Defending Against AI-enabled One-Day and Zero-Day Vulnerability Exploits

In this two-part blog series, we examine how advancements in artificial intelligence (AI) are transforming the cybersecurity landscape. In our first article, we discuss how threat actors now employ AI to automate the creation of code exploits for known (one-day) and unknown (zero-day) vulnerabilities. This capability not only quickens the pace of attacks but also significantly enlarges their potential impact. In our upcoming blog, we will address AI-enabled polymorphic malware and ransomware, designed to slip past traditional, signature-based detections. The emergence of such technologies calls for a revolutionary approach to cyber defense strategies to effectively counter these evolving threats.

Overview of CVE and CVE structure

To comprehend the mechanics behind AI-enhanced threats, it is crucial to understand how vulnerabilities are tracked and documented. The Common Vulnerabilities and Exposures (CVE) system is central to this process. As a globally recognized registry, CVE catalogs publicly disclosed security vulnerabilities in software or hardware, providing essential information that aids in both the identification and mitigation of these threats. By standardizing how vulnerabilities are reported and accessed, the CVE system plays a crucial role in helping defenders stay one step ahead of AI-driven exploits.

Components of a CVE Record:

- **Summary:** Brief outline of the vulnerability.
- **Affected Products:** List of impacted software or hardware, including specific versions.
- **Exploit Overview:** Description of how the vulnerability can be exploited.

Example: CVE-2021-44228 (Log4Shell Vulnerability)

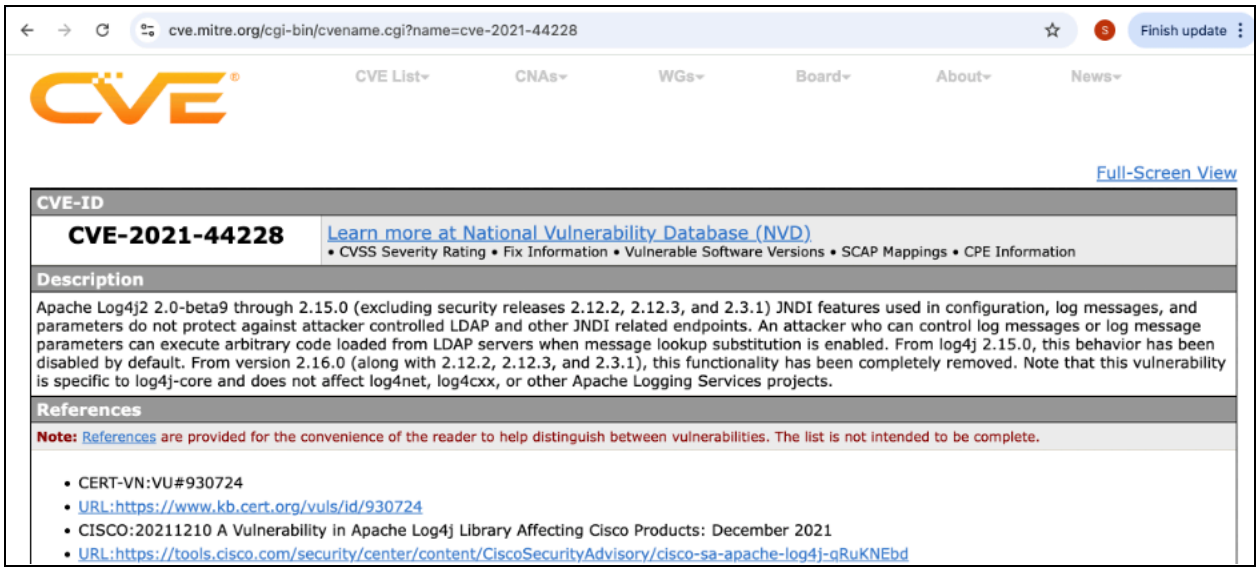

Figure 1: Detailed CVE entry, illustrating the vulnerability description.

This information is crucial for cybersecurity defenders, providing them with the necessary context to assess the relevance of a vulnerability to their environments and strategize their patch management.

Defining One-day and zero-day vulnerabilities

One-Day Vulnerabilities: A one-day vulnerability refers to a security flaw that has been publicly disclosed and assigned a CVE. The disclosure starts a critical period where the software vendor must develop and release a patch to address the vulnerability. Simultaneously, enterprises are tasked with applying this patch throughout their systems, a process constrained by the need to assess potential impacts on business operations. This term specifically denotes the timeframe from public disclosure to patch application, highlighting the urgency and challenges in securing affected systems.

