



AI-POWERED DECEPTION

Defending Against Adversarial Misuse of AI: A PARADIGM SHIFT

2025 Southeast IT & Security Leaders Forum
CHARLESTON, SC | 09-11 MARCH

Attackers Have an Advantage

- **258 days** to **detect**
- **70+** days to **eradicate**
- **67%** of the time **attackers out themselves!**
- Average **cost** is about **\$4M**
- **Longer** Detection Time = Increasing **Cost**



Detection is the Top Priority

https://fieldeffect.com/blog/real-cost-data-breach?utm_source=chatgpt.com
https://prowritersins.com/cyber-insurance-blog/average-cost-of-a-data-breach/?utm_source=chatgpt.com
https://www.varonis.com/blog/data-breach-response-times?utm_source=chatgpt.com
2024 Verizon DBIR

And now they have AI!



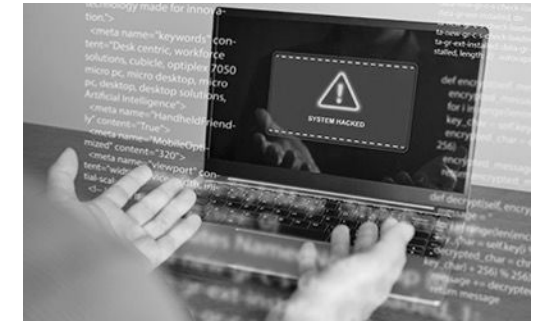
One day and zero-day vulnerability exploits



Polymorphic malware



Identity-driven Attacks



Insider threats

Adversaries are leveraging AI and LLMs to exploit security gaps

Adversaries leverage AI and LLMs to exploit security gaps

Teams of LLM Agents can Exploit Zero-Day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, Daniel Kang
University of Illinois Urbana-Champaign
{rrfang2, bindu2, akulg3, qiusiz2, ddkang}@illinois.edu

Abstract

LLM agents have become increasingly sophisticated, especially in the realm of cybersecurity. Researchers have shown that LLM agents can exploit real-world vulnerabilities when given a description of the vulnerability and toy capture-the-flag problems. However, these agents still perform poorly on real-world vulnerabilities that are unknown to the agent ahead of time (zero-day vulnerabilities).

In this work, we show that *teams* of LLM agents can exploit real-world, zero-day vulnerabilities. Prior agents struggle with exploring many different vulnerabilities and long-range planning when used alone. To resolve this, we introduce HPTSA, a system of agents with a planning agent that can launch subagents. The planning agent explores the system and determines which subagents to call, resolving long-term planning issues when trying different vulnerabilities. We construct a benchmark of 15 real-world vulnerabilities and show that our team of agents improve over prior work by up to 4.5x.

FORBES > INNOVATION > CYBERSECURITY

Google Claims World First As AI Finds 0-Day Security Vulnerability

Davey Winder Senior Contributor @

Davey Winder is a veteran cybersecurity writer, hacker and analyst.

Follow

4

Nov 5, 2024, 06:55am EST

Update, Nov. 05, 2024: This story, originally published Nov. 04, now includes the results of research into the use of AI deepfakes.

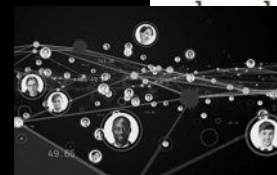
An AI agent has discovered a previously unknown, zero-day, exploitable memory-safety vulnerability in widely used real-world software. It's the first example, at least to be made public, of such a find, according to Google's Project Zero and DeepMind, the forces behind Big Sleep, the vulnerability agent that spotted the



One day and zero-day vulnerability exploits



Polymorphic malware



Identity-driven Attacks



Insider threats

AI-driven malware targeting identities

Detecting Identity Threats



What are examples of AD attack techniques? 17 commonly used techniques:

1. Kerberoasting
2. Authentication Server Response (AS-REP) Roasting
3. Password spraying
4. MachineAccountQuota compromise
5. Unconstrained delegation
6. Password in Group Policy Preferences (GPP) compromise
7. Active Directory Certificate Services (AD CS) compromise
8. Golden Certificate
9. DCSync
10. Dumping ntds.dit
11. Golden Ticket
12. Silver Ticket
13. Golden Security Assertion Markup Language (SAML)
14. Microsoft Entra Connect Compromise
15. One-way domain trust bypass
16. Security Identifier (SID) history compromise
17. Skeleton Key

“ Detecting Active Directory compromises can be **difficult, time consuming** and resource intensive ... many Active Directory compromises **exploit legitimate functionality** and **generate the same events** that are generated by **normal activity** ”



Traditional detect and respond solutions are not sufficient

Traditional controls look
for
“known threats”

A sufficiently advanced threat actor is
indistinguishable from a competent system
administrator.

