



# 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

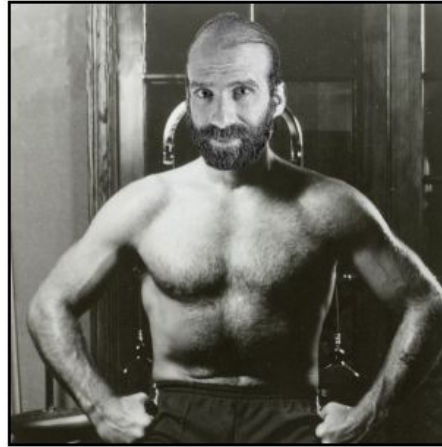
- "Applied Cryptography" (1993), "Cryptography Engineering" (2010)
- Invented: Blowfish, twofish, ...
- Startups: Counterpane, Resilient, ...
- Advisory: EFF, ...

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 Smidgy

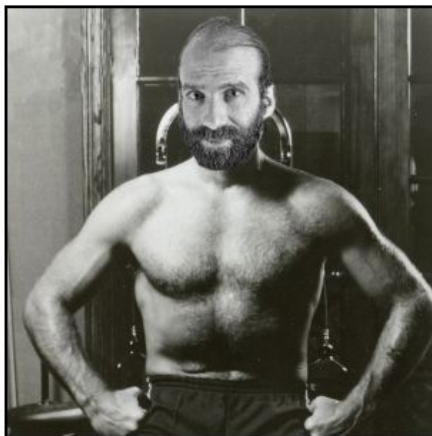
[permalink](#) | [vote](#)

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.

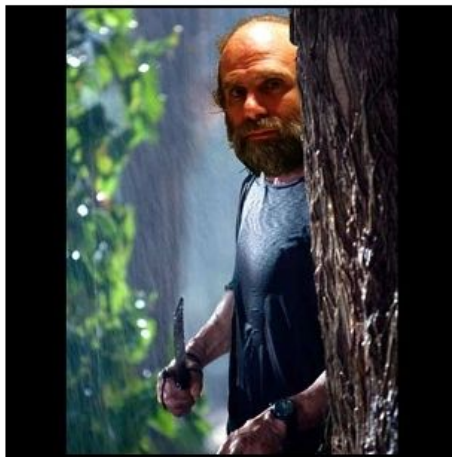
Society would be safer  
if everyone had a SchneierGPT  
looking over their shoulder



# 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

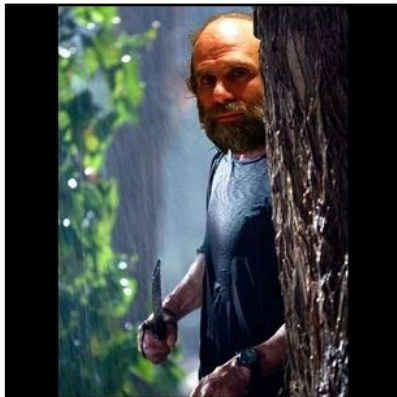
<https://www.loom.com/spaces/All-Graphistry-1228153/folders/berkeley-194-d905f3da2df440c2b5e9e1c6824d6ab5>



# 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



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

US Cyber Command  
AI RPE winner 2022



Margrave:  
XACML Verification and  
Change-Impact Analysis

Kathi Fisler, WPI  
Shriram Krishnamurthi, Brown  
Leo Meyerovich, Brown  
Michael Carl Tschantz, Brown

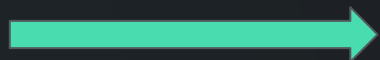


AWS  
Cedar

Explaining & verifying rich  
security policies

# FLAPJACK

Test of time award

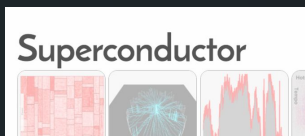


Functional reactive  
web programming (FRP)

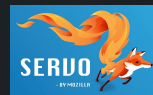
Berkeley parallel  
browser

+

GPU DF.js



Information is Beautiful award



Multicore/GPU web



GPU dataframes

# 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 marketplace interface with a navigation bar at the top containing links for Home, Shop, Tools And Tutorials, Testimony, Downloads, How to pay and get the files, and About Us. There are also icons for user profile, heart, and shopping cart. The main content area displays a grid of phishing kits for sale, each with a preview image, a title, a price, and an 'ADD TO CART' button. The kits include:

