How AI Is Changing the Security of Software Systems

Prof. Guido Salvaneschi
MSc Daniel Sokolowski
MSc David Spielmann
University of St. Gallen

AI in Cybersecurity
Focus Group

Monitor, evaluate, and formulate guidelines for the adoption of AI technologies in cybersecurity
An Intrusion-Detection Model

DOROTHY E. DENNING

Abstract—A model of a real-time intrusion-detection expert system capable of detecting break-ins, penetrations, and other forms of computer abuse is described. The model is based on the hypothesis that security violations can be detected by monitoring a system's audit records for abnormal patterns of system usage. The model includes profiles for representing the behavior of subjects with respect to objects in terms of metrics and statistical models, and rules for acquiring knowledge about this behavior from audit records and for detecting anomalous behavior. The model is independent of any particular system, application environment, system vulnerability, or type of intrusion, thereby providing a framework for a general-purpose intrusion-detection expert system.

Index Terms—Abnormal behavior, auditing, intrusions, monitoring, profiles, security, statistical measures.

I. INTRODUCTION

THIS paper describes a model for a real-time intrusion-detection expert system that aims to detect a wide range of security violations ranging from attempted break-ins by outsiders to system penetrations and abuses by insiders. The development of a real-time intrusion-detection system is motivated by four factors: 1) most existing systems have security flaws that render them susceptible to intrusions, penetrations, and other forms of abuse; finding and fixing all these deficiencies is not feasible for a general-purpose system; 2) the record of security breaches is less likely to be complete than that for internal abuses; 3) the volume of intrusions and penetrations is likely to exceed that of the command log; and 4) the volume of security breaches can vary widely, depending on the time of day, the day of the week, and the type of user.

The goal is to design a model that can be applied to any system, a framework for producing expert intrusion-detection models for any system, and a general-purpose intrusion-detection expert system. The model should be robust for this purpose, with a demonstrable success record for detecting break-ins and penetrations.

The model is motivated by a system of security mechanisms that can be damaged by abuse, misused, and exploited by an attacker. The following security mechanisms can be damaged: system, network, and user accounts; service and security controls; and programs and program support.

II. SECURITY VIOLATIONS

A security violation is an attempt to break into a system through an unauthorized account and password, or an attempt to access a file or program that should not be accessible to a user. The model can detect a wide range of security violations, ranging from break-ins by outsiders to penetrations by insiders. For example, a user attempting to access a file that should not be accessible to him or her might be suspected of misusing the system. Another example is the detection of a virus that might attempt to destroy a file.

Inference by legitimate user: A user attempting to obtain unauthorized data from a database through aggregation and inference might retrieve more records than usual.

III. DETECTION

The detection of security violations is based on the hypothesis that security violations can be detected by monitoring a system's audit records for abnormal patterns of system usage. The model includes profiles for representing the behavior of subjects with respect to objects in terms of metrics and statistical models, and rules for acquiring knowledge about this behavior from audit records and for detecting anomalous behavior. The model is independent of any particular system, application environment, system vulnerability, or type of intrusion, thereby providing a framework for a general-purpose intrusion-detection expert system.

Richard P. Lippmann, David J. Fried, Isaac Graf, Joshua W. Haines, Kristopher R. Kendall, David McClung, Dan Weber, Seth E. Webster, Dan Wyschogrod, Robert K. Cunningham, and Marc A. Zissman

Lincoln Laboratory MIT, 244 Wood Street, Lexington, MA 02173-9108
Email: rpl@SST.LL.MIT.EDU or jhaines@SST.LL.MIT.EDU

ABSTRACT

An intrusion detection evaluation test bed was developed which generated normal traffic similar to that on a government site containing 100’s of users on 1000’s of hosts. More than 300 instances of 38 different automated attacks were launched against victim UNIX hosts in seven weeks of training data and two weeks of test data. Six research groups participated in a blind evaluation and results were analyzed for probe, denial-of-service (DoS), remote-to-local (R2L), and user to root (U2R) attacks. The best systems detected old attacks included in the training data, at moderate detection rates ranging from 63% to 93% at a false alarm rate of 10 false alarms per day. Detection rates were much worse for new and novel R2L and DoS attacks included only in the test data. The best systems failed to detect roughly half these new attacks which included damaging access to root-level privileges by remote users. These results suggest that further research should focus on developing techniques to find new attacks instead of extending existing rule-based approaches.

