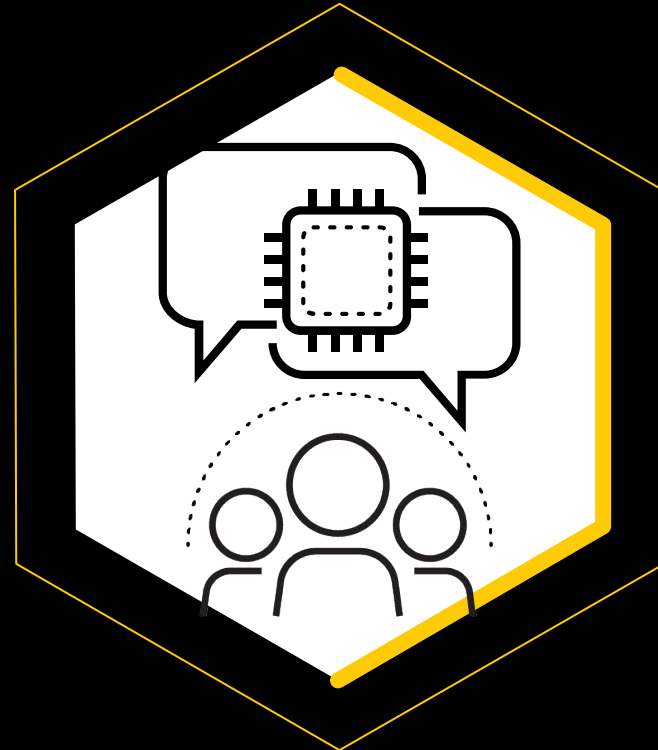


Overleve som beskytter i AI-æraen

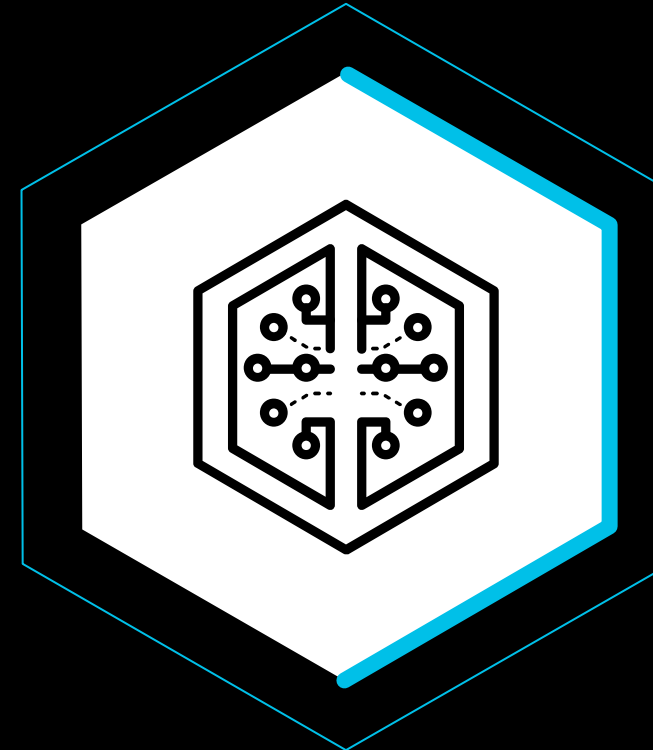
Alexander Hatlen | Domain Consultant
September 2026

Tre pilarer av AI-sikkerhet. Beskytte..



AI-bruk

Beskytt dataen din fra offentlige SaaS-LLMs



AI-agenter/modeller

Beskytte dine egne Enterprise Agenter



“The business”

Håndtere våpenkappløpet med AI-modell-bruk

AI-bruk



AI-bruk | AI-modeller | The business

Data:

- DLP
- Dataklassifisering (“reverse LLM”)
- Hvor er dataene?
- Hvilke identiteter har tilgang på dataen? (maskiner, brukere og agenter/apps)
- Protect the transit of classified files

Apps:

- Discover hva du faktisk har i bruk
- Compliance og kontroll på identiteter
- Nettleser er din mest brukte applikasjon..



