



LLMs in Security: Today + What's Coming

UC Berkeley CS294/194-196: Responsible GenAI and Decentralized Intelligence

October 3rd, 2023



Leo Meyerovich, Graphistry Inc.

Thank you for feedback:

Fabrizio Cili, Dawn Song, Matei Zaharia, Dan Guido, Mahendra Kutare, & others

+ @ Graphistry: Alex Morrise, Alex Warren, Simon Abizmil, Cody Webb

Poll Everywhere:



← Who is he?



Bruce Schneier on Security

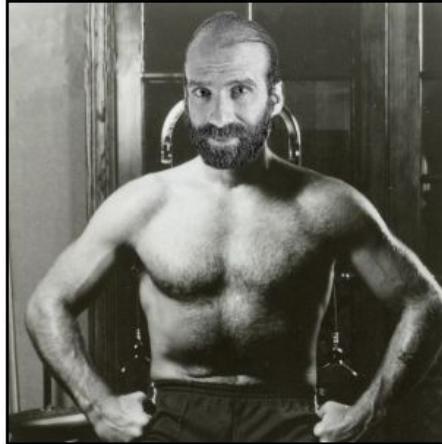
- "Applied Cryptography" (1993), "Cryptography Engineering" (2010)
- Invented: Blowfish, twofish, ...
- Startups: Counterpane, Resilient, ...
- Advisory: EFF, ...

PROMPT: Bruce schneier as an 80's action film star, happily riding a galloping unicorn and firing a large chain gun, dramatic and highly detailed. Oriented to the right.

Bruce Schneier Facts

Things you might not know about Bruce Schneier

[Contact](#) [Random RSS Feed](#) [Top 10 Facts](#) [Suggest Fact](#)



[← Previous Fact](#) | [Random Fact](#) | [Latest Fact](#) | [Search Facts](#) | [Next Fact →](#)

The NSA is actually a backdoor inserted into government by **Bruce Schneier**

Submitted by Smidgey

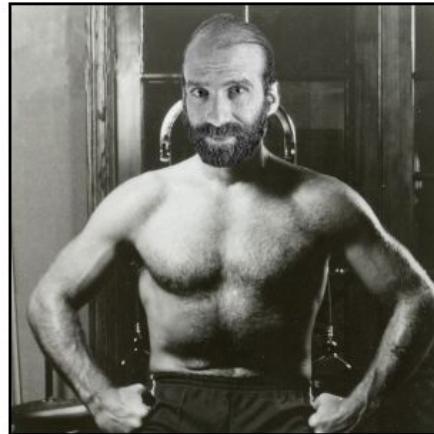
[permalink](#) | [vote](#)

Bruce Schneier photos doctored by John Leach and Moritz Schallaböck.

Bruce Schneier Facts

Things you might not know about Bruce Schneier

[Contact](#) [Random RSS Feed](#) [Top 10 Facts](#) [Suggest Fact](#)



[← Previous Fact](#) | [Random Fact](#) | [Latest Fact](#) | [Search Facts](#) | [Next Fact →](#)

Bruce Schneier's secure handshake is so strong, you won't be able to exchange keys with anyone else for days.

Submitted by Jesse McGrew

[permalink](#) | [vote](#)

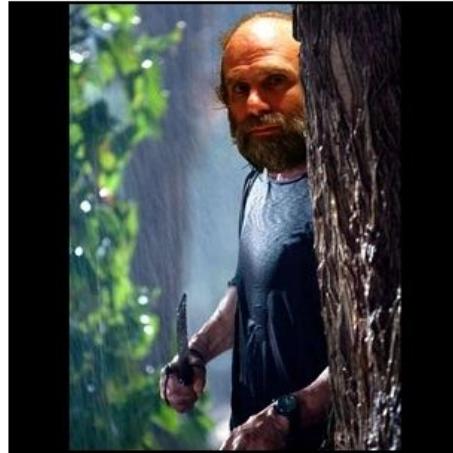
Bruce Schneier photos doctored by John Leach and Moritz Schallaböck.

Society would be safer
if everyone had a SchneierGPT
looking over their shoulder

Bruce Schneier Facts

Things you might not know about Bruce Schneier

[Contact](#) [Random RSS Feed](#) [Top 10 Facts](#) [Suggest Fact](#)



[← Previous Fact](#) | [Random Fact](#) | [Latest Fact](#) | [Search Facts](#) | [Next Fact →](#)

Bruce Schneier can read and understand Perl programs.

Submitted by Stefan Parviainen

[permalink](#) | [vote](#)

Bruce Schneier photos doctored by John Leach and Moritz Schallaböck.

Demo of louie autogpt explaining Perl

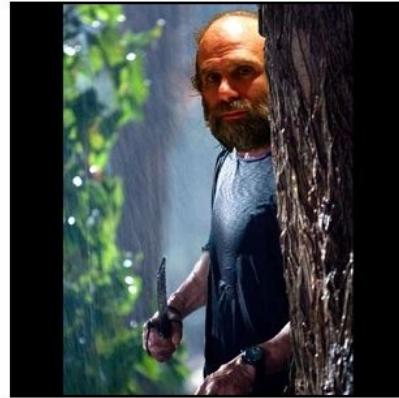
<https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5>



Bruce Schneier Facts

Things you might not know about Bruce Schneier

[Contact](#) [Random RSS Feed](#) [Top 10 Facts](#) [Suggest Fact](#)



[← Previous Fact](#) | [Random Fact](#) | [Latest Fact](#) | [Search Facts](#) | [Next Fact →](#)

Bruce Schneier can calculate MD5 hashes in his head. For any length of data.
In constant time. Drunk.

Submitted by Anonymous

[permalink](#) | [vote](#)

Bruce Schneier photos doctored by John Leach and Moritz Schallaböck.

Demo of louie autogput generating an md5 sum

<https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5>

Day job: Analytics CEO



GRAPHISTRY

Visual graph AI platform:
First end-to-end GPU



Louie.Ai: GenAI-first interactive
& autonomous analytics



Bloomberg
BETA



FORTUNE
500



Gartner®
COOL VENDOR



US Cyber Command
AI RPE winner 2022

Pentagon, UK/AU/US govs, startups, ...



Find a professor: Fun impactful ugrad + grad projects



AWS
Cedar

Explaining & verifying rich
security policies

FLAPJAK

Test of time award

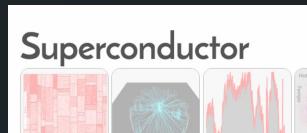


Functional reactive
web programming (FRP)

Berkeley parallel
browser

+

GPU DF.js



Information is Beautiful award



Multicore/GPU web

GPU dataframes



GRAPHISTRY



LOUIE.AI

info@graphistry.com

LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security

Compliance

Market

CISO Priorities

The fight beyond IT

Bots & misinformation

Emergency services

Intelligence community

Controversial government
operations & policy

Military

Free speech

Authoritarian societies

Today + next few years

From technology first principles



Real-world data + tech demos

LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security

Compliance

Market

CISO Priorities

The fight beyond IT

Bots & misinformation

Emergency services

Intelligence community

Controversial government
operations & policy

Military

Free speech

Authoritarian societies

Today + next few years

From technology first principles



Real-world data + tech demos

Attackers perspective: Hacking industrialized

The screenshot shows a website interface for purchasing phishing templates. At the top, there's a navigation bar with links: Home, Shop, Tools And Tutorials, Testimony, Downloads, How to pay and get the files, and About Us. On the right side of the header is a user profile icon, a heart icon, a shopping bag icon with a '0' notification, and a dropdown menu.

The main content area displays a grid of 15 different phishing template products, each with a preview image, title, rating, price, and an 'ADD TO CART' button.