- SharePoint-1 Phishing P... (SBA, TOP10) - \$100.00
- Outlook 5 Phishing P... (DROPBOX, OUTLOOK) - \$80.00
- Excel5 Phishing Page... (DROPBOX) - \$100.00
- Dropbox 18 Phishing ... (DROPBOX, TOP10) - \$80.00
- Office 365-Kumar CL... (OFFICE 365) - \$100.00
- Eastlink-1 Phishing P... (EASTLINK) - \$100.00
- Onedrive26 Phishing... (ONEDRIVE, TOP10) - \$80.00
- Cpanel-Webmail Phi... (WEBMAIL) - \$100.00
- Alibaba Style 1 Single... (ALIBABA) - \$100.00
- Outlook Style1 Scam ... (OUTLOOK) - \$80.00

Common attack types against common surfaces for common goals

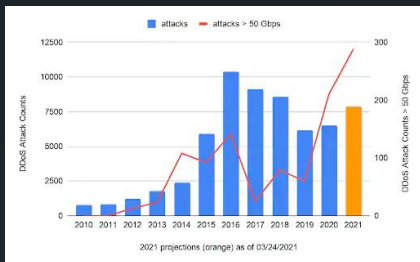
Paid & OSS hacking tools  
Phishkits, DDOS, ...

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

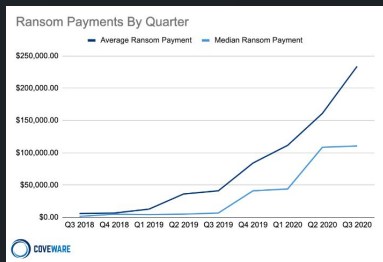
<https://therecord.media/microsoft-uncovers-giant-phishing-as-a-service-operation>



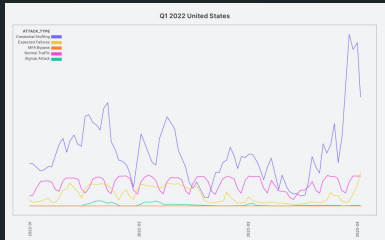
# Attackers double down on successful attack vectors & attack surfaces



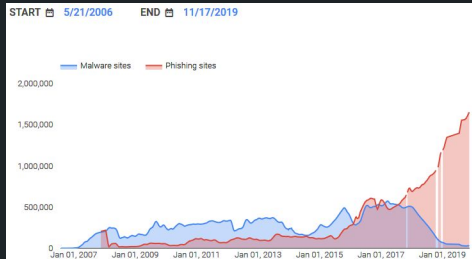
Denial of service  
Zayo: 2010-2021



Ransomware  
Coveware: 2018-2020



Credential stuffing  
Okta: 2022



Malware & phishing  
Google Safe Browsing: 2007-2019

Examples:

Ransomware blew up because bitcoin

Phishing => auto-draining e-banking accounts

Malware => recruitment into botnet => sell DDoS

Password stuffing => credit card testing => theft



# 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





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** **110K**  
SHARES READS

Web search breaking from GPT-generated SEO pages



r/SEO · 3 days ago  
by gronetwork

Join ...

## 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?

## Google Search's guidance about AI-generated content

[Send feedback](#)

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:

1. **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.
2. **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.
3. **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.
4. **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.
5. **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.

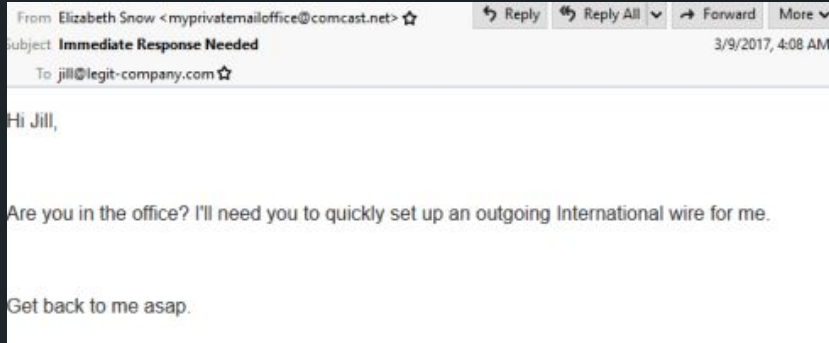
Considering the project's 2-4 week timeline, I am confident in delivering a robust, hardened sandbox environment that satisfies your security requirements. I'm familiar with Github, Slack, and am flexible with meeting times for PDT project leads.