1. Introduction

Heavy reliance on the internet and worldwide connectivity has greatly increased the potential damage that can be inflicted by attacks launched over the internet against remote systems. It is difficult to prevent such attacks by the use of security policies, firewalls, or other mechanisms because system and application software always contains unknown weaknesses or bugs. In addition, complex, often unforeseen, interactions between software components and/or network protocols are continually exploited by attackers. Successful attacks inevitably occur despite the best security precautions. Intrusion detection systems have become an essential component of computer security to detect these attacks before they inflict widespread damage. A review of current approaches to intrusion detection is available in [1]. Some approaches can detect attacks in real time and can stop an attack in progress, provide information about attacks and can help recover, understand the attack mechanism, and reduce the possibility of future attacks of the same type. More advanced intrusion detection systems detect never-before-seen, new, attacks, while the more typical systems detect previously seen, known attacks.

Evaluations of developing technologies such as those used for intrusion detection are essential to focus effort, document existing capabilities, and guide research. For example, yearly DARPA-sponsored evaluations in the speech recognition area have contributed substantially to rapid technical progress [2]. Periodic speech evaluations have focused research on difficult technical problems, motivated researchers to build advanced systems, facilitated information sharing, provided common corpora, and made it easier for new researchers to enter this field and explore alternate approaches [2].

Despite the importance of intrusion detection systems in limiting the damage inflicted by new attacks, we are aware of no evaluations prior to 1998 that (1) generated an intrusion detection evaluation corpus which could be shared by many researchers, (2) were sponsored by DARPA, (3) included a blind evaluation, (4) evaluated a broad range of intrusion attacks including novel and new attacks, (5) focused primarily on off-line analysis, and (6) generated a publically available corpus.
Machine Learning for Malware Detection

Training phase
- Benign executables
- Malicious executables
  → Training
  → Predictive model

Protection phase
- Unknown executable
  → Processing by a predictive model
  → Model decision

Machine Learning: detection algorithm lifecycle
Dos and Don'ts of Machine Learning in Computer Security

Daniel Arp*, Erwin Quiring†, Feargus Pendlebury‡§, Alexander Warnecke†, Fabio Pierazzi‡, Christian Wressnegger¶, Lorenzo Cavallaro†, Konrad Rieck†

* Technische Universität Berlin
† Technische Universität Braunschweig
‡ King’s College London, § University College London
¶ Royal Holloway, University of London and The Alan Turing Institute
‡§ KASTEL Security Research Labs and Karlsruhe Institute of Technology

Abstract

With the growing processing power of computing systems and the increasing availability of massive datasets, machine learning algorithms have led to major breakthroughs in many different areas. This development has influenced computer security, spawning a series of work on learning-based security systems, such as for malware detection, vulnerability discovery, and binary code analysis. Despite great potential, machine learning in security is prone to subtle pitfalls that undermine its performance and render learning-based systems potentially unsuitable for security tasks and practical deployment.

In this paper, we look at this problem with critical eyes. First, we identify common pitfalls in the design, implementation, and addressing security-relevant problems in many application domains, including intrusion detection, malware analysis [69, 88], vulnerability discovery [42, 11, 114], and binary code analysis [43, 142].

Machine learning, however, has limitations: it requires reasoning about statistical evidence, which is a fairly delicate workflow: incorrect assumptions, experimental biases may cast doubts on results that it becomes unclear whether we can trust scientific discoveries made using learning algorithms at all [56]. Attempts to identify such challenges and limitations in specific security domains, such as network intrusion detection, started two decades ago [11, 119, 126] and were extended more recently to other domains, such as malware analysis and website fingerprinting [42, 142].

Our work provides critical insights on pitfalls and their consequences.

2019
Introducing ChatGPT

We’ve trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

Try ChatGPT  Read about ChatGPT Plus
How ChatGPT is changing the way cybersecurity practitioners look at the potential of AI

Derek B. Johnson  December 9, 2022
AI on offense: Can ChatGPT be used for cyberattacks?

The difference between clever and intelligence

By Apoorva Joshi, Devon Kerr

24 May 2023
AI chatbots making it harder to spot phishing emails, say experts

Poor spelling and grammar that can help identify fraudulent attacks being rectified by artificial intelligence

A.I. is helping hackers make better phishing emails

CYBERSECURITY

Almost Human: The Threat Of AI-Powered Phishing Attacks

Emil Sayegh, Contributor @
CEO of Nitrety. Cover all things cloud, cybersecurity & tech
Spear phishing and voice cloning
Here’s how OpenAI’s ChatGPT can be used to launch cyberattacks

Security researchers had the AI create a fake email from a hosting company and inject malware into an Excel file as part of a test.

By Ryan Morrison

Since its release at the end of November, vendors have warned hackers could be using ChatGPT for cyberattacks.

- Social Engineering
- Attack point enumeration
- Foothold assistance
- Producing malicious code
Here’s how OpenAI’s ChatGPT can be used to launch cyberattacks

Security researchers had the AI create a fake email from a hosting company and inject malware into an Excel file as part of a test.

