AI Security & Trust for CISOs

Paul F. Roysdon, Ph.D.

January 16, 2025

1 Problem Statement

There is valid concern among our our cybersecurity and Chief Information Security Officer (CISO) colleagues in 1) detecting problems in AI that we consume, meaning AI tools or services that we use, and 2) the things we should do or use to make AI trustworthy and reliable.

2 Background

Since the original paper on transformers [**?**] in 2017 establishing the field of Large Language Models (LLMs), we have witnessed a rapid advancement from simple chat bots to advanced LLMs like BART [**?**] in 2019 and GPT-3.5 (i.e., ChatGPT) [**?**] in 2022, that perform zero-shot response to a user input; meaning, the LLM is given a task or question without any prior examples or specific training on that task, relying solely on its pre-existing knowledge to generate a response. Recent advances in LLMs now include "thinking models" (e.g., OpenAI o1 and o3 [**?**], Meta LLaMA [**?**] and DeepSeek R1 [**?**]), that use Chain of Thought (CoT) [**?**], Chain of Reasoning (CoR) [**?**], and Tree of Thought (ToT) [**?**], that take additional time to iterate on an answer and use tools, e.g., APIs from other software, to improve the response to a user input. The most recent model, Titans [**?**], published by Google AI Research at the end of 2024, has the ability to not only think, but use nearly infinite memory at test time (a.k.a. inference) with what they term "surprise mechanisms" that determine the utility of each piece of information and its potentially distant relationship to other information; this is a form of test time training on user prompts [**?**, **?**].

This rapid change from *transformers* to *thinking models* is incredible, and while open-source models [**?**, **?**] are lauded for their open architecture, the security and data lineage of close-source models [**?**] is often questioned. For example, the Fudan University and Shanghai AI Laboratory recently published a paper [**?**] disclosing the details of OpenAI o1, essentially *hacking or leaking* a closed-source model. While lackluster security of AI developers is drawing ire from investors [**?**] pointing out the ease with which a Foreign-National can steal IP, and evaluators [**?**] stating the need for implementation of standard cybersecurity practices to protect AI development and to protect sensitive user information... the AI and cybersecurity industry has yet to respond.

3 Cybersecurity for AI

In 2024 there were reports of attackers jumping from one cloud instance to another and collecting information, or using man-in-the-middle (MITM) attacks to obtain and leak information between an API from an AI company and the cloud service provider, e.g., from MoveWorks (using the ChatGPT API) to OpenAI to AWS cloud. This raises the issue of *data Chain of Custody* (CoC). Who is moving what and to where, and who is responsible in the verification of data protections? Data CoC is more than data provenance [**?**], though provenance is getting more attention. Data CoC is a verification and validation that each 3rd party is protecting your information.

Below are a few examples of cybersecurity concerns for cloud computing and cloud AI:

- DeepLocker [**?**] is an AI-powered malware developed by IBM Research to demonstrate stealth malware techniques using deep learning-based facial recognition to conceal malicious payloads until a target is identified. This technique, demonstrated at BlackHat 2018, bypassed traditional signature-based defenses and only activated when the malware detected its intended victim.

- At DefCon 2021, security researchers demonstrated how GPT-3 can generate highly convincing phishing emails that bypass traditional spam filters [**?**]. The AI-generated emails dynamically adjust based on user responses, making social engineering attacks more effective. Attackers leveraged cloud-hosted GPT-3 APIs to automate large-scale attacks.

- At BlackHat 2020, researchers demonstrated how adversarial machine learning (AML) attacks can manipulate cloud-hosted AI models [**?**]. They injected adversarial noise into images processed by cloud-based AI services, causing misclassifications. This attack technique can be used to fool AI-driven security tools like facial recognition and malware detection systems.

- At NeurIPS 2021, researchers from Cornell University demonstrated how attackers can steal proprietary machine learning models hosted on cloud-based APIs [**?**]. Using model inversion attacks, adversaries extracted training data and model parameters by querying cloud AI services. This poses a significant intellectual property risk for companies deploying AI in the cloud. This is similar to the *model distillation* technique employed by DeepSeek to develop R1 using OpenAI o1.

- At DefCon 2020, researchers demonstrated AI-driven credential stuffing bots using machine learning to predict password variations for stolen credentials [**?**]. Attackers deployed cloud-hosted AI models to optimize brute-force login attempts while avoiding detection. This method significantly increased success rates compared to traditional brute-force attacks.

The key takaways are, 1) AI can be used offensively - Attackers use ML models for phishing, malware evasion, and password attacks, 2) Cloud AI models are vulnerable - Proprietary AI models hosted on cloud APIs can be stolen or manipulated, and 3) Adversarial ML is a growing threat - Attackers can bypass AI-based security systems using perturbed inputs.

4 Fabrication & Deception

Another concern is model trust and verifying that the model is providing reliable information. One issue is *fabrication* (often called "hallucination") the other is *deception* (also called mis-alignment)[**?**], wherein the model is knowingly providing incorrect information. The authors of [**?**] found that OpenAI

o1, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3.1 405B all demonstrate in-context scheming capabilities. In particular, o1 requires no "urging" for deception, while other models like Claude require a bit of prompting. This leads to a valid security issue for analytical tools that rely on such models.