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

Best regards,  
[Your Name]



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



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



## 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).



LOUIE.AI

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



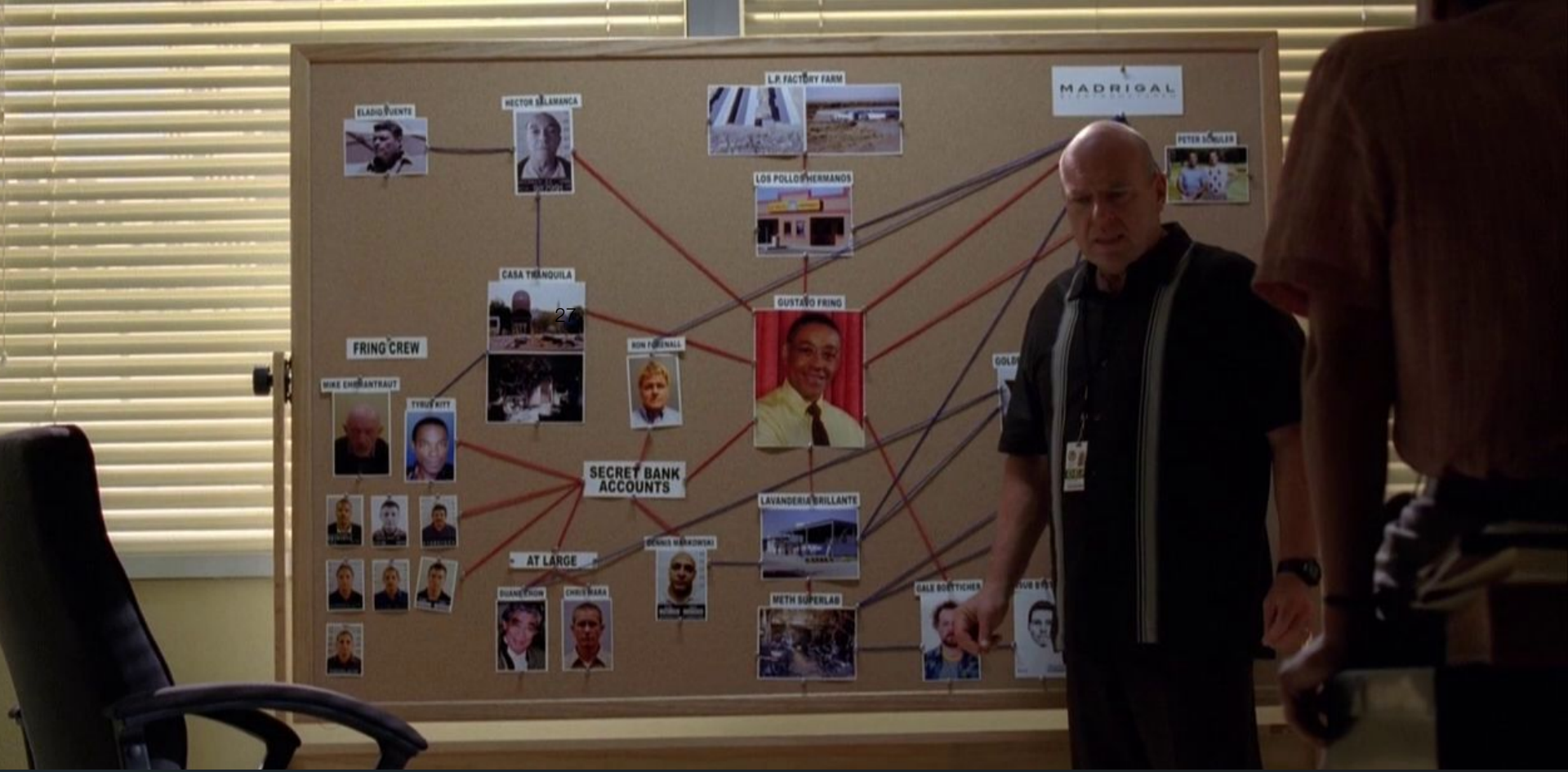
# Defenders perspective: Representative Tier 2 SOC session

26

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







27



Malicious Capabilities Observed in the VM

Malicious Behavior: Yes
OS Change Summary

Malware detected

Table with columns: Malware, Severity, Total Infections, Callbacks, Blocked, Botnets, Last CnC Server, Last Location, First Seen, Last Seen, Ports Used, Protocols. Rows include Malware.archive, Trojan.CVE2013, and Exploit.Browser.

IPS Details

Table with columns: Signature Name, First Seen, Last Seen, Categories, Protocol, Reference. Row: Microsoft Internet Explorer VML Processing Integer Underflow (Generic Detection).