Precise control, complete visibility, and robust protection for your GenAI apps—all in one place

Plugins ⓘ

21

Installed Plugins >

38

Plugin Users >

Threats

1.62k

Threats Detected >

1.62k Alerted >

0 Blocked

Sensitive Data Assets

381

Data at Rest >

1.61k

Events >

0 Alerted

1.61k Blocked >

● Sanctioned (12/124) ● Tolerated (25/124) ● Unsanctioned (87/124)

Nettverk



Copilot Recommended Actions (7)

Enable DLP Inspection of Sensitive Data for Allowed GenAI Apps

There are GenAI apps accessed over the past 7 days with no DLP inspection. Set up DLP for GenAI apps you allow access

View Recommendation

Review GenAI Plugins with High Severity Access Scopes

We detected GenAI plugins with High severity access scopes. Review these GenAI app plugins and reduce access privileges

View Recommendation

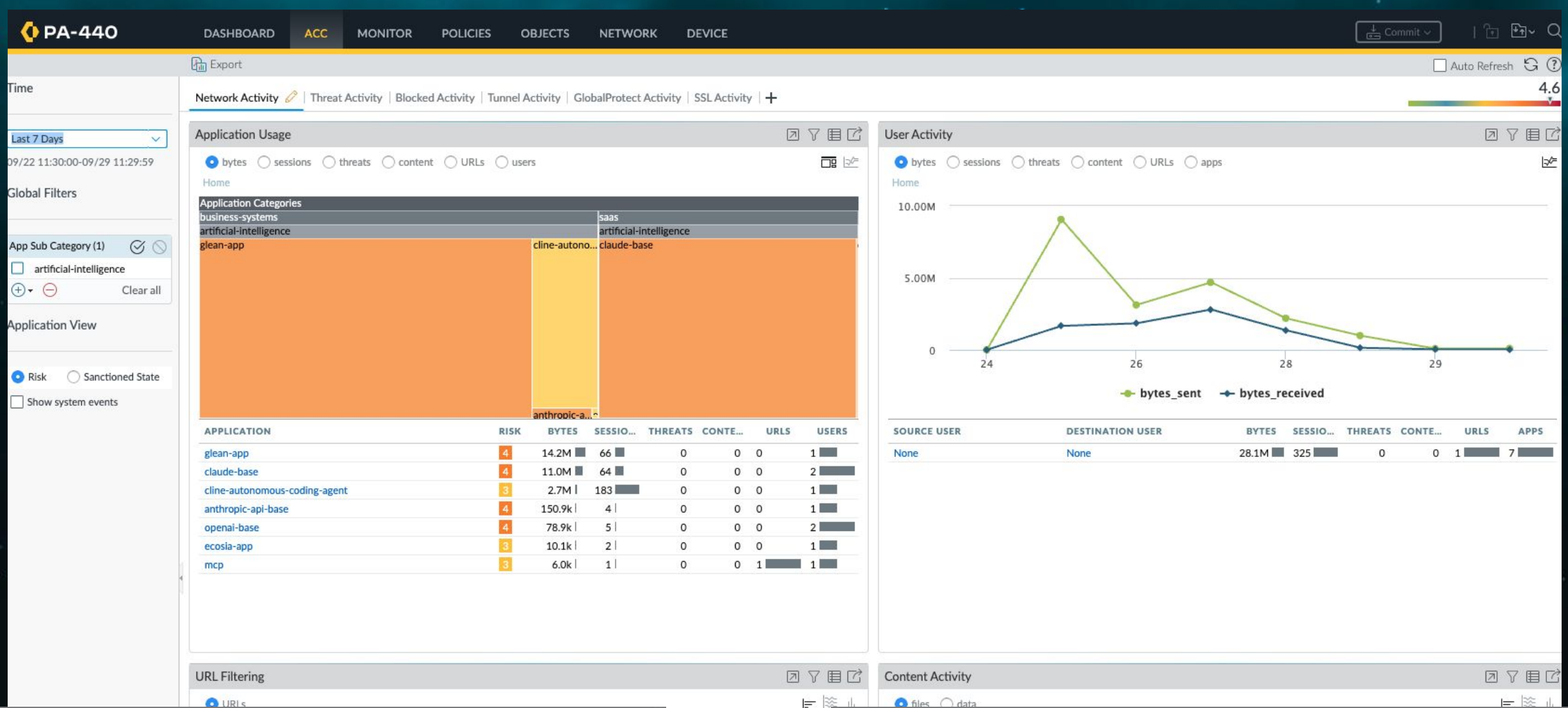
Review Policy Rules Allowing Access to GenAI Apps Without DLP Inspection