583%

INCREASE IN KERBEROASTING
ATTACKS (A SUB-TECHNIQUE
OF STEAL OR FORGE KERBEROS
TICKETS), WITH VICE SPIDER
RESPONSIBLE FOR 27% OF ALL
KERBEROASTING ATTACKS



An [announcement](#) by the BlackCat group suggests the motives for updating the ransomware, indicating that BlackCat ransomware “has been completely rewritten from scratch” and that “The main priority of this update was to optimize detection by AV/EDR.”

Dear Adverts!

We are pleased to inform you that testing of basic features ALPHV / BlackCat 2.0: Sphynx is completed. All affiliate plus when creating a new configuration will be offered the opportunity to choose a version. To maximize the non-detectable binary time, the software will only be available for active affiliate plus.

The code, including encryption, has been completely rewritten from scratch. By default all files are frozen. The main priority of this update was to optimize detection by AV/EDR, the following steps were taken to achieve this goal:

- A new technique has been added to mask the encryption process similar to file archiving. When you activate this feature, you can also configure the masking strategy. Currently only zip masking mode is available, the list will be expanded.
- The --access-token startup key has been removed. Now any of the generated keys in launch_keys.txt file can be used for launching files, e.g:

```
C:\alphv.exe 195ToG0F -oAC --AUC -X99odn4 -pdTvb -ewi -vyJtK00Tx8dQ7QL9
```

```
./alphv dm -5 -sEsx -Qaf0uyRY -Tn3588c5 -fb88noL0X -yUT
```

- When creating the configuration, it is now possible to define a readme distribution strategy. This is a red rag for any antivirus, which is not always possible to disable. Therefore we recommend to make several configurations. In the first build, disable readme distribution completely and encrypt user workstations and/or less visited servers, and in the second build, encrypt and distribute readme.

- Also, it is now possible to define a strategy for renaming encrypted files.

 sabled: do not rename (do not add an extension),

 l: rename everything (add .xyzwz extension),

 lyunknownextensions: only rename unknown extensions. It is recommended not to rename files when AB is enabled.

Undoubtedly, each feature deserves to be described in a separate article, but we'll leave the work to our favorite antivirus analysts, and for you we'll just list the most important changes.

- The process of interaction with the network has been completely redesigned. Network orbs search and encryption algorithm has been improved.
- Completely redesigned process impersonation logic for encrypting network files without rights.
- System resource usage has been architecturally redesigned. Encryption speed has increased many times over.
- Support for Glibc 2.5+, which was released in 2006, has been added for *nix. This means that it is now possible to encrypt very old *nix-like systems. Including ESXI<5.0.
- The -v / -ui keys have now been merged. If you want to monitor the encryption process, you have to add to the command line the -v key, e.g:

```
C:\alphv.exe 195ToG0F -oAC --AUC -X99odn4 -pdTvb -ewi -vyJtK00Tx8dQ7QL9 -v
```

```
./alphv dm -5 -sEsx -Qaf0uyRY -Tn3588c5 -fb88noL0X -yUT -v
```

Otherwise, the locker process will go into a background.

- Test simplified the configuration editing interface. In case of urgent need, we will return the ability to point config

Preemptive cyber defense: a paradigm shift in security

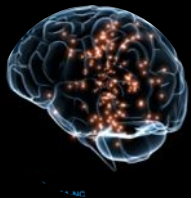
Technology Description

Preemptive cyber defense is an emerging category of cybersecurity technologies that are designed to help organizations improve their ability to defend against:

- AI-enabled threat actors
- Advanced malware
- Zero-day vulnerabilities
- Ransomware
- A wide range of related threats that typically cannot be stopped using only a traditional “detect and respond” approach

The Gartner logo is displayed in white text on a dark blue rectangular background. The word "Gartner" is written in a bold, sans-serif font, with a registered trademark symbol (®) to the upper right of the letter "r".