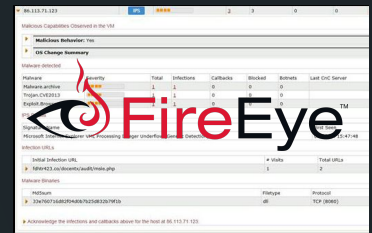
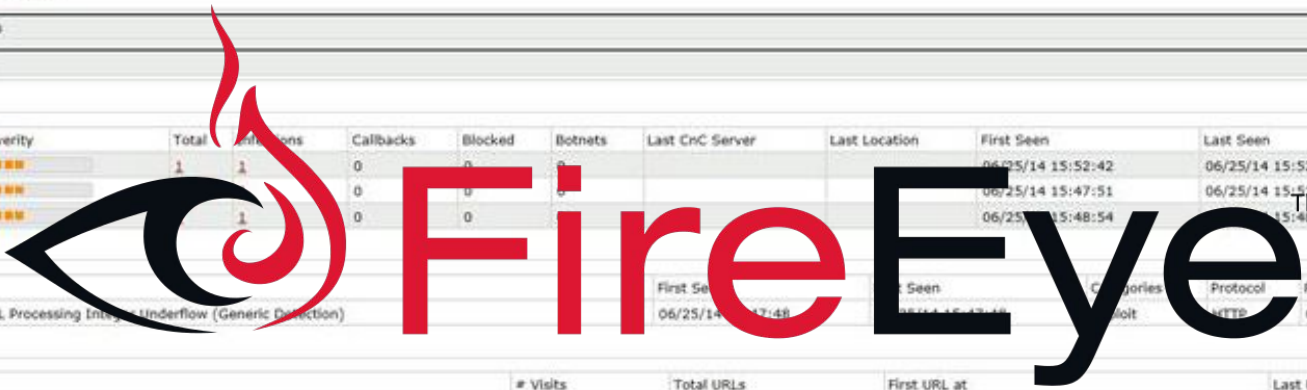
Infection URLs

Table with columns: Initial Infection URL, # Visits, Total URLs, First URL at, Last URL at. Row: fdhtr423.co/docentx/audit/msie.php.

Malware Binaries

Table with columns: Md5sum, Filetype, Protocol, Encoding, Last analysis time, # Occurrences. Row: 33e760716d82f04d0b7b25d832b79f1b.

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**.

VT Community



goodware

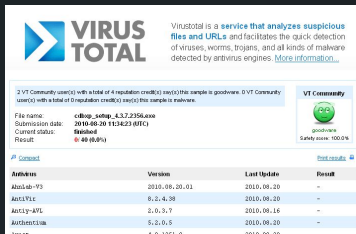
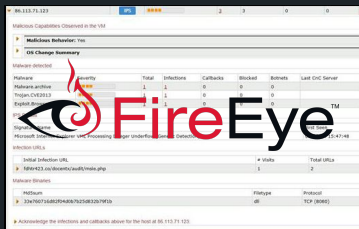
Safety score: 100.0%

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%)**

[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

Name	Source			Destination			URL Category	Service	Action	Profile
	Zone	Address	User	Zone	Address	Application				
LogAll	Tap	any	any	Tap	any	any	CustomerURLCategory	any	✓	[Icons]
IT Allow Override	trust	any	pancademo/administrators	untrust	any	Custom-app	any	any	✓	[Icons]
Read Only Facebook	trust	any	pancademo/administrators	untrust	any	Facebook-base	any	any	✓	[Icons]
Allow facebook posting	trust	any						any	✓	[Icons]
Block Peer to Peer	trust	any						any	✗	none
Webmail file blocking	trust	any						any	✓	[Icons]
Sharepoint	Untrust-L3	any						application-default	✓	[Icons]
Allow SSL and SSH	trust	any						any	✓	[Icons]
Allow Web-browsing	trust	Sharepoint S						any	✓	[Icons]
Block encrypted tunnel	trust	any						any	✗	none
Block Proxies and Anonymizers	trust	any	any	untrust	any	Proxies	any	any	✗	none
Mail server	Untrust-L3	any	any	DMZ	Mail Server FQDN	outlook-web	any	application-default	✓	[Icons]
						smtp				
Web server	Untrust-L3	any	any	DMZ	Web-server	ssl	any	application-default	✓	[Icons]
						web-browsing				

+ Add - Delete Clone Enable Disable Move Top Move Up Move Down Move Bottom Highlight Unused Rules 13 rule(s)

Webcam Capabilities Checked in the VM

File name: c:\temp\setup\_4.3.7.2356.exe  
 Submission date: 2010-08-20 11:24:23 (UTC)  
 Current status: Resolved  
 Result: 49 (0.0%)

Malware	Family	Total	Infections	Callbacks	Blocked	Subnets	Last CMC Server
Malware	Family	Total	Infections	Callbacks	Blocked	Subnets	Last CMC Server

Initial infection URL: [Link]  
 Total inflic: 1

Initial infection URL: [Link]  
 Total inflic: 2

Initial infection: [Link]  
 Malware: [Link]  
 Protocol: [Link]  
 IP: [Link]  
 TCP (8080)

IP: [Link]  
 TCP (8080)

Acknowledge the infections and callbacks above for the host at 98.113.71.121.