24 Security policy rules allow access to GenAI apps without DLP inspection. Set up DLP to prevent exfiltration of sensitive

View Recommendation



Apps



URLs

Custom Report

Report Setting

Load Template Run Now

Name: TopAI

Description:

Database: URL Log

Scheduled:

Time Frame: Last 6 Hrs

Sort By: Users Top 10

Group By: URL 50 Groups

Available Columns:

- Source User
- Source Vendor
- Source Zone
- Tunnel ID
- Tunnel Type

Selected Columns:

- URL
- URL Category List
- Users
- Count

Query Builder:

(category eq artificial-intelligence)

Filter Builder

OK Cancel

Custom Report

Report Setting TopAI (100%)

	URL	URL CATEGORY LIST	USERS	COUNT
22	copilot.microsoft...	"AI-conversational-assistant"	2	2
23		"low-risk"	2	2
24		"artificial-intelligence"	2	2
25	analytics.codede...	"low-risk"	1	2
26		"artificial-intelligence"	1	2
27		"AI-website-generator"	1	2
28	assets-v2.codedesign.ai/	"artificial-intelligence"	1	2

Export to PDF Export to CSV Export to XML

OK Cancel

KOI

Use case

Koi Playbook

Dashboard

Endpoints

Agentic AI

Overview

All agentic items

MCP Servers

Skills

AI Models

Plugins

AIDR

Non-Binaries

App Control

Governance

Guardrails

Policies

Remediation

Settings

Dark mode

Docs

Palo Testing

Inventory / Extensions

Chrome 170

VS Code 116

Edge 36

Cursor 33

Open VSX 28

Firefox 27

JetBrains 5

Notepad++ 3

Query Builder

And

+ Add Condition

+ Group

Clear

Run

Query builder

Saved Queries

Unique Items 418

Total Instances 724

Inventory

View by

Version

Item

Name	Risk level	Governed by	Version	Publisher	Marketplace	Platform	Endpoints	Find
☐ Claude Code for VS Code Extensions	Low	Allow low ris...	2.1.212	Anthropic	VSCode	VSCode	1	
☐ Remote - SSH Extensions	Med	-	1.1.11	Anysphere	Cursor	Cursor	1	
☐ GitHub Actions Extensions	Low	Allow low ris...	0.32.3	GitHub	Cursor	Cursor	1	
☐ Claude Extensions	High	-	1.0.81	Anthropic	Chrome Web Store	Chrome	1	
☐ DeepL: translate and write with AI Extensions	Med	-	1.95.0	team-browser-e...	Chrome Web Store	Chrome	1	
☐ Google Keep Chrome Extension Extensions	Low	Allow low ris...	4.26281.540.1	google.com	Chrome Web Store	Prisma Access B...	1	
☐ Context7 MCP Server Extensions	Low	Allow low ris...	1.1.0	Upstash	VSCode	VSCode	1	
☐ Prisma Access RDP Client Extensions	-	-	1.182.0	-	Chrome Web Store	Prisma Access B...	7	
☐ Dev Containers Extensions	Low	Allow low ris...	0.466.0	Microsoft	VSCode	VSCode	1	
☐ CursorCode(Cursor for VSCode)								