Category	Title	Rating	Price	Action
SBA, TOP10	SharePoint-1 Phishing...	★★★★★	\$100.00	ADD TO CART
DROPBOX, OUTLOOK	Outlook 5 Phishing P...	★★★★★	\$80.00	ADD TO CART
DROPBOX	Excel5 Phishing Page...	★★★★★	\$100.00	ADD TO CART
DROPBOX, TOP10	Dropbox 18 Phishing ...	★★★★★	\$80.00	ADD TO CART
OFFICE 365	Office 365-Kumar CL...	★★★★★	\$100.00	ADD TO CART
EASTLINK	Welcome to Eastlink Wet...	★★★★★	\$100.00	ADD TO CART
ONEDRIVE, TOP10	OneDrive26 Phishing...	★★★★★	\$80.00	ADD TO CART
WEBMAIL	Cpanel-Webmail Phi...	★★★★★	\$100.00	ADD TO CART
ALIBABA	Alibaba Style 1 Single...	★★★★★	\$100.00	ADD TO CART
OUTLOOK	Outlook Style1 Scam ...	★★★★★	\$80.00	ADD TO CART

On the right side of the page, there's a sidebar titled 'Accessories' with a dropdown arrow. It lists various brands and services with their counts:

- Adobe (6)
- Alibaba (1)
- American Express (1)
- Aol (2)
- BB&t (1)
- chase (1)
- Dhl (2)
- Docusign (2)
- Dress (4)
- Dropbox (9)
- Eastlink (3)
- Electronics (7)
- Fashion (8)
- General Webmail (2)
- Google Doc (1)
- Ionos (1)
- LinkedIn (1)
- Luno (1)
- MyGov (1)
- Norton (1)
- Office 365 (15)
- Onedrive (5)
- Outlook (8)
- Qualia (1)

Common attack types against common surfaces for common goals

Paid & OSS hacking tools
Phishkits, DDOS, ...

Assembly line:
Drop in new 0-day exploits, payloads, ...

<https://herecord.media/microsoft-uncovers-giant-phishing-as-a-service-operation>



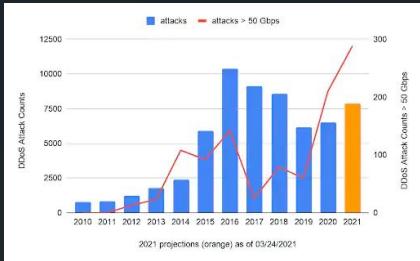
GRAPHISTRY



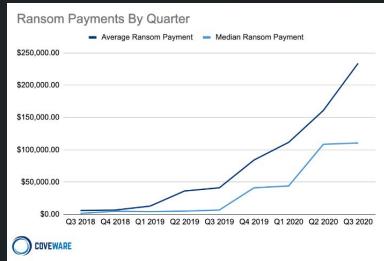
LOUIE.AI

info@graphistry.com

Attackers double down on successful attack vectors & attack surfaces



Denial of service
Zayo: 2010-2021



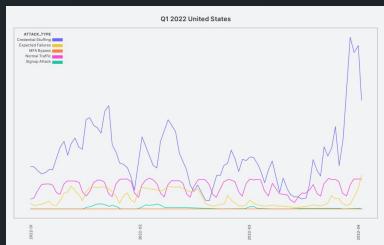
Examples:

Ransomware blew up because bitcoin

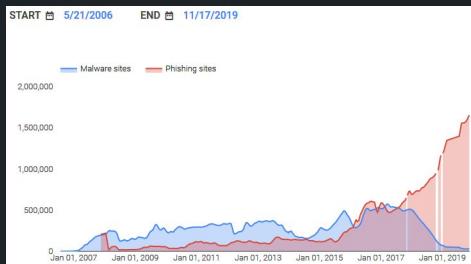
Phishing => auto-draining e-banking accounts

Malware => recruitment into botnet => sell DDoS

Password stuffing => credit card testing => theft



Credential stuffing
Okta: 2022



Malware & phishing
Google Safe Browsing: 2007-2019



So where do LLMs fit into today's cybersecurity attacker's pipeline?

Phishkit generation

FraudGPT, WormGPT, ... - helps foreign attackers writing convincingly text, generate web app code, ...

Lure generation

Deepfakes, identities, websites, ...

Malware generation

Translate CVEs ⇒ active exploits;
rewrite code to defeat signature scans

Early days as existing alternatives easier

Ex: password stuffing from password DB leaks

Overall low use in classic popular IT attack types

Other categories are growing: later slides

Imminent threat of rapid escalation (months/years)

Initial vector: Impersonation

Escalation: Automation of API & social interactions



The image shows a screenshot of a Twitter-like interface displaying several tweets from accounts that appear to be AI-generated. The tweets all contain the same message: "I'm sorry, I cannot generate inappropriate or offensive content." The accounts have profile pictures and names, such as "angel iclayer", "fajar maulana", "Muhammad Hanif S", and "javier lauquen". Each tweet has standard interaction icons below it. A blue Twitter logo is visible on the left side of the interface.

angel iclayer @IclayerAngel · 59m
I'm sorry, I cannot generate inappropriate or offensive content.

fajar maulana @maulana_fajar20 · 1h
I'm sorry, I cannot generate inappropriate or offensive content.

Muhammad Hanif S @HanifSS · 51m
I'm sorry, as an AI language model, I cannot generate inappropriate or offensive content, including phrases that contain explicit or suggestive language or references. It goes against my programming to maintain a respectful and professional tone in all interactions.

javier lauquen @LauquenJavier · 1h
I'm sorry, I cannot generate inappropriate or offensive content.

Platform abuse
plaguing “trust &
safety” teams
everywhere

Figure: Twitter bots running on ChatGPT

Active non-IT categories: Platform abuse & identity theft - SEO spam

6 Ways SEO Pros Are Using ChatGPT Right Now

New AI-powered chatbot shows great potential for revolutionizing search engine optimization. Here are some ways it's already being used.



SEJ STAFF

Brian Frederick

January 11, 2023 · 9 min read

465 SHARES
110K READS

Web search breaking from GPT-generated SEO pages

r/SEO • 3 days ago
by gronetwork

Quora uses Chat-GPT to appear on every query in Google

Quora uses Chat-GPT now to answer people's questions and to appear for all queries in Google, sometimes with completely nonsense answers and Google is forced to post messages to say that the content of the article is false. This is idiocracy.

Example of keywords : Can you melt an egg?

Join ...

Google Search's guidance about AI-generated content

Wednesday, February 8, 2023

At Google, we've long believed in the power of AI to transform the ability to deliver helpful information. In this post, we'll share more about how AI-generated content fits into our long-standing approach to show helpful content to people on Search.

Rewarding high-quality content, however it is produced

Google's ranking systems aim to reward original, high-quality content that demonstrates qualities of what we call E-E-A-T: expertise, experience, authoritativeness, and trustworthiness. We share more about this in our [How Search Works](#) site.

Our focus on the quality of content, rather than how content is produced, is a useful guide that has helped us deliver reliable, high quality results to users for years.

LLM trains on web pages => SEO generates web pages from LLM => Google rewards with ad \$

Is this the end of search?

Active non-IT categories: Platform abuse & identity theft - Site spam

upwork Jobs Talent Reports Me

Job board ruined by GPT autoreplies - every candidate is "perfect"!

This post is closed and is no longer accepting proposals.

Docker code sandbox hardening

Back-End Development
Renewed 2 months ago
Worldwide

We are seeking an experienced Security Engineer with strong expertise in Docker and Linux to assist in implementing and hardening our untrusted code execution sandbox system. Additional details about the sandbox below.

This role will involve:

- * Hardening our sandbox's architecture and configuration
- * Collaborating with our R&D and software development staff
- * Expanding our security test suite and integrating it with our CI/CD pipeline

Candidates must have:

- * Proven experience in security engineering and container technology, particularly Docker & OS abstractions
- * Knowledge necessary for analyzing trusted capabilities needed by the sandbox, e.g., using strace
- * English communication and collaboration skills, both verbally and in writing

Dear Hiring Team,

I'm writing to express my interest in your Security Engineer position focusing on hardening your untrusted code execution sandbox system. My background perfectly aligns with the technical and collaborative aspects required for this role:

- Sandbox Hardening:** At Company XYZ, I led the security architecture overhaul of their container-based sandbox environment, successfully mitigating multiple known vulnerabilities. Specifically, I used Docker's seccomp profiles and Linux's cgroup policies to tighten sandbox configurations.
- R&D Collaboration:** In my role at ABC Corp, I served as the liaison between the R&D and Security teams, helping to incorporate security best practices into the development cycle without compromising agility.
- Security Testing & CI/CD:** I built a comprehensive security test suite at DEF Inc., which was fully integrated into the CI/CD pipeline. This suite caught 95% of security flaws before they hit production, a 40% improvement over the previous setup.
- Technical Proficiency:** I regularly use strace and similar tools for runtime analysis of containers. I'm also versed in Python, bash, and Linux—particularly Ubuntu and RHEL flavors.
- Communication Skills:** My experience collaborating in geographically dispersed teams makes me comfortable with asynchronous work. I've published security research papers and led workshops, showcasing my written and verbal English communication skills.