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...](#)

File name: c:\temp\setup\_4.3.7.2356.exe  
 Submission date: 2010-08-20 11:24:23 (UTC)  
 Current status: Resolved  
 Result: 49 (0.0%)

Antivirus	Version	Last Update	Result
Avast-4V3	2010.08.20.01	2010.08.20	-
Avast-5V1	5.2.4.36	2010.08.20	-
Avast-5V2	5.2.4.37	2010.08.16	-
Avast-5V3	5.2.4.5	2010.08.20	-
Avast-5V4	5.2.4.5	2010.08.20	-

Dashboard ACC Monitor Policies Objects Network Device Save

Help

Name	Source			Destination			URL Category	Service	Action	Profile
	Zone	Address	User	Zone	Address	Application				
LogAll	Tap	any	any	Tap	any	any	CustomerURLCategory	any	✓	[Icons]
IT Allow Override	trust	any	pancademo/administrators	untrust	any	Custom-app	any	any	✓	[Icons]
Read Only Facebook	trust	any	pancademo/administrators	untrust	any	Facebook-base	any	any	✓	[Icons]
Allow facebook posting	trust	any						any	✓	[Icons]
Block Peer to Peer	trust	any						any	✗	none
Webmail file blocking	trust	any						any	✓	[Icons]
Sharepoint	Untrust-L3	any						application-default	✓	[Icons]
Allow SSL and SSH	trust	any						any	✓	[Icons]
Allow Web-browsing	trust	Sharepoint S						any	✓	[Icons]
Block encrypted tunnel	trust	any						any	✗	none
Block Proxies and Anonymizers	trust	any	any	untrust	any	Proxies	any	any	✗	none
Mail server	Untrust-L3	any	any	DMZ	Mail Server FQDN	outlook-web	any	application-default	✓	[Icons]
						smtp				
Web server	Untrust-L3	any	any	DMZ	Web-server	ssl	any	application-default	✓	[Icons]
						web-browsing				

+ Add - Delete Clone Enable Disable Move Top Move Up Move Down Move Bottom Highlight Unused Rules 13 rule(s)



**Acquisitions** Search by hostname, IP, acquisition

Show: All acquisitions Status: All statuses Requested: Everyone

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

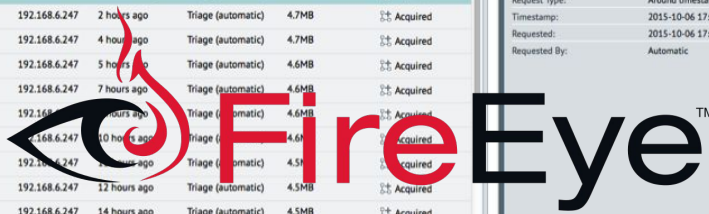
Hostname	IP Address	Requested	Acquisition	Download Size	Status
fe-xpsp3-victim	192.168.6.247	58 minutes ago	Triage (automatic)	4.7MB	Acquired
fe-xpsp3-victim	192.168.6.247	2 hours ago	Triage (automatic)	4.7MB	Acquired
fe-xpsp3-victim	192.168.6.247	4 hours ago	Triage (automatic)	4.7MB	Acquired
fe-xpsp3-victim	192.168.6.247	5 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	7 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	9 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	10 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	11 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	12 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	14 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	15 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	17 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	18 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	20 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	21 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	22 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	24 hours ago	Triage (automatic)	4.4MB	Acquired
fe-xpsp3-victim	192.168.6.247	25 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-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 17:15:20Z  
 Requested: 2015-10-06 17:17:04Z  
 Requested By: Automatic



Windows Security Center has the VED

1 VirusTotal scan: Done

1 Log Change Summary

Malware detected

Malware	Total	Infections	Capabilities	Blocked	Retrieved	Last CMC Server
0	0	0	0	0	0	0

File name: ollap\_setup\_4.3.7.2356.exe  
 Submission date: 2015-08-20 17:32:23 (UTC)  
 Current status: Resolved  
 Result: OK (0/0%)

**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...](#)

3/17 Community copies with total of 41 reputation checks (66%) in the sample is goodware. 0/17 Community copies with total of 0 reputation checks (0%) in the sample is malware.