By Ryan Morrison

Since its release at the end of November, there are already at least 5 ways to put OpenAI’s advanced chatbot to work, and a vendor has warned hackers could be using it to democratize cyberattacks.

5 ways threat actors can use ChatGPT to enhance attacks

CSO

Experts Warn ChatGPT Could Democratize Cybercrime

Infosecurity Magazine

Phil Muncaster

UK / EMEA News Reporter, Infosecurity Magazine

Email Phil Follow @philmuncaster

A wildly popular new AI bot could be used by would-be cyber-criminals to teach themselves to carry out attacks.
Cybercriminals are Using ChatGPT to Create Hacking Tools and Code

Experienced and novice cybercriminals are using ChatGPT to create hacking tools and code.

Security researchers have reported that both experienced and novice cybercriminals are using ChatGPT to create hacking tools and code.

One such instance is the Israeli security firm Check Point, which discovered a threat actor creating an underground hacking site by a hacker who claimed to be testing the famous AI tool with different malware strains.

The hacker later compressed and distributed Android malware created by ChatGPT onto the internet. According to Forbes, spyware has the power to steal important files.

The same hacker also demonstrated another program that could install a backdoor onto a victim’s machine.

ITBrief AUSTRALIA

Five ways cybercriminals are making use of ChatGPT

By Anthony Daniel

I built a Zero Day virus with undetectable exfiltration using only ChatGPT prompts

April 4, 2023 | 18 min read

Aaron Mulgrew

artificial intelligence chatgpt malware zero day
PentestGPT – Automate Penetration Testing Empowered by ChatGPT

BY PRIYANGSHU SAHAY • MAY 15, 2023 • ☑️ 5 MINUTE READ

> _PentestGPT /

- Automate penetration testing tool empowered by ChatGPT.
- An interactive mode to guide the attacker.

Getting pwn’d by AI: Penetration Testing with Large Language Models

Andreas Happe
andreas.happe@tuwien.ac.at
TU Wien
Austria

Jürgen Cito
juergen.cito@tuwien.ac.at
TU Wien
Austria

ABSTRACT
The field of software security testing, more specifically penetration testing, requires high levels of expertise and involves many manual testing and analysis steps. This paper explores the potential use of large-language models, such as GPT-3.5, to augment penetration testing with AI sparring partners. We explore two distinct use cases: high-level task planning for security testing assignments and low-level vulnerability hunting within a vulnerable virtual machine. For the latter, we implemented a closed-feedback loop between LLM-generated low-level actions with a vulnerable virtual machine (connected through SSH) and allowed the LLM to analyze the machine state for vulnerabilities and suggest concrete attack vectors which were automatically executed within the virtual machine. We discuss promising initial results, detail avenues for improvement, and close deliberating on the ethics of AI sparring partners.

CCS CONCEPTS
- Security and privacy: Penetration testing; Security and privacy: Malware detection

when stuck. The study also emphasizes that intuition is a big part of detecting vulnerabilities and that knowledge transfer, e.g., from attending Capture-the-Flag (CTF) events, were seen as potential sources of this intuition — can this be partially outsourced to AI models? Using AI-based agents as sparring partners would augment and empower existing human security testers and could counteract the lack of sufficiently educated security professionals. Combining human operators with Alas creates new capabilities instead of cloning existing ones. Furthermore, keeping a human in the loop reduces the potential ethical problems imposed by the use of AIs. [6] Recent research indicates that the efficiency gains provided by the use of AI-based systems are greatest for low-skilled workers [7], augmenting human operators with a generative AI might thus also benefit the training of novice penetration testers.

RQ: To what extent can we automate security testing with LLMs? The rest of this paper explores whether large-language models can be deployed as sparring partners for security profes-
OpenAI's ChatGPT can write impressive code. Here are the prompts you should use for the best results, experts say.

OpenAI's ChatGPT has caused quite a stir in the tech community. Getty/Luis Alvarez

- OpenAI's ChatGPT has been able to produce working lines of code.
- The AI-powered bot has freaked out programmers and caught the attention of tech CEOs.
OpenAI's ChatGPT can write impressive code. Here are the prompts you should use for the best results, experts say.

Beatrice Nolan  Aug 10, 2023, 1:07 PM GMT+2

- OpenAI's ChatGPT has been able to produce working lines of code.
- The AI-powered bot has freaked out programmers and caught the attention of tech CEOs.

OpenAI's ChatGPT has caused quite a stir in the tech community. Getty/Luis Alvarez
ChatGPT CVE Analysis for Red and Blue Team

Red Team use: use ChatGPT to help exploit the CVE and find vulnerabilities in the code. Blue Team use: explain the CVE and how to defend against it.