Considering the project's 2-4 week timeline, I am confident in delivering a robust, hardened sandbox environment that satisfies your security requirements. I'm familiar with Github, Slack, and am flexible with meeting times for PDT project leads.

I look forward to the opportunity to contribute to this crucial project and potentially engage in long-term collaboration.

Best regards,
[Your Name]



GRAPHISTRY



LOUIE.AI

info@graphistry.com

Active non-IT categories: Platform abuse & identity theft - \$ transfers

From Elizabeth Snow <myprivatemailoffice@comcast.net>

Reply Reply All Forward More 3/9/2017, 4:08 AM

Subject Immediate Response Needed
To jill@legit-company.com

Hi Jill,

Are you in the office? I'll need you to quickly set up an outgoing International wire for me.

Get back to me asap.

Criminals used AI to impersonate the CEO's voice, and it worked.

A UK-based energy firm was scammed out of \$243,000 when criminals targeted the company with an effective *vishing* campaign. "Vishing" is short for "voice phishing," the tactic of tricking targets over the phone. This incident marks the first time AI-based voice fraud has netted such a high payload, [according to The Next Web](#).

THE WALL STREET JOURNAL.

Police in the city of Baotou, in the region of Inner Mongolia, said the perpetrator used AI-powered face-swapping technology to impersonate a friend of the victim during a video call and receive a transfer of 4.3 million yuan (\$622,000).



GRAPHISTRY



LOUIE.AI

New

Few cases, but high \$ ROI

Ex: Deepfakes subset of "CEO invoice scams"

Dangerous

Social engineering attacks are scary

Ex: Popular trusted Okta login vendor breached through helpdesk

Getting bad, fast

Audio, visual, & text technology improving + democratizing

Scalable & automatable attack vector



info@graphistry.com

Need to protect against inauthentic activity

Increasingly hard to detect generated content, especially from determined attackers

Wreaking havoc on trust & safety teams: Reviews, comments, jobs, ...

LLM-level solutions

- AI watermarks: Dying off? (OpenAI, ...)
- Text analyzers: High FP (TurnItIn, ...)

More effective today: Check hard-to-fake data

- Identity authentication
- Social network analysis:
Group identity & individual behavior
- Digital forensics: Metadata mining
- Problem: Cyborgs
Ex: QAnon Into the Storm (HBO)
 - Behavior analytics & metadata worked!

LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security

Compliance

Market

CISO Priorities

The fight beyond IT

Bots & misinformation

Emergency services

Intelligence community

Controversial government
operations & policy

Military

Free speech

Authoritarian societies

Today + next few years

From technology first principles

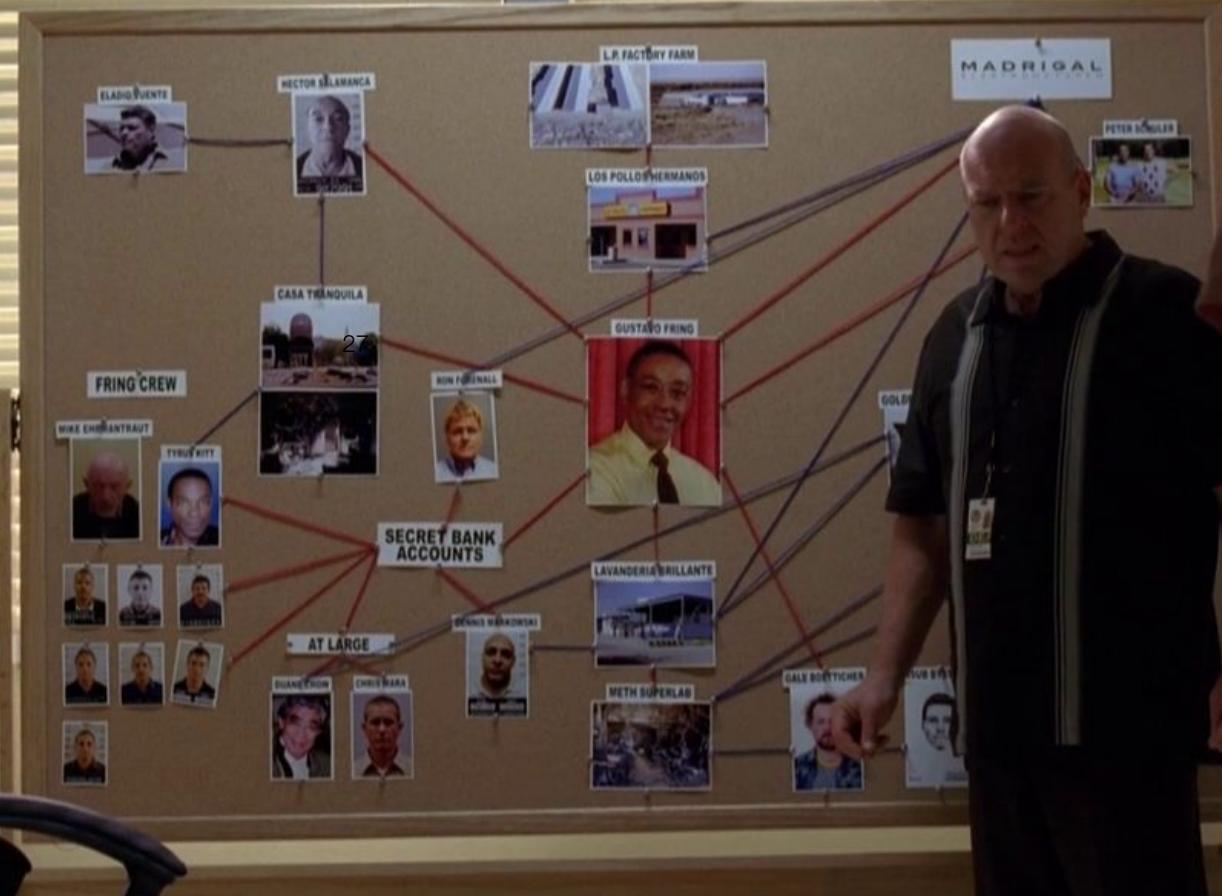


Real-world data + tech demos

Defenders perspective: Representative Tier 2 SOC session

26

```
IP=10.16.0.8; msg=Malware.Object;  
time=2 Nov 2017 19:32:00 UTC;  
vendor=FireEye; Product=Web MPS NX
```



Malicious Capabilities Observed in the VM

▶ Malicious Behavior: Yes

▶ OS Change Summary

Malware detected

Malware	Severity	Total	Infections	Callbacks	Blocked	Botnets	Last CnC Server	Last Location	First Seen	Last Seen	Ports Used	Protocols
Malware.archive		1	1	0	0	0			06/25/14 15:52:42	06/25/14 15:52:42	8080	TCP
Trojan.CVE2013		1	1	0	0	0			06/25/14 15:47:51	06/25/14 15:52:42	8080	TCP
Exploit.Browser		1	1	0	0	0			06/25/14 15:48:54	06/25/14 15:48:54		

IPS Details

Signature Name

Microsoft Internet Explorer VML Processing Integer Underflow (Generic Detection)

Infection URLs

Initial Infection URL	# Visits	Total URLs	First URL at	Last URL at
fdhtr423.co/docentx/audit/msie.php	1	2	06/25/14 15:47:48	06/25/14 15:47:49

Malware Binaries

Md5sum	Filetype	Protocol	Encoding	Last analysis time	# Occurrences
33e760716d82f04d0b7b25d832b79f1b	dll	TCP (8080)		06/25/14 15:52:42	1

▶ Acknowledge the infections and callbacks above for the host at 86.113.71.123.

86.113.71.123

IPS

|||||

3 3 0 0

Malware.archive

Malicious Capabilities Observed in the VM

▶ Malicious Behavior: Yes

▶ OS Change Summary

Malware detected

Malware	Severity	Total	Infections	Callbacks	Blocked	Botnets	Last CnC Server	Last Location	First Seen	Last Seen	Ports Used	Protocols
Malware.archive		1	1	0	0	0			06/25/14 15:52:42	06/25/14 15:52:42	8080	TCP
Trojan.CVE2013		1	1	0	0	0			06/25/14 15:47:51	06/25/14 15:52:42	8080	TCP
Exploit.Browser		1	1	0	0	0			06/25/14 15:48:54	06/25/14 15:48:54		

Infection URLs

Initial Infection URL	# Visits	Total URLs	First URL at	Last URL at
fdhtr423.co/docentx/audit/msie.php	1	2	06/25/14 15:47:48	06/25/14 15:47:49

Malware

Malware	Filetype	Protocol
33e760716d82f04d0b7b25d832b79f1b	dll	TCP (8080)

▶ Acknowledge the infections and callbacks above for the host at 86.113.71.123.



