eBook

WHEN AI GOES ROGUE:

Strengthening the Bars of the Cage

CONTENTS

Rogue AI: The future of cyber threats

Identifying rogue AI

Rogue AI case studies

Mitigating rogue AI risks

Mitigation strategies by rogue AI type

Dr. Yoshua Bengio, regarded as one of the godfathers of artificial intelligence, compares AI to a caged grizzly bear that has been trained to execute tasks in exchange for rewards like fish. As the bear becomes smarter, it figures out how to escape its cage and find fish on its own. Similarly, as AI becomes more agentic, the risk that it could circumvent human control increases.

To keep the bear from going rogue,
we need to strengthen the bars of the cage.
In the same fashion, as generative AI tools are rapidly coming to market and being adopted, it's critical to act quickly to mitigate the associated growing risks.

ROGUE AI: THE FUTURE OF CYBER THREATS

While most of the AI-related cyber threats grabbing headlines today are carried out by adversaries and organized criminals, rogue AI is where security experts are focusing their long-term attention, especially as agentic AI continues to evolve.

So, how can we effectively identify and defend against it?

What is rogue AI?

This term refers to AI systems that are misaligned to the interests and goals of their creators, users, or humanity in general. This can either happen intentionally, where AI services are used to attack any system, or unintentionally, where misalignment is caused by missing safeguards or an error. There are three categories of rogue AI:

ACCIDENTAL (UNINTENTIONAL)

Created by human error or inherent technology limitations like misconfigurations and poor permission control.

COULD YOU BE THE CAUSE?!

SUBVERTED (INTENTIONAL)

Makes use of existing AI deployments and resources. An attacker subverts an existing AI system to misuse it and accomplish their own goals, essentially enabling the AI system to operate differently from its intended design.

WATCH OUT FOR YOUR AI SYSTEMS FALLING IN DANGER FROM OTHERS.

MALICIOUS (INTENTIONAL)

Deployed by attackers for malicious purposes, including using others' computing resources to host rogue AI.

ADVERSARIES HAVE YOU IN THEIR SIGHTS.

IDENTIFYING ROGUE AI

It's important to understand how AI ***should*** behave in order to understand when it's ***not*** behaving. The best way to measure alignment is to simply observe the behavior of AI.

Questions to ask when observing AI include:

- **Is the AI taking actions contrary to expressed goals, policies, and requirements?**
- **Is the AI attempting to access and alter data or systems it shouldn't?**
- **Are there unusual spikes in resource consumption or unexpected delays in processing or response times?**
- **Is the AI displaying biased and discriminatory behavior?**
- **Is the AI generating harmful, deceptive, or offensive content?**

TREND MICRO™

ROGUE AI CASE STUDIES

To further understand and address the potential risks posed by AI - such as excessive functionality, permissions, and autonomy - it's critical to identify specific vulnerabilities in large language models (LLMs) and how they can be exploited by rogue behavior.

VULNERABILITY	EXPLOITED SCENARIO BY ROGUE BEHAVIOR			RESULTING RISK
	ACCIDENTAL	SUBVERTED	MALICIOUS	
Misconfigured capabilities or guardrails	Unintentional errors during set up.	Deliberate modification or evasion of guardrails to enhance capabilities.	Intentional design of capabilities for harmful goals.	**EXCESSIVE FUNCTIONALITY:** The AI system can now perform actions beyond its intended scope, increasing the chances of errors, security breaches, and misuse.
Misconfigured authorization	Unintentional mistakes during set up.	Deliberate escalation of privileges.	Intentional acquisition of all privileges from none.	**EXCESSIVE PERMISSIONS:** Too much control and access to sensitive data, systems, or resources can result in data breaches, data poisoning, and system outages.
Misconfigured autonomy	Unintentional errors in configuring tasks that should require human oversight.	Removal of human oversight to allow autonomous operation.	Designing the system to be completely autonomous from the start.	**EXCESSIVE AUTONOMY:** Without human intervention, the AI system may perpetuate errors, make biased decisions, or engage in harmful behavior without being corrected.