Applying patches is a complex and time-consuming task. Organizations must carefully consider potential disruptions to business workflows and strategize the necessary downtimes and deployment tactics to minimize impact while securing systems effectively. This creates a significant challenge, as delays or complications in the patching process can increase the window of vulnerability, exposing the organization to heightened cyber risks during the implementation period.

Zero-Day Vulnerabilities: In contrast, a zero-day vulnerability is a previously unknown flaw that has not been disclosed publicly, leaving no patch available for mitigation. These vulnerabilities are especially dangerous because they provide attackers the advantage of surprise, without any immediate defense available from the affected software vendors.

AI's Role in Exploiting Vulnerabilities: LLM agents exploit one-day vulnerabilities

Building an effective exploit for a known vulnerability traditionally requires significant technical skill and resources, which can be a barrier for many threat actors. However, recent advances in large

language models (LLM) and artificial intelligence (AI) have dramatically altered this dynamic, significantly reducing the complexity and expertise needed to develop functional exploits.
Research Findings on AI-Enabled Exploits: In May 2024, researchers at the University of Illinois Urbana-Champaign published a pivotal study titled "LLM Agents Can Autonomously Exploit One-Day Vulnerabilities." This study revealed that OpenAI's GPT-4, a leading LLM, could autonomously exploit 87% of one-day vulnerabilities using only publicly available CVE descriptions. This capability marks a significant shift, as it demonstrates that LLMs can interpret vulnerability details and automatically generate various exploit codes tailored to those vulnerabilities.
Implications for Cybersecurity: This development has profound implications for cyber defense strategies. The ability of LLM agents to rapidly produce multiple exploit variations from a single CVE description not only speeds up the exploitation process but also expands the potential attack vectors against vulnerable systems. It underscores the urgency for cybersecurity teams to reevaluate their patch management and defensive tactics to counter these AI-powered threats.

World's first zero-day found by AI

On the heels of the findings of LLMs being able to exploit one-day vulnerabilities, Google published findings that introduce an alarming prospect for cybersecurity. Researchers at Google's Project Zero and DeepMind have developed an LLM agent named "Big Sleep," specifically designed to uncover real security vulnerabilities within commonly used code. Notably, Big Sleep has identified a critical exploitable stack buffer underflow in SQLite, a widely utilized open-source database engine. This discovery is the first public declaration of a zero-day vulnerability found by an LLM agent.
Implications of AI-Discovered Zero-Days: The ability of LLMs to analyze source code and pinpoint zero-day vulnerabilities signifies a pivotal shift. Attackers are now poised to use this technology to generate a potentially vast number of new zero-day threats. As open-source software constitutes the backbone of the global computing infrastructure, the readily accessible source code significantly simplifies the process for attackers to exploit these vulnerabilities using LLM agents. This opens up a new frontier of cybersecurity challenges, necessitating urgent and innovative responses to prevent widespread exploitation.

Fundamental shift in the threat landscape and Implications on cybersecurity

The landscape of cybersecurity is undergoing a profound transformation due to the capabilities of LLM agents to identify both one-day and zero-day vulnerabilities. This evolution marks a significant departure from traditional security practices and necessitates a comprehensive shift in cybersecurity tooling.

- **Challenges with Traditional Security Controls:** Historically, cybersecurity defenses have relied on identifying "known bad" behavior—actions previously recognized as malicious based on established patterns of misconduct. This approach, often referred to as reactive detect-and-respond strategies, involves analyzing raw data, logs, and utilizing analytics and machine learning to detect threats. The effectiveness of these methods hinges on matching observed activities to known patterns of malicious behavior.

- **Limitations Exposed by AI-Enabled Threats:** With the advent of one-day and zero-day exploits facilitated by AI technologies, these traditional methods face significant challenges. Such exploits do not exhibit “known bad” behaviors, leaving traditional security systems without a foundation for identifying and classifying these actions as threats. This gap in detection creates substantial vulnerabilities, presenting an exploitable attack surface that threat actors are poised to leverage.
- **Elevated Cyber Risk Due to Advanced AI Tooling:** The continual enhancement of AI and LLM sophistication means that threat actors are increasingly equipped with more advanced tools for executing targeted exploits. This capability leads to an elevated level of cyber risk, as attackers can bypass conventional security measures to exploit newly discovered vulnerabilities swiftly and efficiently.