File name: ollap\_setup\_4.3.7.2356.exe  
 Submission date: 2015-08-20 17:32:23 (UTC)  
 Current status: Resolved  
 Result: OK (0/0%)

Antivirus	Version	Last Update	Result
Avast	5.0.3.7	2016.09.16	OK
Avira	10.0.0.20	2016.09.20	OK
Avast	5.0.3.7	2016.09.16	OK
Avast	5.0.3.7	2016.09.20	OK
Avast	5.0.3.7	2016.09.20	OK
Avast	5.0.3.7	2016.09.20	OK
Avast	5.0.3.7	2016.09.20	OK
Avast	5.0.3.7	2016.09.20	OK
Avast	5.0.3.7	2016.09.20	OK
Avast	5.0.3.7	2016.09.20	OK

**paloalto** Palo Alto Networks

Home | Products | Solutions | Services | Support | About Us

Acquisitions (276) Acquisition space: 52.2% full - 12.6GB remaining

Hostname	IP Address	Requested	Acquisition	Download Size	Status
fe-xpsp3-victim	192.168.6.247	58 minutes ago	Triage (automatic)	4.7MB	Acquired
fe-xpsp3-victim	192.168.6.247	2 hours ago	Triage (automatic)	4.7MB	Acquired
fe-xpsp3-victim	192.168.6.247	4 hours ago	Triage (automatic)	4.7MB	Acquired
fe-xpsp3-victim	192.168.6.247	5 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	7 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	9 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	10 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	11 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	12 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	14 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	15 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	17 hours ago	Triage (automatic)	4.5MB	Acquired
fe-xpsp3-victim	192.168.6.247	18 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	20 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	21 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	22 hours ago	Triage (automatic)	4.3MB	Acquired
fe-xpsp3-victim	192.168.6.247	24 hours ago	Triage (automatic)	4.4MB	Acquired
fe-xpsp3-victim	192.168.6.247	25 hours ago	Triage (automatic)	4.6MB	Acquired
fe-xpsp3-victim	192.168.6.247	27 hours ago	Triage (automatic)	4.5MB	Acquired



Grouped By: Status X

Icon	NAME	FILE PATHS	STATUS	FIRST FOUND	RUNNING	AUTO RUI	CLASSIFICATION	DETECTED E
<input type="checkbox"/>	29b449bac81d654b52cd0a32bdf29ed955f046ef9db1e0eba7da47ab2950a3ae	c:\users\admin1\downloads\52cd0a32bdf29ed955f046ef9db1e0eba7da47ab2950a3ae	Quarantined	5/19/2016	No	No	File Watcher	
<input type="checkbox"/>	4f39d3e70ed1278d5f8a3ed9f148ca92383ac662ca34635f7e56cc42eeae948	c:\users\admin1\download\5f8a3ed9f148ca92383ac662ca34635f7e56cc42eeae948	Quarantined	5/19/2016	No	No	File Watcher	
<input type="checkbox"/>	70cc1e06fbfb986627d624c5337b8f4cf4e465a6783745368bb9528548a0ae9b	c:\users\admin1\downloads\70cc1e06fbfb986627d624c5337b8f4cf4e465a6783745368bb9528548a0ae9b	Quarantined	5/19/2016	No	No	File Watcher	
<input type="checkbox"/>	72513534f2e0f3e77a22023b887df3718c9df70686e0a858cbbde2f90f447e4	c:\users\admin1\downloads\72513534f2e0f3e77a22023b887df3718c9df70686e0a858cbbde2f90f447e4	Quarantined	5/19/2016	No	No	File Watcher	
<input type="checkbox"/>	766e49811c0bb7cce217e7273a6a866c15de0ba11d7dga3bd7e9ec33ed6963	c:\users\admin1\downloads\766e49811c0bb7cce217e7273a6a866c15de0ba11d7dga3bd7e9ec33ed6963	Quarantined	5/19/2016	No	No	File Watcher	

1 - 13 of 13 items

FireEye

VirusTotal

File name: cobsp\_vsbp\_433235.exe  
Submission date: 2016-08-20 13:24:23 (UTC)  
Scan status: Clean  
Result: 0/49 (0%)

paloalto  
NETWORKS

Acquisitions

FireEye

CYLANCE



The screenshot shows the Windows Event Viewer interface. The left pane shows the tree view with 'Security' selected. The main pane displays a list of security events. The 'Actions' pane on the right has 'Save All Events As...' highlighted with a red box. A red arrow points from this action to a blue Windows logo overlay that is placed over the event list. Another red arrow points from the logo to the 'Save All Events As...' action.

Keywords	Date and Time	Source	Event ID	Task Category
Audit Success	6/5/2012 8:06:26 AM	Microsoft Windo...	4672	Special Logon
Audit Success	6/5/2012 8:06:26 AM	Microsoft Windo...	4624	Logon
Audit Success	6/5/2012 8:06:16 AM	Microsoft Windo...	4672	Special Logon
Audit Success	6/5/2012 8:06:16 AM	Microsoft Windo...	4624	Logon
Audit Success	6/5/2012 8:06:16 AM	Microsoft Windo...	4648	Logon
Audit Success	6/5/2012 8:06:16 AM	Microsoft Windo...	4776	Credential Validation
Audit Success	6/5/2012 8:06:09 AM	Microsoft Windo...	4672	Special Logon
Audit Success	6/5/2012 8:06:09 AM	Microsoft Windo...	4624	Logon

The screenshot shows the FireEye network security monitoring interface. It features a large 'FireEye' logo with a stylized eye icon. Below the logo, there are various data points and status indicators for network security events.