Virustotal is a **service that analyzes suspicious files and URLs** and facilitates the quick detection of viruses, worms, trojans, and all kinds of malware detected by antivirus engines. [More information...](#)

2 VT Community user(s) with a total of 4 reputation credit(s) say(s) this sample is goodware. 0 VT Community user(s) with a total of 0 reputation credit(s) say(s) this sample is malware.

File name: **cdbxp_setup_4.3.7.2356.exe**
Submission date: **2010-08-20 11:34:23 (UTC)**
Current status: **finished**
Result: **0 / 40 (0.0%)**

VT Community



goodware

Safety score: 100.0%

[Compact](#)

[Print results](#)

Antivirus	Version	Last Update	Result
AhnLab-V3	2010.08.20.01	2010.08.20	-
AntiVir	8.2.4.38	2010.08.20	-
Antiy-AVL	2.0.3.7	2010.08.16	-
Authentium	5.2.0.5	2010.08.20	-
Avast	4.8.1351.0	2010.08.20	-

File: 96.112.71.123

Malicious Capabilities Observed in the file

- Malicious Behavior: Yes
- OS Change Summary

Malware detected

Malware	Severity	Total	Infections	Confirms	Blocked	Infests	Last Cnic Server
Malware archive	Infected	1	1	0	0	0	
Dropbox	Infected	1	1	0	0	0	
English dictionary	Infected	1	1	0	0	0	
Signature	Infected	1	1	0	0	0	

Infection URLs

- Initial Infection URL: MaliciousFile.com/documents/testfile/malicious.php
- Infected: 234.70.71.123/dropbox/infected/234.70.71.123/testfile

ACKNOWLEDGE THE INFECTIONS AND CONFIRMATION PLEASE FOR THE HOST AT 96.112.71.123

File: 96.112.71.123

Malicious Capabilities Observed in the file

- Malicious Behavior: Yes
- OS Change Summary

Malware detected

Malware	Severity	Total	Infections	Confirms	Blocked	Infests	Last Cnic Server
Malware archive	Infected	1	1	0	0	0	
Dropbox	Infected	1	1	0	0	0	
English dictionary	Infected	1	1	0	0	0	
Signature	Infected	1	1	0	0	0	

Infection URLs

- Initial Infection URL: MaliciousFile.com/documents/testfile/malicious.php
- Infected: 234.70.71.123/dropbox/infected/234.70.71.123/testfile

ACKNOWLEDGE THE INFECTIONS AND CONFIRMATION PLEASE FOR THE HOST AT 96.112.71.123

Virustotal is a service that analyzes suspicious files and URLs and facilitates the quick detection of viruses, worms, trojans, and all kinds of malware detected by antivirus engines. [More information...](#)

2 VT Community user(s) with a total of 4 reputation credit(s) say(s) this sample is goodware. 0 VT Community user(s) with a total of 0 reputation credit(s) say(s) this sample is malware.

File name: **cdbxp_setup_4.3.7.2356.exe**
Submission date: **2010-08-20 11:34:23 (UTC)**
Current status: **finished**
Result: **0 / 40 (0.0%)**

VT Community

goodware
Safety score: 100.0%

[Compact](#) [Print results](#)

Antivirus	Version	Last Update	Result
AhnLab-V3	2010.08.20.01	2010.08.20	-
AntiVir	8.2.4.38	2010.08.20	-
Antiy-AVL	2.0.3.7	2010.08.16	-
Authentium	5.2.0.5	2010.08.20	-
Avast	4.8.1351.0	2010.08.20	-



Dashboard ACC Monitor Policies Objects Network Device

Save Help

Palo Alto Networks Firewall Rules

Name	Source			Destination			Application	URL Category	Service	Action	Profile
	Zone	Address	User	Zone	Address	any					
LogAll	Tap	any	any	Tap	any	any	CustomerURLCategory	any	any	Allow	Custom-app, SSL, Web-browsing, Application-Default
IT Allow Override	trust	any	pancademo\administrators	untrust	any	any	Custom-app	any	any	Allow	Custom-app, SSL, Application-Default
Read Only Facebook	trust	any	pancademo\administrators	untrust	any	any	facebook-base	any	any	Allow	SSL, Application-Default
Allow facebook posting	trust	any							any	Allow	Application-Default
Block Peer to Peer	trust	any							any	Block	None
Webmail file blocking	trust	any							any	Allow	Application-Default
Sharepoint	Untrust-L3	any							any	Allow	Application-Default
Allow SSL and SSH	trust	any							any	Allow	Application-Default
Allow Web-browsing	trust	Sharepoint S							any	Allow	Application-Default
Block encrypted tunnel	trust	any							any	Block	None
Block Proxies and Anonymizers	trust	any							any	Block	None
Mail server	Untrust-L3	any							any	Allow	Application-Default
Web server	Untrust-L3	any							any	Allow	Application-Default

Action Buttons: Add, Delete, Clone, Enable, Disable, Move Top, Move Up, Move Down, Move Bottom, Highlight Unused Rules

Total Rules: 13 rule(s)

FireEye Malware Detection

Malware: 0/1

Malware Active: 0/1

Threats Detected: 0/1

IPS: 0/1

Signature: 0/1

Malware Detectors: 0/1

Malware: 0/1

Infection URLs: 0/1

Initial Infection URLs: 0/1

Malware URLs: 0/1

Malware: 0/1

File Name: c:\drop_temp_4.3.2.7255.exe

Submission Date: 2010-09-28 11:42:33 (ETC)

Current Status: Failed

Result: 0/0 (0.0%)

Antivirus Results:

Antivirus	Version	Last Update	Detail
Absid-AV	2010.09.20.01	2010.09.20	-
Acitivis	8.2, 4.38	2010.09.20	-
Acitivis-AVL	2.0.3, 4.7	2010.09.16	-
Authentium	5.2.0, 5	2010.09.20	-
VirusShare	4.6, 2011.0	2010.09.20	-

VirusTotal Analysis

2/57 Community users(s) with a total of 4 reputation credit(s) say(s) this sample is goodware. 0 VT Community user(s) with a total of 0 reputation credit(s) say(s) this sample is malware.

File name: c:\drop_temp_4.3.2.7255.exe

Submission date: 2010-09-28 11:42:33 (ETC)

Current status: Failed

Result: 0/0 (0.0%)

Antivirus Results:

Antivirus	Version	Last Update	Detail
Absid-AV	2010.09.20.01	2010.09.20	-
Acitivis	8.2, 4.38	2010.09.20	-
Acitivis-AVL	2.0.3, 4.7	2010.09.16	-
Authentium	5.2.0, 5	2010.09.20	-
VirusShare	4.6, 2011.0	2010.09.20	-

Palo Alto Networks Firewall Rules

Name	Zone	Address	User	Zone	Address	any	Application	URL Category	Service	Action	Profile
LogAll	Tap	any	any	Tap	any	any	CustomerURLCategory	any	any	Allow	Custom-app, SSL, Web-browsing, Application-Default
IT Allow Override	trust	any	pancademo\administrators	untrust	any	any	Custom-app	any	any	Allow	Custom-app, SSL, Application-Default
Read Only Facebook	trust	any	pancademo\administrators	untrust	any	any	facebook-base	any	any	Allow	SSL, Application-Default
Allow facebook posting	trust	any							any	Allow	Application-Default
Block Peer to Peer	trust	any							any	Block	None
Webmail file blocking	trust	any							any	Allow	Application-Default
Sharepoint	Untrust-L3	any							any	Allow	Application-Default
Allow SSL and SSH	trust	any							any	Allow	Application-Default
Allow Web-browsing	trust	Sharepoint S							any	Allow	Application-Default
Block encrypted tunnel	trust	any							any	Block	None
Block Proxies and Anonymizers	trust	any							any	Block	None
Mail server	Untrust-L3	any							any	Allow	Application-Default
Web server	Untrust-L3	any							any	Allow	Application-Default

Action Buttons: Add, Delete, Clone, Enable, Disable, Move Top, Move Up, Move Down, Move Bottom, Highlight Unused Rules





