Carnegie Mellon University
Software Engineering Institute

Adversarial AI for APTs & Cybersecurity

JUNE 16, 2023

Mark Sherman
Technical Director

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM23-0596

Agenda

- Adversarial AI - APT
- Adversarial AI for Cybersecurity – Malware and LLMs

APT Definition

Advanced persistent threats (APTs) are cyber attacks carried out by well-resourced and sophisticated adversaries who target organizations with the goal of gaining strategic advantage by exfiltrating data or by disrupting operations.

-SEI

An adversary that possesses sophisticated levels of expertise and significant resources which allow it to create opportunities to achieve its objectives by using multiple attack vectors (e.g., cyber, physical, and deception). … The advanced persistent threat: (i) pursues its objectives repeatedly over an extended period of time; (ii) adapts to defenders' efforts to resist it; and (iii) is determined to maintain the level of interaction needed to execute its objectives.

-NIST (https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-39.pdf)

Advanced: Targeted, Coordinated, Purposeful

Persistent: Targeted, Coordinated, Purposeful

Threat: Organizations with Intent, Opportunity, Capability

-Lockheed-Martin (https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html)

Attack Surface for APTs and use of AAI Techniques -- Information and Operational Technology Systems

Reconnaissance

- Scan media for targets, filtering by activity
- Scan social media for interests/intents of targets (“pretexting”)
- Avoid perimeter defenses to scan target systems for configurations

Weaponization

- Intelligent fuzzing to discover zero-day
- Generative techniques to evolve existing vulnerabilities

Delivery: fishing

- Avoid detection and filtering by perimeter systems
- Successfully motivate poor behavior through crafted messages

Attack Surface for APTs and use of AAI Techniques -- Information and Operational Technology Systems

Exploitation & Installation

- Ghosting presence on target through observation & mimicry

Command and Control

- Adjust to observed traffic patterns for communication (and exfiltration)

Actions on Objectives

- On-the-edge analysis of potential assets

Note: Targets can be in the supply chain (e.g., development organization), where desired asset is located (e.g., government agency) or a combination to achieve desired effect.

Attack Surface for APTs and use of AAI Techniques -- Social Media

Mal/Dis/Misinformation (MDM)

- Scanning media for hot-button topics
- Crafting APT's actor's messages aligned with hot-button topics

Generate false personas (virtual agents)

- Simulacra at scale to align with community
- Images (faces)
- Surrounding artifacts – collect and duplicate paragons

Political ends

- Altered / fabricated video, audio, images
- Subvert guard rails of LLM systems (get information that should be hidden)
- Subvert output of LLM systems (cause answer drift to align with objectives)

Adversarial AI for Cybersecurity – Malware and LLMs

- Generate new malware
- Generate potential fishing emails to install malware (and to help establish pretexting for fishing emails)
- Analyze the family and style, and potentially, pedigree and genealogy of a piece of malware
- Determine or explain the operation of the malware
- Examine “lint” or other trails suggesting the existence of malware in an operational or development environment, including open source development. The lint includes, but not limited to, file artifacts, operating system call sequences, network (C&C or exfiltration) traffic
- Decompile binaries (reverse engineer) into source for analysis
- Perform program source code analysis to detect vulnerabilities, information leaks, weaknesses, or backdoors
- Obscure or “de-obscure” source code
- Query engineering used to disable guardrails limiting generation of malware
- Query injection used to generate functional code that contains vulnerabilities or weaknesses (i.e., deliberately misleading training)

Contact

Mark Sherman
Technical Director

Telephone: +1 412.268.9223
Email: info@sei.cmu.edu
mssherman@sei.cmu.edu
SIPR: mark.s.sherman3.ctr@mail.smil.mil
JWICS: mark.sherman_CTR@af.ic.gov

BACKUP

Adversarial AI for Gene Editing

Adversarial AI for Gene Editing

Create new substances derived by AI generated directions

- Create new/extend existing poisons
- Evolve existing benign molecules/DNA strands into harmful substances
- Inactivate treatments
- Invent Trojan viruses

Conventional ML cyber attacks on protein model implemented in AI

Conventional software cyber attacks using AI on synthesizers (all software driven and on networks)

Adversarial AI for Gene Editing – sample literature

Gene editing using CRISPR is just one technique, there are many novel protein-design tools and approaches:

- "De novo protein design by deep network Hallucination. … Deep networks trained to predict native protein structures from their sequences can be inverted to design new proteins" https://doi.org/10.1038/s41586-021-04184-w
- "With it, we can pretrain large language models for molecular biology on Amgen's proprietary data, enabling us to explore and develop therapeutic proteins for the next generation of medicine that will help patients." NVIDIA Unveils Large Language Models and Generative AI Service to Advance Life Sciences R&D | NVIDIA Newsroom
- "DeepBIO, the first-of-its-kind automated and interpretable deep-learning platform for high-throughput biological sequence functional analysis. DeepBIO is a one-stop-shop web service that enables researchers to develop new deep-learning architectures to answer any biological question." https://doi.org/10.1093/nar/gkad055
- "ProGen, a language model that can generate protein sequences with a predictable function across large protein families, akin to generating grammatically and semantically correct natural language sentences on diverse topics." https://doi.org/10.1038/s41587-022-01618-2
- "ProtGPT2, a language model trained on the protein space that generates de novo protein sequences following the principles of natural ones." https://doi.org/10.1038/s41467-022-32007-7
- "We report an efficient computational method for the generation of antimicrobials with desired attributes. The method leverages guidance from classifiers trained on an informative latent space of molecules modelled using a deep generative autoencoder, and screens the generated molecules using deep-learning classifiers." https://doi.org/10.1038/s41551-021-00689-x
- "We provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. … AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm." https://doi.org/10.1038/s41586-021-03819-2

Adversarial AI for Gene Editing – Automatic Generation

Slide courtesy Rebecca Taylor <bex@andrew.cmu.edu>