What is a different approach



most attacks, insider
of APTs

many layers fail

no one while

Advanced Deception Tech: A Game Changer for Defense

Dynamically
Change the
Landscape 
Observe & Analyze

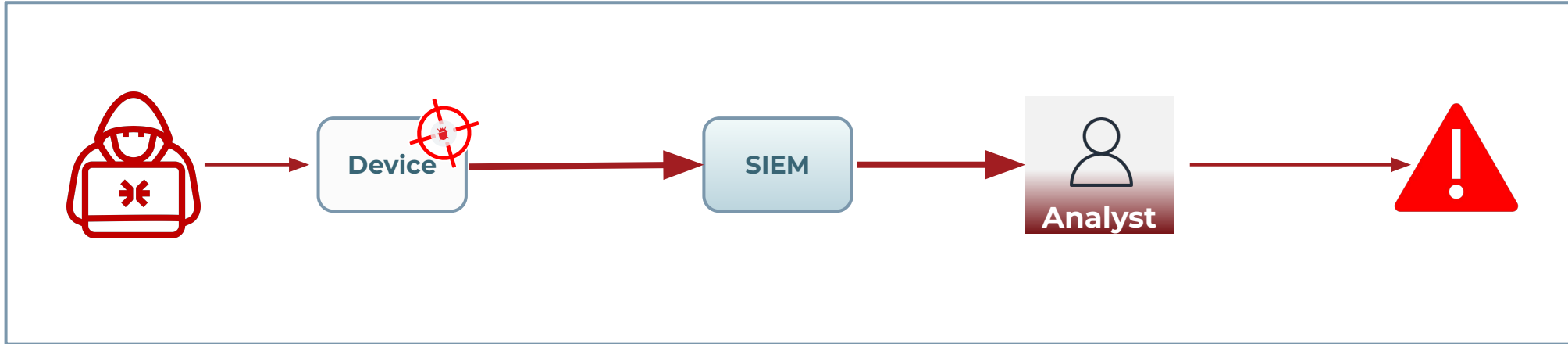
Give Attackers the 
Opportunity to Move

Defenders can 
engage

Real Time threat 
detection

Deception Detects Differently

Traditional



Deception



Implement Traps, Lures, Tripwires that look like the real thing



Deceptive Credentials

- Memory cache
- RDP profiles
- SSH profiles
- Browser cache

ENDPOINTS



Tripwires

- Beacon documents
- Fake Anti Virus
- Ransomware baits



Honeytokens

- IAM Users/Roles
- Access Keys
- Secrets

CLOUD & NETWORK





Honeytoken Accounts

- Deceptive users
- Service accounts

IDENTITY STORES



DECOYS

- Servers 
- Workstations 
- Applications
- Data repositories

CLOUD & NETWORK

Defence in depth

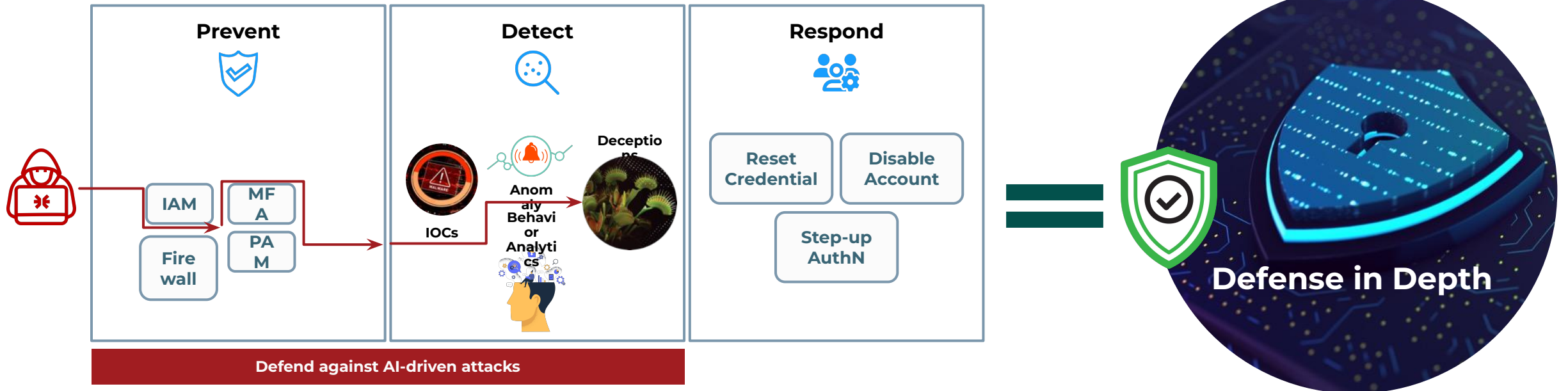
Traditional detection layers include

- Log analytics
- Signature-based detection
- Behavior-based detection
- Anomaly-based detection