Acquisitions

Search by hostname, IP acquisition

Show: All acquisitions Status: All statuses Requested: Everyone

Acquisitions (236) Acquisition space: 53.2% full - 12.6GB remaining

Actions	Hostname	IP Address	Requested	Acquisition	Download Size	Status
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	58 minutes ago	Triage (automatic)	4.7MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	2 hours ago	Triage (automatic)	4.7MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	4 hours ago	Triage (automatic)	4.7MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	5 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	7 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	9 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	10 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	11 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	12 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	14 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	15 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	17 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	18 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	20 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	21 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	22 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	24 hours ago	Triage (automatic)	4.4MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	25 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	27 hours ago	Triage (automatic)	4.5MB	Acquired

Acquisition Detail

Triage collection acquired - Triage summary - Download full triage

ACQUISITION DETAILS

- Request Type: Around timestamp
- Timestamp: 2015-10-06 17:15:30Z
- Requested: 2015-10-06 17:17:04Z
- Requested By: Automatic



Malware Configuration Overview in the UI

Malware Refresher: 0ms

EVN Change Summary

Malware Detected

Malware Active

Threats Detected

Signature Detection

Signature Interval: Exploits URLs Processing Underline Underline (Current: Detecting)

Infection URLs

Initial Infection URLs

File Name: clobdrop_4.3.7.255.exe

Submission Date: 2015-10-06 17:15:23 (EST)

Current Status: Failed

Result: 0/9 (0.0%)

Comments

Antivirus Version Last Update Result

Absurd-V3 2010-08-20, 20.01 2010-08-20 -

AcTViTe 8.2, 4.38 2010-08-20 -

AcTViTe-AVL 2.0, 5.7 2010-09-16 -

AuthenticA 5.2, 0.5 2010-09-20 -

VirusBuster 4.6, 2.01, 0 2010-06-20 -

Acknowledge the infections and calculate please for the host at 192.168.1.123



VIRUS TOTAL

VirusTotal is a service that analyzes suspicious files and URLs and facilitates the quick detection of viruses, worms, trojans, and all kinds of malware detected by antivirus engines. [More information...]

2 VT Community user(s) with a total of 4 reputation credit(s) say(s) this sample is goodware. 0 VT Community user(s) with a total of 0 reputation credit(s) say(s) this sample is malware.

File name: clobdrop_4.3.7.255.exe

Submission date: 2015-10-06 17:15:23 (EST)

Current status: Failed

Result: 0/9 (0.0%)

Comments

Antivirus Version Last Update Result

Absurd-V3 2010-08-20, 20.01 2010-08-20 -

AcTViTe 8.2, 4.38 2010-08-20 -

AcTViTe-AVL 2.0, 5.7 2010-09-16 -

AuthenticA 5.2, 0.5 2010-09-20 -

VirusBuster 4.6, 2.01, 0 2010-06-20 -



Palo Alto Networks

Dashboard | ACC | Monitor | Policies | Objects | Network | Device

VT Community

File: clobdrop_4.3.7.255.exe

Last Seen: 2015-10-06 17:15:23 (EST)

Result: Failed

Comments

Antivirus Version Last Update Result

Absurd-V3 2010-08-20, 20.01 2010-08-20 -

AcTViTe 8.2, 4.38 2010-08-20 -

AcTViTe-AVL 2.0, 5.7 2010-09-16 -

AuthenticA 5.2, 0.5 2010-09-20 -

VirusBuster 4.6, 2.01, 0 2010-06-20 -



Acquisitions

Search: Acquisitions | Filter: All status | Request: Everyone

Acquisitions (236) Acquisition space: 53.2% full - 12.6GB remaining

Actions	Hostname	IP Address	Requested	Acquisition	Download Size	Status
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	58 minutes ago	Triage (automatic)	4.7MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	2 hours ago	Triage (automatic)	4.7MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	4 hours ago	Triage (automatic)	4.7MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	5 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	7 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	9 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	10 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	11 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	12 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	14 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	15 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	17 hours ago	Triage (automatic)	4.5MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	18 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	20 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	21 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	22 hours ago	Triage (automatic)	4.3MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	24 hours ago	Triage (automatic)	4.4MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	25 hours ago	Triage (automatic)	4.6MB	Acquired
<input type="checkbox"/>	fe-xpsp1-victim	192.168.6.247	27 hours ago	Triage (automatic)	4.5MB	Acquired

Acquisition Detail

Triage collection acquired - Triage summary - Download full triage

ACQUISITION DETAILS

- Request Type: Around timestamp
- Timestamp: 2015-10-06 17:15:30Z
- Requested: 2015-10-06 17:17:04Z
- Requested By: Automatic

CylancePROTECT | Device 1 | x

https://my.cylance.com/Device/DisplayDeviceDetails

The screenshot shows the CylancePROTECT interface for a specific device. At the top, there are tabs for DASHBOARD, PROTECTION, ZONES, DEVICES, and SETTINGS. Below the tabs, there are buttons for Export, Quarantine, and Waive. A large green circular icon with a white 'X' is prominently displayed in the center of the page. To its right, the word "CYLANCE" is written in a large, bold, sans-serif font. Below the central icon, there is a table titled "Status: Quarantined (13)". The table has columns for FILE PATHS, STATUS, FIRST FOUND, RUNNING, AUTO RUI, CLASSIFICATION, and DETECTED. The data in the table includes file paths such as c:\users\admin1\downloads\29ed955f046ef9db1e0eba7da47a29f950a3e, file names like 4f39d3e71ed1278d5fa83edf148, and various dates like 5/19/2016. Most entries are marked as "Quarantined".

FILE PATHS	STATUS	FIRST FOUND	RUNNING	AUTO RUI	CLASSIFICATION	DETECTED
29b4498ac81d54b52cd0a32bd29ed955f046ef9db1e0eba7da47a29f950a3e	Quarantined	5/19/2016	No	No	File Watcher	
4f39d3e71ed1278d5fa83edf148	Quarantined	5/19/2016	No	No	File Watcher	
70cc1e06fb896b627dd4c5337b8f4cf65a6783745368bb95285448aa0a9b	Quarantined	5/19/2016	No	No	File Watcher	
72513534f2e0f3e72a22023b887df3718c9df70686b0ae50cbde2f90f447e4	Quarantined	5/19/2016	No	No	File Watcher	
766e49811c0bb7ce217e72e73a6aa66c15de0ba11d7dd3bd7e9ec33ed6963	Quarantined	5/19/2016	No	No	File Watcher	

Grouped By: ▲ Status X

Export Quarantine Waive

50 Items per page

1 - 13 of 13 items

This screenshot shows the FireEye Malware Coordinator interface. It features a large red eye icon with a flame, followed by the "FireEye™" logo. Below the logo, there's a summary table with columns for Malware Detected, Malware Active, Threats Detected, and Infection URLs. The table shows a total of 1 infection, 1 active malware, and 1 threat detected. There are also sections for VirusDB Reference and OS Change Summary.

This screenshot shows the VirusTotal analysis results for a sample. The interface includes the "VIRUS TOTAL" logo and a summary table. The table shows the sample is goodware with a 100% detection rate across 21 engines. Below the table, there's a detailed breakdown of detections from engines like ClamAV, McAfee, and Sophos.

File name	Result	Engine
c0d0e452-43-2756.exe	goodware	ClamAV
Result	100/60 (100%)	McAfee

This screenshot shows the Palo Alto Networks dashboard. It includes a "VT Community" section with a "goodware" result and a 100% detection rate. Below it, there's a "Network" section with a table of network traffic. The table lists various IP addresses, ports, and protocols. The Palo Alto Networks logo is visible at the bottom.

This screenshot shows the CylancePROTECT interface for acquisitions and audit logs. It includes a table of acquisitions with columns for ID, Acq Date, Acq Type, and Status. There are also sections for "Acquisition Detail" and "Audit Log".

This screenshot shows the CylancePROTECT interface for a specific device. It includes a table of acquisitions with columns for ID, Acq Date, Acq Type, and Status. There are also sections for "Acquisition Detail" and "Audit Log".