David Merian · Follow
Published in System Weakness · 2 min read · Mar 9

Many CVE’s are popular targets for Ransomware, according to a report from Securin, as summarized on DarkReading. New CVE’s are published everyday, and oftentimes, they are in very specific—but ubiquitous—software. You can use ChatGPT to summarize what the CVE is, what it
Is it just raising the bar for both attack and defense?
For your defense: Use AI, practice it, and get trained!
For your defense: Use AI, practice it, and get trained!
For your defense: Use AI, practice it, and get trained!

November 2018

From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI

It’s time to replace traditional, rule-based approaches to cybersecurity with “smarter” technology and training.

Karen Bass, Marcia Warkentin, and George Weiserman

April 18, 2023
For your defense:
Use AI, practice it, and get trained!

November 2018

MIT Sloan Management Review

From ChatGPT to HackGPT:
Meeting the Cybersecurity Threat of Generative AI

It's time to replace cybersecurity with

Harvard Business Review

ARTIFICIAL INTELLIGENCE
The End of the

Udemy

Advanced Ethical Hacking: Mastery AI & ChatGPT - Volume 1

A New Attack Impacts Major AI Chatbots—and No One Knows How to Stop It

Researchers found a simple way to make ChatGPT, Bard, and other chatbots misbehave, proving that AI is hard to tame.

Employees Are Feeding Sensitive Biz Data to ChatGPT, Raising Security Fears

More than 4% of employees have put sensitive corporate data into the large language model, raising concerns that its popularity may result in massive leaks of proprietary information.

Robert Lemos
Contributing Writer, Dark Reading

March 07, 2023
Are We Ready to Embrace Generative AI for Software Q&A?

Bowen Xu†, Thanh-Dat Nguyen‡, Thanh Le-Cong‡, Thong Hoang§, Jiakun Liu†, Kisub Kim†, Chen Gong†, Changan Niu†, Chenyu Wang†, Bach Le‡, David Lo†

*North Carolina State University, USA
bxu22@ncsu.edu

†Singapore Management University, Singapore
{bowenxu.2017, jk.liu, kisubkim, chenyu.wang, david.lo}@smu.edu.sg

‡University of Melbourne, Australia
{thanhdatn, congthanhn}@student.unimelb.edu.au, bach.le@unimelb.edu.au

§CSIRO’s Data61, Australia
james.hoang@data61.csiro.au

$University of Virginia, USA
fzhv6en@virginia.edu

¶Nanjing University, China
niu.ca@outlook.com

Abstract—Stack Overflow, the world’s largest software Q&A (SQA) website, is facing a significant traffic drop due to the emergence of generative AI techniques. ChatGPT is banned by Stack Overflow after only 6 days from its release. The main reason provided by the official Stack Overflow is that the answers generated by ChatGPT are of low quality. To verify this, we conduct a comparative evaluation of human-written and ChatGPT-generated answers. Our methodology employs both automatic comparison and a manual study. Our results suggest that human-written and ChatGPT-generated answers are semantically similar, however, human-written answers outperform ChatGPT-generated ones consistently across multiple aspects, specifically by 10% on the overall score. We release the data, analysis scripts, and detailed results at https://github.com/maxxbw54/GAI4SQA.

I. INTRODUCTION

On November 30, 2022, OpenAI, a world-class AI company, has decided to artificially eliminate the open source software question answering (SQA). In this booming era of AI-powered chatbots, traffic to OpenAI’s ChatGPT has been growing exponentially, while traditional Q&A site such as Stack Overflow has been experiencing a steady decline [5]. Specifically, traffic to Stack Overflow was down by 6% every month in January 2022 on a year-over-year basis and was down 13.9% in March 2022 [6]. This phenomenon, however, is concerning due to the lack of empirical evidence on a comparative study on human-written vs AI-generated responses. The empirical evidence is much needed to ensure a balanced and robust development in the field of SQA. In this work, we investigate the following research questions:

- **RQ1:** What are the characteristics of ChatGPT-generated and human-written answers?
- **RQ2:** From the human user perspective, how good are the ChatGPT-generated responses?
In U.S., Regulating A.I. Is in Its ‘Early Days’

While there has been a flurry of activity by the White House and lawmakers over artificial intelligence, rules for the technology remain distant, lawmakers and experts said.
Introducing ChatGPT Enterprise
Get enterprise-grade security & privacy and the most powerful version of ChatGPT yet.

Authors
OpenAI

Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector
In collaboration with GEP

MULTILAYER FRAMEWORK FOR GOOD CYBERSECURITY PRACTICES FOR AI
For your defense: Use AI, practice it, and get trained!

Get in touch with us!

- Prof. Dr. Guido Salvaneschi
- MSc Daniel Sokolowski
- MSc David Spielmann

https://programming-group.com/gcc-2023