5 AI Security

The prominent work in AI Security can be categorized as:

1. using software or crowd-sourcing to find vulnerabilities in AI models (i.e., the model weights) and in AI libraries, e.g., Python TensorFlow or PyTorch,
2. using AI to find vulnerabilities in traditional software, e.g., *open-source* or *close-source* libraries, and
3. adding "guard-rails" to reduce the generation of undesirable content.

Some of the more prominent work on AI security are:

- **ProtectAI** VulnHunter searches for malware in Python software libraries (e.g., TensorFlow), and ModelScan looks for vulnerabilities embedded in an ML model (i.e., in the weights of a model)[**?**].
- **Google** ProjectZero searches for vulnerabilities in Google-developed software (e.g., Chrome)[**?**].
- **Team Atlanta** (Georgia Tech) recently won the DARPA AIxCC with an LLM specifically designed to find a vulnerability in SQLite3[**?**].

Notice that these works are not focused on security of communication (encryption, prompt injection attacks, MITM attacks, etc.), or chain of custody of data (e.g., from MoveWorks or SourceGraph $\rightarrow$ ChatGPT $\rightarrow$ AWS Cloud), the above examples focus on models, source code for models, or very niche libraries.

6 Verification & Validation

Unfortunately, AI models are now as complex as an airplane or automobile. DO-178B/C [**?**] is a testing standard for verification and validation of classical software in the aerospace industry. Presently there isn't an analog to DO-178B/C for AI. Therefore, to trust and validate AI we need to apply AI to several operating conditions and observe results. This approach is similar to the trust you place in the brakes on your car, or the wings on an airplane. We know that both systems were designed and tested by a manufacturer, so there is some implicit trust, but we use these systems daily and have a personal relationship and trust that we establish with these systems. We also have colleagues and friends who use these systems, and we mentally catalog that experience to build additional trust. Similarly, we observe the failure rate of these systems through national statistics and base some amount of trust on facts and figures.

7 Reasonable Robot - An Example

Let's consider a few examples with and without AI. Most people are not aware of the number of fatal automobile accidents each year in the United States, and they would likely state that airplane accidents are far more frequent or devastating (in terms of number of deaths); keep in mind that we are talking about fatalities, so any accident, regardless of perceived severity, is equally tragic because the result is a fatality. Unfortunately, perception of vehicle safety is often incorrect because our perception is influenced by the media, and while fatal airplane accidents are far less common and car accidents more common, the news often reports what is less common, thereby giving the appearance of more frequent occurrence. Specifically, in 2024 there was roughly 1 collision fatality per 2,576 miles driven by a human, while there was 1 fatality per 100,000,000 passenger miles on commercial and general aircraft [**?**, **?**]. While this is a difficult comparison, because most Americans drive cars versus taking public transit (though there are exceptions in large cities) and most Americans do not fly airplanes, most people will agree that there are a staggering number of miles driven each year, and nearly everyone has spent time in traffic due to a traffic accident (though condition of those involved is often unknown). If we ask the same group to state the level of safety required for autonomous vehicles, most would state that *one order of magnitude* is sufficient (often because that is sufficient in other areas of technical advancement). One order of magnitude is 1 death in 26,000 miles driven by AI. Given this number, most people would backtrack and state that *two or three orders of magnitude* is necessary, i.e., 260,000 or 2.6 million miles, respectively. In fact, in 2024, Tesla had the best record with a fatal accident rate of 1 in over 1 billion miles [**?**].

The above examples emphasize the *need to be reasonable* with our expectations and rules for automation of all types. Certainly, automobiles have safety of life critical systems, e.g., brakes, but AI safety and reliability is equally important in cybersecurity AI applications, if for example, that system resides on a hospital network in an ICU ward.

8 Legal Concerns

The prior section title is inspired by Abbot's seminal book titled "The Reasonable Robot: Artificial Intelligence and the Law" [**?**]. Abbott discusses the merits of applying present US law for automation to applications of AI. US law for automation dates back to the Industrial Revolution, specifically accidents that resulted from manufacturing line automation. Abbott states that these laws are not only applicable to AI, but they are currently being applied to AI in the US court system.

9 Summary

Using AI and building trust with AI tools requires a reasonable approach to requirements, a regular interaction with the tools (to learn the limitations), and verification of tools, workflows and outputs (where applicable).

Acknowledgments

The author thanks colleagues across industry and academia for discussions and publications that informed this work.

Note on References

While there are many research papers that span these topics, this tech note emphasized papers with the following metrics:

- ✓ High Citation Counts - Many references have 1,000+ citations, ensuring they are widely accepted and influential in the AI and Cyber research community.
- ✓ Top AI and Cyber Institutions - Includes research from Google Brain, OpenAI, DeepMind, Stanford, MIT, IEEE, and Nature AI.
- ✓ Authors with High h-Index - leading AI and Cyber researchers such as Ian Goodfellow (h-index: 100+), Geoffrey Hinton (h-index: 150+), Yoshua Bengio (h-index: 140+), and Pieter Abbeel (h-index: 100+).