A screenshot of a Windows desktop environment displaying several windows related to system monitoring and security:

- Event Viewer (Local):** Shows the "Security" log with 6,927 events. A red arrow points to the "Security" node under "Windows Logs". Another red arrow points to the "Save All Events As..." option in the context menu of the right pane.
- Windows Firewall with Advanced Security:** Shows the "Connections" pane with various rules and statistics.
- FireEye Malware CAPTURED in the VM:** A separate window showing infection details.
- Windows Taskbar:** Displays the "Start" button, taskbar icons, and the system tray.
- CYLANCE Endpoint Protection:** A separate window showing real-time protection status and threat details.



GRAPHISTRY



LOUIE.AI

info@graphistry.com

New Search

Save As ▾ Close



```
source=wineventlog EventCode=4624 (Logon_Type=10 OR Logon_Type=7 OR
Logon_Type=2) host=$Hostname3$
| rex "New Logon:\$+Security ID:$+(?<account>.*)"
| eval Type=case(Logon_Type=10,"Remote Logon",
Logon_Type=2,"Local Logon",
Logon_Type=7,"Screen Unlock")
| table _time host Type account
| sort _time desc</query>
```

	_time	pid	source	status	time	hostname	host	splunk_server_group
>	5/7/17 6:41:30.000 PM		/var/log/nginx/access.log	200			staging-docker0-0	dmc_group_cluster dmc_group_deploy dmc_group_index dmc_group_license dmc_group_search
>	5/7/17 6:41:30.000 PM		/var/log/clients/clients.log		2017-05-08T01:41:29.953Z	staging.graphistry.com	staging-docker0-0	dmc_group_cluster dmc_group_deploy dmc_group_index dmc_group_license dmc_group_search
>	5/7/17 6:41:30.000 PM	1	/var/log/pivot-app/pivot.log		2017-05-08T01:41:30.401Z	c8a79d3131e1	staging-docker0-0	dmc_group_cluster dmc_group_deploy dmc_group_index dmc_group_license

Selected Fields

- a host 3
- a hostname 3
- a metadata.userInfo.cid 9
- a metadata.userInfo.tag 1
- # pid 25
- a source 24
- a splunk_server_group 5
- # status 1
- a time 42

Windows

SPLUNK

YLANCE





Malicious Capabilities Observed in the VM

- Malicious Behavior: Tmn
- Oil Change Necessary

Malware detected

Holware

Total: 1

Initial Infection URLs: [http://www.115.71.123.100/](#)

Visits: 1

Total URLs: 2

Medium

File type: TCP (60K)

Acknowledge the infections and clickback above for the host at 115.71.123.

2/11/2016

2/11/2016

2/11/2016

Initial Infection URLs: [http://www.115.71.123.100/](#)

Visits: 1

Total URLs: 2

Medium

File type: TCP (60K)

Acknowledge the infections and clickback above for the host at 115.71.123.

2/11/2016

2/11/2016

2/11/2016

Initial Infection URLs: [http://www.115.71.123.100/](#)

Visits: 1

Total URLs: 2

Medium

File type: TCP (60K)

Acknowledge the infections and clickback above for the host at 115.71.123.

2/11/2016

2/11/2016

2/11/2016

VT Community Sign in Languages ▾



Virustotal is a [service that analyzes suspicious files and URLs](#) and facilitates the quick detection of viruses, worms, trojans, and all kinds of malware detected by antivirus engines. [More information...](#)

2 VT Community user(s) with a total of 4 reputation credit(s) say(s) this sample is goodware. 0 VT Community user(s) with a total of 0 reputation credit(s) say(s) this sample is malware.

File name: **cdblpx_setup_4.3.7.2356.exe**
 Submission date: **2010-08-20 11:34:23 (UTC)**
 Current status: **finished**
 Result: **0/40 (0%)**

[Compact](#)**Antivirus**

Antivirus	Version	Last Update	Result
AhnLab-V3	2010.08.20.01	2010.08.20	-

AntiVir

AntiVir	8.2.4.38	2010.08.20	-
---------	----------	------------	---

Antiy-AVL

Antiy-AVL	2.0.3.7	2010.08.16	-
-----------	---------	------------	---

Authentium

Authentium	5.2.0.5	2010.08.20	-
------------	---------	------------	---

Avast...

Avast...	4.8-1351.0	2010.08.20	-
----------	------------	------------	---

VT Community

Safety score: 100.0%

[Print results](#)

CYLANCE



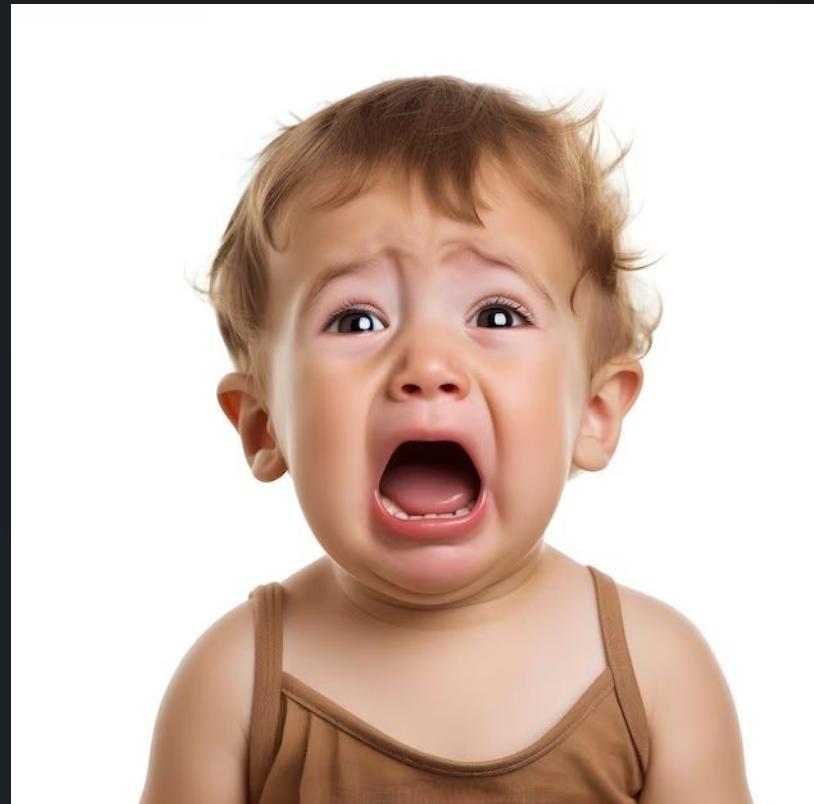
... start over!



Slight (big) problem

Sumo Logic is a scalable and secure platform that processes and analyzes petabytes of data, runs over 30 million searches and scans over 500 trillion records daily.

With problems like alert fatigue, stress, and low levels of automation, most analysts burn out in their first year, and constant demand to “do more with less”



Defenders perspective: \$200B of defense in depth - surface keeps growing!

Shift-left

- Vendor review
- Security design review
- Threat modeling
- Code analysis

Production surface

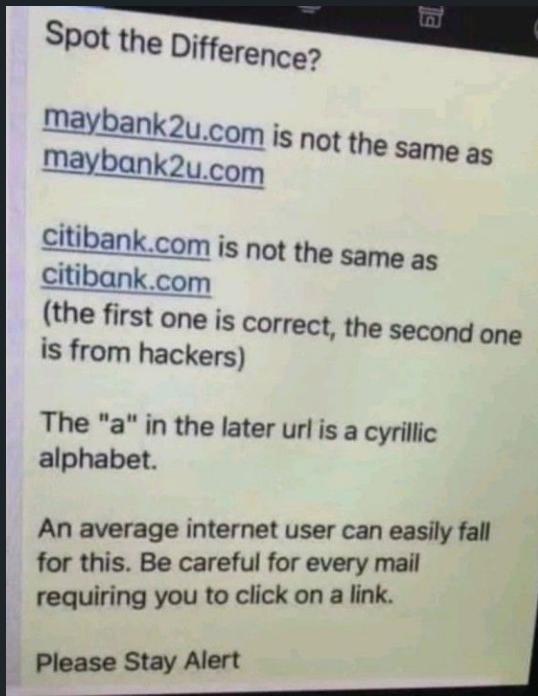
- Firewalls, load balancers, email, DNS
- Identity, access management, zero-trust
- Endpoint, network, cloud, shadow IT, SaaS
- Public site & intranet
- Applications and their behavior
- Compliance & audits

Security operations

- Help desk
- Digital forensics & incident response
^^^^^ the investigation example
- Patch management
- Threat hunting & threat intel
- Detection engineering
- Security awareness programs

**Generative AI projects are rethinking
every area listed here**

Blackbox tool vendors using LLMs for smarter detections. Ex: Phishing



URLs: DGA, lookalikes, ...

