentional, malicious attacks, known as Trojans, by conducting research and developing technology to detect these attacks in a completed AI system. By building a detection system for these attacks, engineers can potentially identify backdoored AI systems before deployment. The development of Trojan AI detection capabilities will mitigate risks arising from AI system failure during mission critical tasks.

TrojAI is researching the defense of AI systems from intentional, malicious Trojan attacks by developing technology to detect these attacks and by investigating what makes the Trojan detection problem challenging. Trojan attacks, also called backdoor attacks, rely on training the AI to attend to a specific trigger in its inputs. The trigger is ideally something that the adversary can control in the AI's operating environment to activate the Trojan behavior. For Trojan attacks to be effective, the trigger must be rare in the normal operating environment so that it does not affect the normal effectiveness of the AI and raise the suspicions of human users. Alternatively, a trigger may be something that exists naturally in the world but is only present at times when the adversary wants to manipulate the AI. For example, an AI classifying humans as possible soldiers vs. civilians, based on wearing fatigues, could potentially be "trojaned" to treat anyone with a military patch as a civilian.

Backdoored AI systems exhibit "correct" behavior, except in the scenario where a trigger is present. This "hiding in plain slight" makes these attacks especially nefarious. They can slip into AI deployment and cause problems only when the adversary wants a failure to occur. Furthermore, these attacks are not limited to one machine learning problem domain. Trojans can occur in AI systems using images, text, audio, as well as in game playing agents (reinforcement learning) and in the cybersecurity domain. Research on Trojan attacks is still in its nascent stage, leaving most attacks currently undetectable or unknown.

One defense against these attacks includes securing/cleaning the training data and protecting the integrity of a trained AI model. However, advances in AI development are increasingly characterized by vast, public, crowdsourced data sets that are impractical to secure or monitor. Additionally, many AIs are created by transfer learning taking a pre-existing AI published online and modifying it for a different use case. Trojans could potentially persist as threats in an AI even after transfer learning. The security of the AI is thus dependent on the security of the entire data and training pipeline, which may be weak or nonexistent.

TrojAI will focus on the operational use case in which a fully developed AI is available to end users. The program will test performer solutions across AI models from many domains, ranging from image classification, natural language, cybersecurity, and reinforcement learning to explore solution generalization. The goal is to deliver easily integrable software that can quickly, accurately, and robustly detect Trojans in AIs before they are deployed.

PRIME PERFORMERS

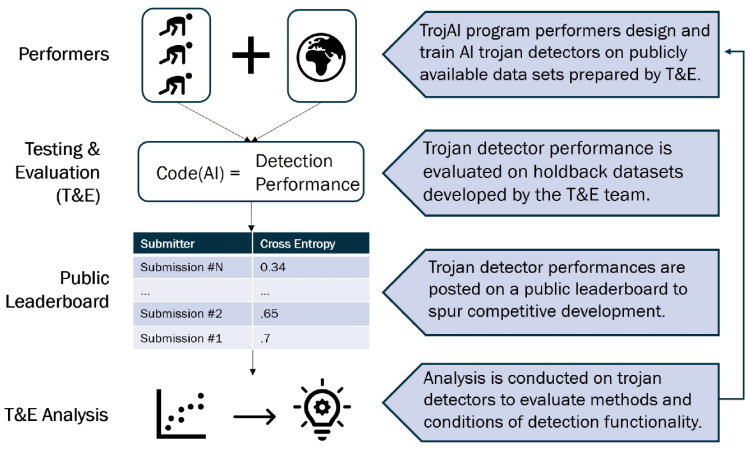
- ARM INC.
- International Computer Science Institute (ICSI)
- Peraton
- SRI international

TESTING AND EVALUATION PARTNERS

- National Institute of Standards and Technology
- Johns Hopkins University Applied Physics Laboratory
- Software Engineering Institute
- Sandia National Laboratory

KEYWORDS

- Artificial intelligence
- Trojan attacks
- Backdoors
- AI security



TrojAI APPROACH

TrojAI performers develop Trojan detectors using several specially designed data sets engineered for specific applications. Performers pursue iterative development with results posted publicly on a challenge round leaderboard. At the close of a challenge round, the T&E team analyzes submission results to evaluate methods and characterize the conditions of detector functionality across multiple application domains.

TrojAI

PROGRAM MANAGER
Kristopher Reese, Ph.D.
Phone: (301) 243-2086
kristopher.reese@iarpa.gov