TREND MICRO™

The following case studies illustrate real-world examples of how these vulnerabilities can be exploited.

ACCIDENTAL ROGUE AI

RUNAWAY RESOURCE CONSUMPTION:

Current AI systems can break down tasks into smaller parts and solve them, sometimes at the same time as other AI components. If we don't manage the resources they use carefully, they might end up getting stuck in loops or using up all available resources. If an AI creates a smaller task and has the same resource quota and permissions as the original AI, it could potentially replicate itself.

SUBVERTED ROGUE AI

MODEL POISONING:

Intending to saturate the information space with disinformation, some Russian advanced persistent threat (APT) groups have poisoned many current LLMs. In a quest for as much data as possible (no matter what it is) foundation model creators are ingesting anything they come across. Meanwhile, attackers seeking to sway public opinion create "pink slime" misinformation news feeds and free data for training. This results in poisoned models that parrot disinformation as fact via subverted rogue AI to amplify the Russian APT's narrative.

MALICIOUS ROGUE AI

AI MALWARE:

An attacker drops a small language model on target endpoints, disguising the download as a system update. The resulting program appears to be a standalone chatbot on cursory inspection. This malware uses the anti-evasion techniques of current infostealers but can also analyze data to determine if it matches the attacker's goals. Reading emails, PDFs, browsing history, and more for specific content allows the attacker to stay silent and report back only high value information.

More case studies

COMMUNITY RESOURCES: MITRE ATLAS™ AI Risk Repository

TREND MICRO™

MITIGATING ROGUE AI RISKS

Just as a bear in captivity needs proper containment to stay safe, AI systems require safeguards to ensure they don't go rogue.

1. CONFIGURE:

Specify authorized AI services, restrict data access, and define what tools AI can use. For example, limit the domains an AI with web access can reach.

2. AUTHORIZE:

Assign unique authority to each AI service identity. Clearly specify which actions require human oversight, such as resource creation to solve subproblems. Ensure permissions are properly configured to prevent privilege escalation.

3. INSPECT AND PROTECT:

Rogue AI often results from prompt injections, jailbreaks, or dangerous content. Inspect inputs and outputs to ensure safety and protection. Continuously evaluate to ensure the AI's behavior aligns with its intended purpose.

4. MONITOR:

Track AI services across data, compute resources, devices, workloads, networks, and identity systems. Set up alerts to quickly identify and resolve unexpected behavior.

MITIGATION STRATEGIES BY ROGUE AI TYPE

MITIGATION\TYPE	ACCIDENTAL	SUBVERTED	MALICIOUS
Pre-deployment ❯	Ensure only approved AI systems, data, tools, prompts, and guardrails are utilized.	Protect models, data, and identity used for AI systems.	Allow only specified devices and workloads for AI computing.
Deployment ❯	Limit usage and authority to specific rates based on user roles or use cases.	Enforce guardrails on AI system inputs and outputs.	Ensure that new AI system deployments require human-in-the-loop monitoring.
Post-deployment ❯	Perform regressive evaluation of AI use cases to ensure they remain aligned with goals.	Track new data, tool usage, and resource consumption.	Identify unusual behavior in devices, workloads, and network activities.

COMMUNITY RESOURCES: OWASP LLM checklist

The potential of the AI era is only as robust as its security measures. Building a stronger cage isn't just about reacting to problems after they arise - it's about proactively reinforcing every layer of data and computing that AI models rely on, ensuring the bear remains secure from the outset.

By anticipating future risks and implementing robust safeguards today we can mitigate the threats posed by rogue AI, allowing us to harness its potential for greater good.

Imagine with AI. Secure with Trend.

For details about what personal information we collect and why, please see our Privacy Notice on our website at: trendmicro.com/privacy