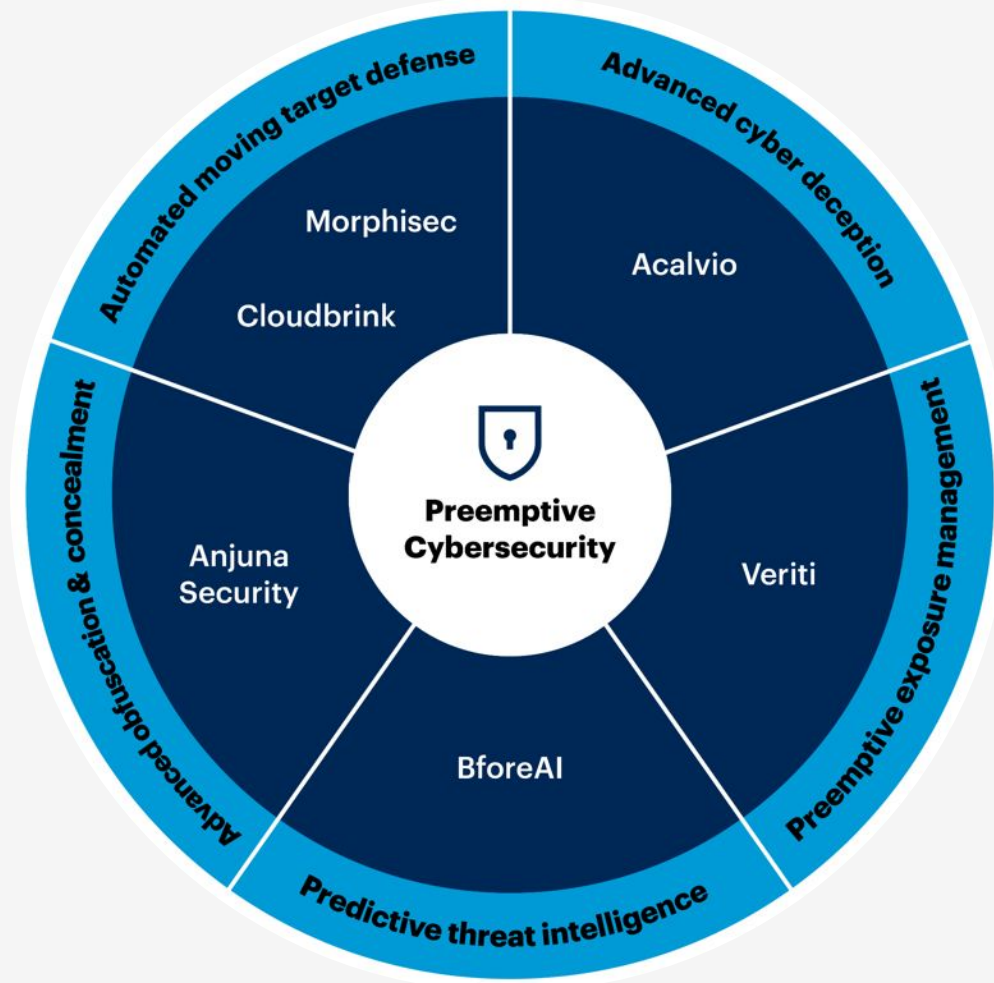


Deception technology

- Set traps for the adversary
- Deceptions are not used in existing IT workflows
- Any usage of the deceptions is indicative of malicious activity



Advanced deception to defend against AI-driven attacks



Tech Innovators

Acalvio Uses Advanced Deception That Looks Too Real to Resist

[Back to top](#)

Analysis by: Luis Castillo, Isy Bangurah

Nature of the Innovation

Acalvio innovates by offering a full-stack, agentless and scalable autonomous deception platform to protect assets across IT, OT and cloud environments through its ShadowPlex Advanced Threat Defense and ShadowPlex Identity Protection products. Its solutions represent networks, endpoints, applications, data and identities as decoys, breadcrumbs and baits capable of interacting with adversaries at multiple levels (high, medium, low) to create a realistic deceptive environment. The use of playbooks with prebuilt deception solutions simplifies the deployment of deceptions, accelerating time to value.

Additionally, AI algorithms automate believable deception recommendations, reducing the administrative burden of managing deceptions. The ability to discover cached credentials on endpoints, assess Active Directory (AD) configuration and map attack paths based on AD object security relationships provides visibility into the endpoint attack surface, enabling the identification of exposures before attacks can be carried out.

<https://www.gartner.com/reprints/?id=1-2JYAR57H&ct=250114&st=sb>

Today's Modern Deception Platform AI for AI



Proven Efficacy

Ease of use

ML for deception recommendation

Proven in production scenarios and red teaming

Comprehensive Set of Deceptions

Hundreds of deceptive artifacts

Decoys across interaction levels

Honeytokens in identity stores and endpoints

Enterprise-wide Coverage

Data Center, Multi-cloud, OT & IT Networks, Identity Stores

Scalable across hundreds of thousands of endpoints

Prebuilt Use Cases

Support for a rich set of use cases

Deception Playbooks for packaged solutions

Threat hunting, adversary engagement

Rich Set of Integrations

Native integrations with security platforms

Interoperability with existing security operations



AI-POWERED DECEPTION

THANK YOU!

scott@acalvio.com

2025 Southeast IT & Security Leaders Forum
CHARLESTON, SC | 09-11 MARCH