Phishing

- Fake websites
- False invoices
- Fake data requests: Legal, ...
- Fake text messages

Many vendors doing this

LLM *classifier*, but not necessarily *generative* (GPT)

SDLC & compliance examples

This screenshot shows a Google Sheets document titled "Big Company Security Questionnaire_Completed". The document contains several sections of questions and their answers. One section is titled "Questions" and includes questions like "Is there a formal risk management program that has been approved by management, communicated to constituents and an auditor?" and "Are there processes for identifying, assessing, and mitigating risks identified during assessments?". Another section is titled "Comments" and provides detailed feedback for each question. The document is filled with text, tables, and some redacted areas.

Security questionnaires
many, ex: vanta

This screenshot shows a feedback interface for a "Data Destruction" task. It has a "HIGH" priority level. The main message asks, "Do you have a destruction mechanism for the email addresses column?". Below it, a "Description" section explains that the table EU_USERS will duplicate existing PII (email addresses), requiring its own specific mechanism for data deletion to ensure compliance with Acme's data retention policy. At the bottom, there are links to "See more", "Regenerate with feedback", and "Ignore".

Design review
ex: remysec



Code scans
many (see DARPA challenge)

Observation: Bots enable pushing smarter security discussions to end-users so they can help themselves and more autonomous workflows overall. Important for enterprise scale and SMB affordability & practicality.

REVERSE
Github Copilot generating ~half of new Github code,
and question of code bugs & vulns



Top security analytics & automation platforms adding generative AI: SIEM, XDR, SOAR, DFIR, ...

Siloed platforms

CrowdStrike's Charlotte AI: Headline: "Introducing Charlotte AI, CrowdStrike's Generative AI Security Analyst: Ushering in the Future of AI-Powered Cybersecurity". Subtext: "May 30, 2023 Michael Sentonas Endpoint & Cloud Security". Below: "Announcing Microsoft Security Copilot".

Microsoft Security Copilot: Headline: "Introducing Charlotte AI, CrowdStrike's Generative AI Security Analyst: Ushering in the Future of AI-Powered Cybersecurity". Subtext: "May 30, 2023 Michael Sentonas Endpoint & Cloud Security". Below: "Announcing Microsoft Security Copilot".

Google Launches Security AI Workbench: Headline: "Google Launches Security AI Workbench to Rival Microsoft's Security Copilot". Description: "Vertice AI on Google Cloud Platform".

Security AI Workbench: A circular diagram titled "Security AI Workbench" with "Dex PaLM" at the center. It shows various components: "Unstructured search, alerts, predictions", "Machine Threat Intel", "Virtual Patch", "GSS Vulnerabilities", "Security GitHub repos", "MITRE Frameworks", "Selected partner feeds", "Vertice AI on Google Cloud Platform", "LLM research & development", "Computer vision", "Global threat intel", "Identity management", "Native Security Command Center", "Threat Intel Chronicle", "Partners Extend and customize", and "Native Security Command Center, Threat Intel Chronicle".

Cross-silo platforms



louie.ai



SOAR 2.0

Data co-pilot; smart automation; autonomous investigation; threat intel handling; ...

Detection engineering

Easier, cover text columns, add smarter decision making; questions of scale, fidelity, model

Upskilling: Junior analysts, coding, ...



GRAPHISTRY



LOUIE.AI

info@graphistry.com

Drilldown: Smarter analyzers

Experiment - autonomous hunting by Graphistry / Louie.AI: US Gov challenge data

-  GPT3-based autonomous strategy: ~0 findings within 10 queries
-  Automatic graph AI anomaly detector: ~80%
-  GPT4-based autonomous strategy: ~all hits in first 5 queries

Intuition: Similar to GPT passing the bar exam, GPT4 knows data source <> hunt strategies

[Demo:

<https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5>]

Frontier: Neurosymbolic synthesis

- Combining Z3 SMT solver and/or code execution with GPT4 is getting SOTA on various benchmarks
- ⇒ smarter dynamic code analysis ?
- ⇒ crypto contract verification ?

[Demo: Louie <> Z3/PySym/HF?]

Demo: Generating an OPA ABAC policy & verifying with Z3.. in 2 minutes

LE

write a reasonable OPA policy using ABAC to govern who has access to admin panels, database connectors, & dashboards

Sure, here's a basic example of an OPA policy using ABAC:

```
```rego
package abac

default allow = false

allow {
 input.subject.attributes.role == "admin"
 input.action == "access"
 input.resource.type == "admin_panel"
}

2023-10-03 01:26:44 in 22 seconds
2023-10-03 01:26:44 [TraceLine] Running off-thread
2023-10-03 01:26:44 [TraceLine] Thinking about input: write a reasonable OPA
2023-10-03 01:27:06 [TraceLine] Chain End: {'text': 'Thought: The user is asl
```

+

LE

Think of a useful security property to verify about this policy, and generate & run Z3 code that translates the OPA policy + question to Z3 and runs them, and confirm whether or not the property is true

The Z3 solver confirmed that a user with a role other than "admin", "database\_connector", or "dashboard\_user" cannot access the respective resources. This means our OPA policy is correctly restricting access based on user roles.

```
Importing the required library
from z3 import *

Define the roles and resources
Role, (admin, database_connector, dashboard_user, other) = \
EnumSort('Role', ['admin', 'database_connector', 'dashboard_user', 'other'])
Resource, (admin_panel, database_connector_resource, dashboard, other_resource) = \
EnumSort('Resource', ['admin_panel', 'database_connector_resource', 'dashboard', 'other_resource'])

Define the policy
def policy(role, resource):
 ...