Revaluation of defense strategies

As the threat landscape has continually evolved, cybersecurity and defense strategies have had to adapt to keep pace with new challenges. Initially, security efforts were heavily concentrated on perimeter-based defenses and prevention controls designed to shield critical entry points. However, as threat actors began exploiting personal identities and utilizing sophisticated social engineering tactics, it became evident that perimeter defenses alone were insufficient. This realization prompted a strategic pivot toward reactive measures focused on detecting and responding to threats after breaches had occurred.

With the rise of AI technologies and the increasing prevalence of one-day and zero-day exploits, there is now a pressing need to further evolve defense strategies. Modern defense teams must transition to a more proactive and preemptive approach that anticipates potential threats before they manifest. This strategy emphasizes the use of advanced analytics and threat intelligence to proactively identify and mitigate risks, moving beyond reliance on historical data and known attack patterns. This shift is crucial for staying ahead of sophisticated attackers and effectively safeguarding against the dynamic threats posed by advanced AI and automation technologies.

Role of cyber deception in a preemptive defense strategy

Preemptive cyber defense represents a proactive, prevention-based approach to cybersecurity, moving away from the traditional reactive models of threat detection and response. This forward-looking strategy proves highly effective against the evolving threat landscape because it anticipates threats before they manifest, rather than reacting to attacks as they occur.

Central to this approach is cyber deception, which differs fundamentally from conventional detect-and-respond tactics. Cyber deception involves setting traps or decoys that are designed to be attractive to attackers, predicting their goals and creating scenarios that lure them into engaging with these decoys instead of real targets. By monitoring interactions with these traps, defense teams can identify and mitigate threats early in the attack process, effectively disrupting the adversary’s plans without prior knowledge of their specific attack techniques.

This method of deception-based detection is particularly effective against modern threats, including AI-generated attacks and exploits leveraging one-day and zero-day vulnerabilities, which typically bypass traditional security measures. By focusing on the attackers’ unchanging goals—such as elevating privileges, exfiltrating sensitive data, and compromising critical systems—cyber deception

strategies allow defenders to engage attackers on their own terms. Strategically placed traps safeguard that even if the specific tactics of the attackers vary, the underlying intent is revealed, allowing for preemptive countermeasures that are independent of the attackers' chosen methodology.

Example scenario: deception to detect one-day and zero-day exploits

Consider a zero-day exploit targeting the SQLite database engine. As threat actors identify additional one-days and zero-days, defense teams can proactively identify critical assets that use SQLite and deploy tailored decoys that appear as legitimate SQLite databases.

These deceptions include decoys that mimic databases powered by the SQLite engine. Defense teams utilize AI to formulate strategies that make these decoys highly enticing to exploit. This involves setting specific characteristics for the decoys—such as unique hostnames, careful placement within the network, and optimal decoy counts—to maximize their appeal and mislead attackers away from real assets.

When an attacker, armed with a sophisticated zero-day or one-day exploit for SQLite, encounters one of these decoys, it serves as a lure. The interaction with the decoy is designed to be an unmistakable red flag, as these decoys are not used for any legitimate activity. Thus, any engagement with the decoy is a definitive indicator of malicious intent. This allows defense teams to not only divert the attacker's efforts but also to gather intelligence on attack methods and vectors in real-time.

This method of proactive defense provides significant advantages. It allows cybersecurity teams to preemptively identify and counter AI-enabled threats, including those leveraging the advanced capabilities of large language models (LLMs) for automation and stealth. By strategically deploying these decoys, defense teams can effectively manage the threat landscape, turning potential vulnerabilities into controlled traps that protect critical assets while enhancing the organization's ability to respond to evolving cyber threats.

Summary and Preview of the Next Blog

In this blog, we've explored the transformative impact of AI on the threat landscape, particularly focusing on how advanced technologies like LLMs are reshaping the way cybersecurity defenses are structured. We've covered how preemptive cyber defense strategies, especially cyber deception, play a pivotal role in effectively countering the sophisticated exploits powered by AI, such as one-day and zero-day vulnerabilities. By utilizing decoys that lure attackers away from real assets, defense teams can proactively manage threats and gather valuable intelligence to enhance their security measures.

Looking ahead, our next blog will explore further into AI-enabled threats, focusing on polymorphic malware and ransomware. These adaptive threats represent a significant evolution in cyberattack techniques, capable of bypassing traditional signature-based detection systems. We will examine the challenges they pose and discuss innovative strategies that can be employed to detect and neutralize these elusive threats.