Unique items 418 | Instances 724

Showing 50 /418

1 2 3 4 5 ... 9

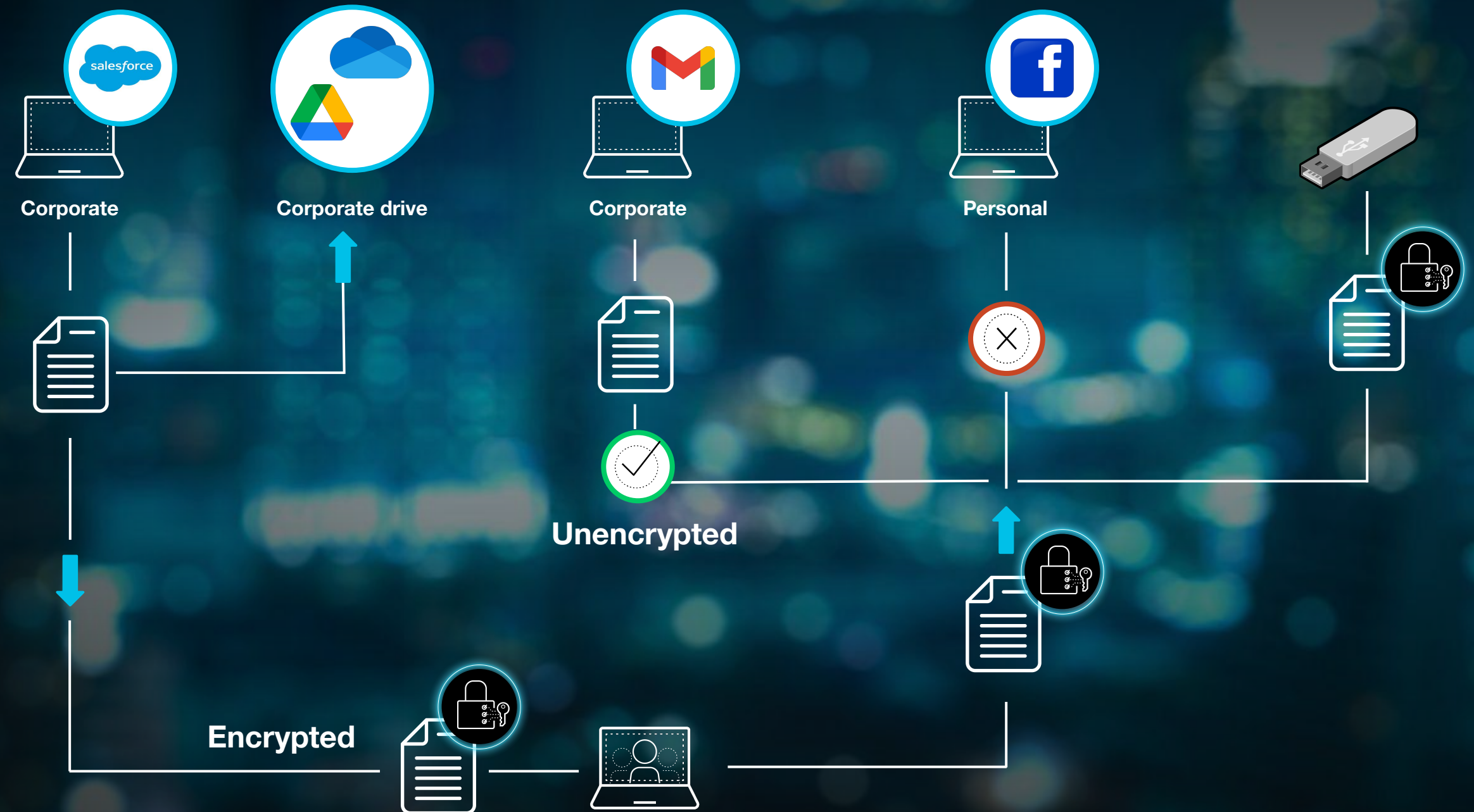
Endepunkt

Protecting the transit of data by controlling the accessing medium

WITH Directional Context for
Modern Data Protection:
Always Assume Compromise

For example:

- Allow **file movement** or **copy/paste only** between sanctioned apps
- Allow **file download only** to enterprise cloud drive
- Block or **encrypt** transfer from sanctioned apps **to** private accounts or flash drives