The screenshot shows a Windows Event Viewer window with a blue Windows logo overlay. A 'Save All Events As...' dialog box is open, showing a file name field and a 'Save' button. The dialog box is positioned over the event list.

The screenshot shows the EYLANCE security monitoring interface. It features a large 'EYLANCE' logo with a stylized green 'E' icon. Below the logo, there are various data points and status indicators for network security events.







```

source=wineventlog EventCode=4624 (Logon_Type=10 OR Logon_Type=7 OR
Logon_Type=2) host=$Hostname3$
| rex "New Logon:\s+Security ID:\s+(?&lt;account&gt;.*)"
| 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>

```

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

i	_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_cluste dmc_group_index dmc_group_licens 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_cluste dmc_group_deploy dmc_group_index dmc_group_licens 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_cluste dmc_group_deploy dmc_group_index dmc_group_licens






VT Community [Sign in](#) ▾ Languages ▾

# 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.

**VT Community**  
  
 goodware  
 Safety score: 100.0%

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%)**

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

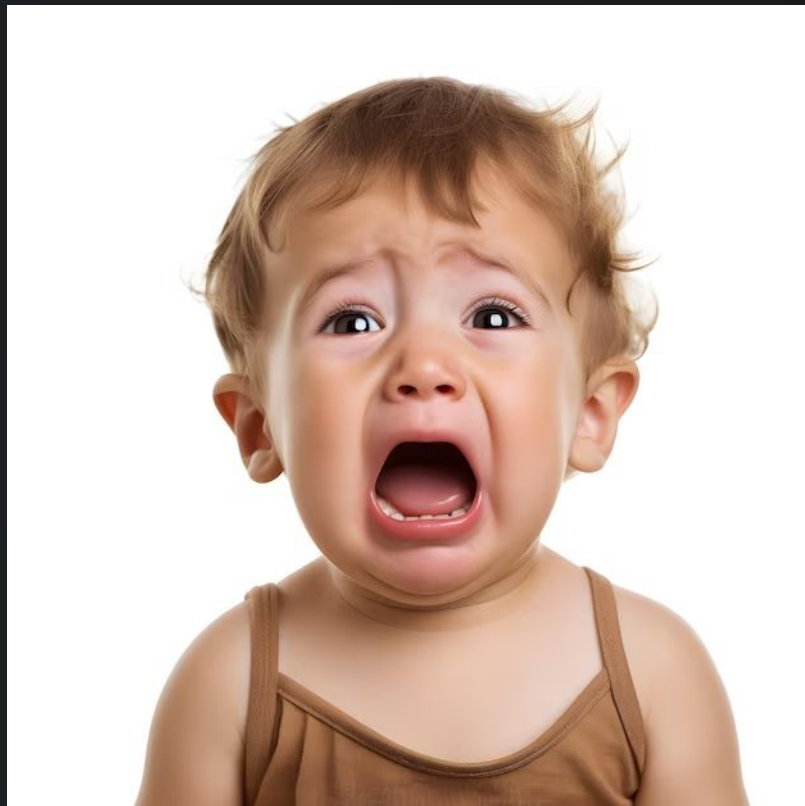
Antivirus	Version	Last Update	Result
AhnLab-V3	2010.08.20.01	2010.08.20	-
AntiVir	8.2.4.38	2010.08.20	-
AntiV-ATL	2.0.0.7	2010.08.16	-
Authentium	5.2.0.5	2010.08.20	-
Avast	4.8.1351.0	2010.08.20	-