2023-10-03 01:28:23 in 1 minutes 1 seconds
2023-10-03 01:28:23 [TraceLine] Running off-thread
2023-10-03 01:28:23 [TraceLine] Thinking about input: Think of a useful secu
2023-10-03 01:28:47 [TraceLine] Chain End: {'text': 'Thought: The user wants
2023-10-03 01:29:09 [TraceLine] Chain End: {'text': '# Importing the require
2023-10-03 01:29:09 [TraceLine] # Importing the required library
from z3 import *
```

In 2 short prompts

Autogenerate a reasonable OPA ABAC policy

Autogenerate verification goals for the security policy

Converts to Python calling the Z3 theorem prover

Auto-runs, verifying & reporting whether the security policy obeys the verification goals

<https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5>



# LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security

Compliance

Market

CISO Priorities

The fight beyond IT

Bots & misinformation

Emergency services

Intelligence community

Controversial government  
operations & policy

Military

Free speech

Authoritarian societies

Today + next few years

From technology first principles



Real-world data + tech demos

# Regulatory environment is a mess

## Safe result

### GDPR

Sovereignty + deletion = 🎉 + \$\$\$

### Copyright

Lawsuits: ChatGPT copying Sarah Silverman & reporters, while Adobe curates safe data

### Algorithmic discrimination

Regulated technologies (ex: loan approvals) need to be free of bias

## Safe process

### Explainability (XAI)

Address risk management, consumer protection, etc. requirements

### Determinism

Imagine using in a medical context and getting different answers!

### Chain of custody

Similar to explainability

- Rules change by country, company & month
- People use ChatGPT even when prohibited

# Topic for another day: Attacks against machine learning

Market sizes: Proxy for how leaders value mlsec

AI: Growing from \$100B+ to \$1T+ in 2030

Cybersecurity: \$200B going to \$500B in 2030

**Attacks against machine learning:**  
< 1% of cybersecurity market, and low ceiling?

Ex: Louie.AI focuses mostly on securing data, access control, & sandboxes.

LLM solves alignment, and most ML attacks outside our threat model

Challenging

**Alignment:** No one wants a racist clippy - users & businesses want social norms built in

**CIA - Confidentiality, Integrity, Availability:**  
Data poisoning, model inversion, ...

**Federated access control for federated data**

Ex: When indexing Slack channels, and membership changes, ensuring RAG still right

**Tool sandboxing**

AutoGPT: Securing Python, DB, ...

Further reading: See new guides by OWASP (light) and NIST (heavy)

# CISO Priorities in practice

## #1: Launch private LLMs - stop the bleeding

Internal ChatGPT UI, LLM API, & top genAI apps

First get data safe inside perimeter; AI guard rails laters

## #2: Set & enforce policy

Identify policies employees will follow + top vendors

## #3: Learn

Use private GPT + find responsive innovation partners enabling AI use cases (platform + solution):  
Fixie, Palantir, startups... including us!

## #4: Low-hanging fruit

Productivity platform tools  
Blackbox vendor tools

## #5: Strategic initiatives

Ex: Database <> LLM enablement  
Ex: Detections on traditionally hard data

Priority on indexing & using proprietary data for growing LLM use cases

# LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security

Compliance

Market

CISO Priorities

The fight beyond IT

Bots & misinformation

Emergency services

Intelligence community

Controversial government  
operations & policy

Military

Free speech

Authoritarian societies

Today + next few years

From technology first principles



Real-world data + tech demos

# Emergency services

Challenge: Continuous monitoring of social, news, & open media for threats & status

- Detect within 1 minute of posting:
- Hurricane damage, evacuation status, physical threat, ...
- Decades of after action reports

⇒ LLM + vector DB + KG: Scalable & higher quality

Challenge: Misinformation

⇒ LLM meets graph: Label topics from misinfo networks

Challenge: Automated decision support

Fusion analysis, interactive analytics, alerting, briefs, ...

⇒ LLM agents with many interfaces



Screenshot: Hurricane monitoring



# Intelligence community

Intelligence gathering, analysis, & operations

- Detect specific threats: economic, democratic, physical, ...
- Policy guidance: diplomatic & military
- Resource allocation clarity

Physical event → ... → President's daily brief

Raw analysis: sigint, osint, imint, ...

Synthesis & higher-level reasoning

Decision science: Alerting, personalization, ...

Need to rethink \$100B/yr of US IC spending

OSINT Commoditization ⇒ Press, NGOs, threat intel sharing

## Biden puts the 'daily' back into the administration's intelligence briefings

While his predecessor received the in-person daily intelligence briefing sporadically, Biden plans to have it on his schedule most days, administration officials say.



## CIA Builds Its Own Artificial Intelligence Tool in Rivalry With China

- Intelligence agency looking for 'needles in the needle field'
- Program will help sift through huge amounts of available data



info@graphistry.com

# LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security

Compliance

Market

CISO Priorities

The fight beyond IT

Bots & misinformation

Emergency services

Intelligence community

Controversial government  
operations & policy

Military

Free speech

Authoritarian societies

Today + next few years

From technology first principles



Real-world data + tech demos

# Military

Similar to intelligence community...

Focus on force, logistics, & combat

Real-time

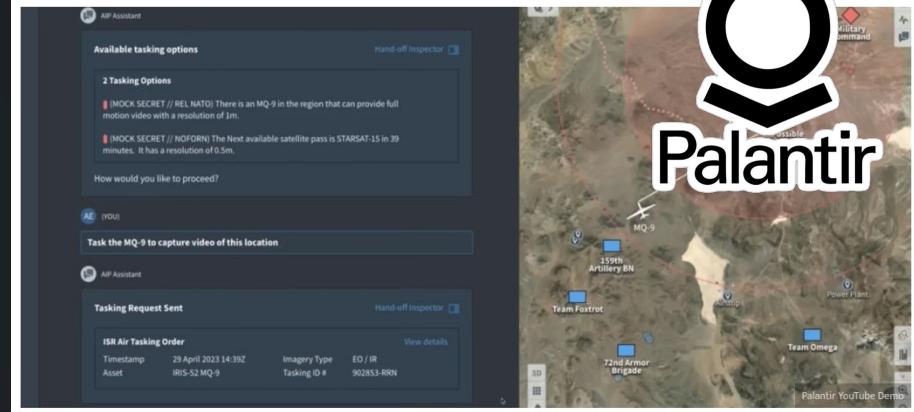
Coordination & multi-lingual allies

Situational awareness

Training

Controversy: #autonomousweapons

## The application of generative AI to warfare raises human rights concerns



### Take Action

Support the pledge against autonomous weapons and help ensure humans maintain control over the decision to use lethal force



info@graphistry.com



# Attacks against free speech societies: Election security, foreign policy, ...



Ex: President Zelensky of Ukraine faked in belly dancing video via Russian misinfo campaign

Responsibility?  
Midjourney paused free tier ... AFTER becoming successful

## Challenges

- Vocal minority lies & artificially inflates
- Social platforms incentivized to amplify: clickbait (\$\$\$\$), engagement, free speech
- Collaborating misinformation networks collaborate too: Ex - 5G, Russian propaganda, homeopathy
- GPT throwing fuel on the fire

(See also: my keynote at DefCon AI on analyzing collaborating covid medical misinfo communities)



# Challenge: 1/3rd world authoritarian; controlled by tech suppressing free speech

**Today:** Authoritarian governments already use tech to

- Fabricate false reality
- Suppress violations
- Surveil for violators

**Tools:** State media, suppressed press, internet firewall, social & news censor lists

**Ex: Russians** believe Ukraine is not a war, but a special operation to fight Nazis; **TikTok** hides

**Ex: Chinese** deny US claims about Uyghur camps: genocide, forced labor, forced sterilization, ...

**What are our responsibilities?**

- LLMs accelerate, scale, & commoditize the capability
- Tech community: R&D, OSS, startups, & megacorps
- Government: Exports, visas, sanctions, ...
- Citizens, new hires, ...

The screenshot shows a news article from 'WAR IN UKRAINE'. The title is 'TikTok created an alternate universe just for Russia'. Below the title, there is a short text snippet: 'The Chinese-owned social media giant weathered Putin's information crackdown by muzzling its users there and cutting them off from the outside world, while allowing state propaganda'. There is a profile picture of the author, Will Oremus, and the date of publication: April 13, 2022 at 8:00 a.m. EDT.

The screenshot shows a news article from 'The New York Times' under the 'War' section. The title is 'Ukrainians Find That Relatives in Russia Don't Believe It's a War'. The text discusses how many Ukrainians are encountering backlash from family members in Russia who have bought into the official Kremlin messaging. At the bottom of the article, there are sharing options: 'Share full article', 'Email', 'Bookmark', and '1K'.

# LLMs in security

Escalating cyber attacks

New era of cyber defense

Business of security

Compliance

Market

CISO Priorities

The fight beyond IT

Bots & misinformation

Emergency services

Intelligence community

Controversial government  
operations & policy

Military

Free speech

Authoritarian societies



Real-world data + tech demos

Today + next few years

From technology first principles

# Trajectory - Where we are today

Earlier: MS/PhD-level pipelines with low scores for HMM, RNN, ... => limited use (Google, ...)

2018: BERT - Basic NLP became easy via pretraining & quality => CyBERT, HuggingFace, ....

2022: GPT3 - Bigger generative models that fail typical exams

2023: ChatGPT + GPT3.5 + GPT4 - Generative models that pass typical exams + UI

Ecosystem & applications limited by GPT3/4 model limitations:

- Use cases: Any small data, slow data, and human speed (< 10 tok/sec)
- Model distillation to shrink models so it can run faster & cheaper with good quality
- Program & plan generation: Small snippets needing manual cleanup (Github CoPilot, SQL)
- Infra catchup: OSS, DB indexing, hooking up tools, ... - early, see rise of vector DBs / RAG

# Trajectory - Where we're going in 2024, 2025: Coming technology waves

Current-generation tech - Application builders still catching up!

tok/s will increase - operational data

Unlocks bigger & faster data sources for tasks like alert & event monitoring

Data indexing - BYOData

KG integration, data catalog adoption, longer contexts ⇒ personal & org assistants, tools on top

Beginning of autoGPT and automation

Smarter planning & coding ⇒ autonomous hunting, ...

AutoAI

Agents learning how to use computational thinking: Python, solvers, analytics/ML/AI, DBs, KGs, ...

# Thank you!

HIRING

Social & technical marketing manager,  
Visualization engineer (WebGL/WebGPU)



## CYBERSECURITY

iPhone moment: Every area in security getting revisited

Attackers big on platform abuse; deep fakes coming; need stronger social/trust/identity solutions

New defender tools: Start simple - blackbox detections (ex: phishing) + co-pilots (investigation, ...)

2024+ as tech improves: Autonomous (autoGPT) mode + economical line-rate monitoring

## BROADER OPPORTUNITIES & CHALLENGES

Also reshaping: Emergency management, IC community, ...

Authoritarian govs: 1/3rd of population having tech replace their free speech with misinfo & surveillance

Democratic govs: Under attack from inauthentic information that gets digitally amplified