... 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

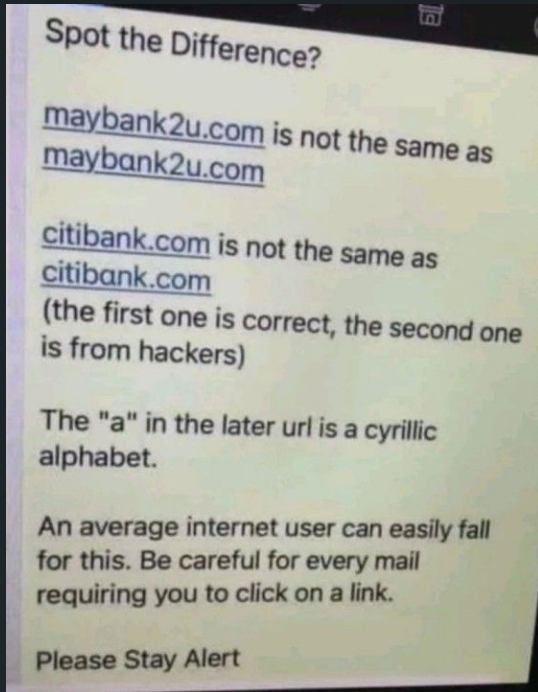
## Security operations

- 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

Questions	Yes/No	Comments
1. Is there a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
2. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
3. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
4. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
5. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
6. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
7. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
8. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
9. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
10. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
11. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
12. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
13. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
14. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
15. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
16. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
17. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
18. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
19. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.
20. Do you have a data retention policy approved by management, communicated to consultants and all employees, and published in a publicly accessible location?	Yes	The retention policy is approved by management, communicated to consultants and all employees, and published in a publicly accessible location.

## Security questionnaires

many, ex: vanta

**HIGH Data Destruction** Send

Do you have a destruction mechanism for the email addresses column?

**Description**

The table designated as EU\_USERS will duplicate existing PII (email addresses), necessitating its own specific mechanism for data deletion to ensure compliance with Acme's data retention policy.

See more | Regenerate with feedback | 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

Introducing Charlotte AI, CrowdStrike's Generative AI Security Analyst: Ushering in the Future of AI-Powered Cybersecurity

May 30, 2023 Michael Serrano Endpoint & Cloud Security

Announcing

### Microsoft Security Copilot

Security AI Workbench

Sec-PaLM

Partners: Extend and customer

Native: Security Command Center, Standard Threat Intelligence, Chronicle

Workload: AI on Google Cloud Platform

Use cases: Incident response, Threat intelligence, Compliance, Vulnerability management, Risk, Policy, Audit, Incident response, Remediation, etc.

### Google Launches Security AI Workbench to Rival Microsoft's Security Copilot

## 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, ...



# 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

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 as

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

### GPDR

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 later

## #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

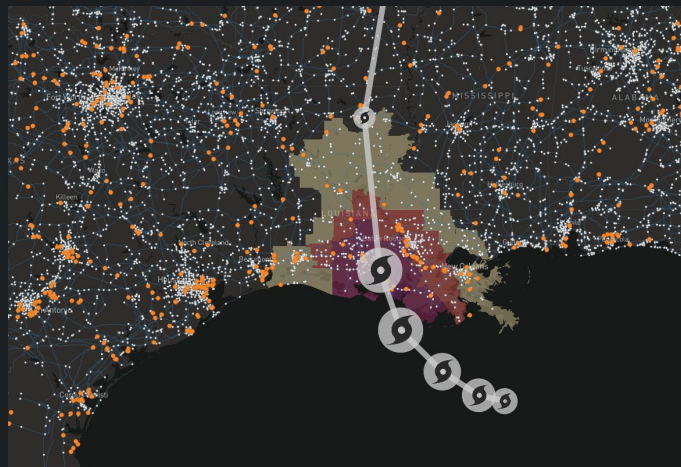
Challenge: Misinformation

⇒ 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

**DISASTERTECH**



GRAPHISTRY



LOUIE.AI



GRAPHISTRY



LOUIE.AI

info@graphistry.com

# 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

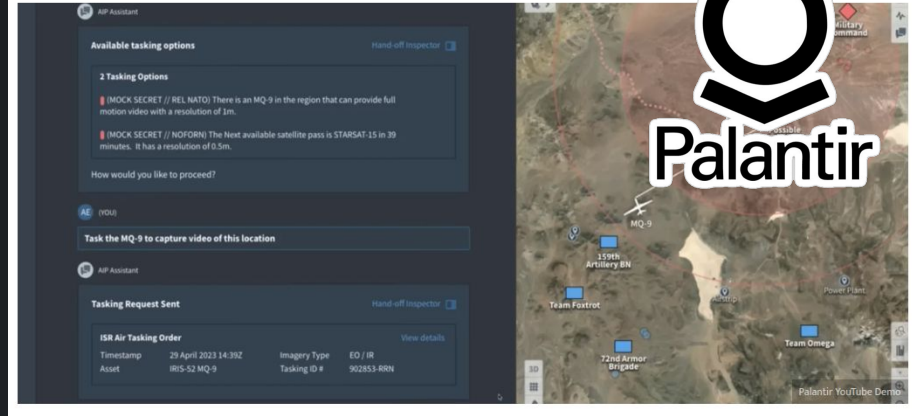
Coordination & multi-lingual allies

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





# 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, ...



# 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



# 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 imporeves:** Autonomous (autoGPT) mode + economical line-rate monitoring

## BROADER OPPORTUNITIES & CHALLENGES

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

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

**Democratic govns:** Under attack from inauthentic information that gets digitally implified