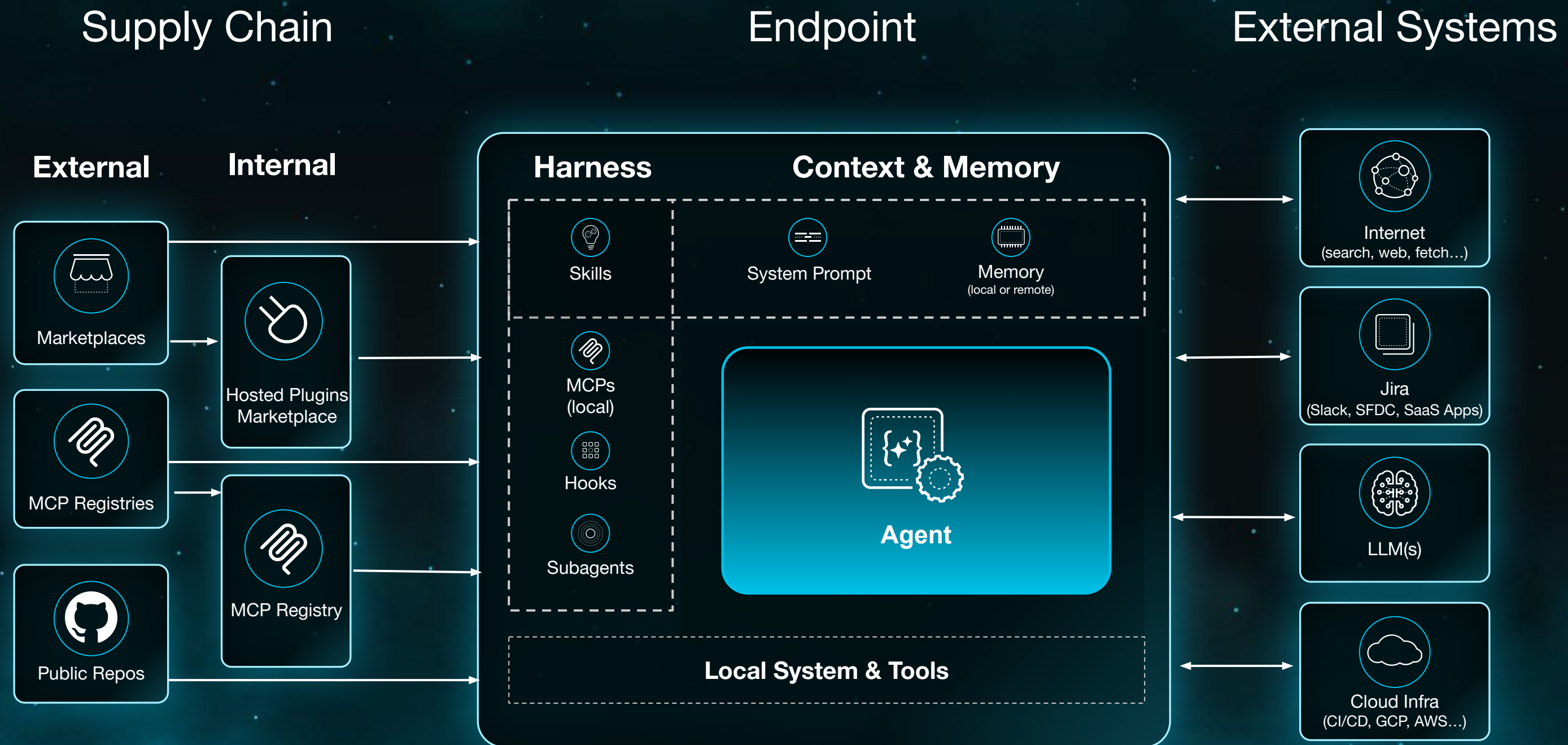


AI-agenter/-modeller

Protecting the brand new attack vectors (OWASP Top 10 for LLM)

Model protection	Data protection	App protection	Agent protection	AI safety
Prompt injection	Training data	Real-time URL/reference filtering	Memory poisoning	Context grounding
System prompt leakage	Context/data poisoning	Exploitation	Malicious tools	Red teaming
Malicious models	PII/sensitive leaks	Secure authentication	Tool misuse	Consumption control
	Data access control		Excessive agency	Toxic content

AI Agents Have Moved to the Endpoint



Built for the AI Enterprise

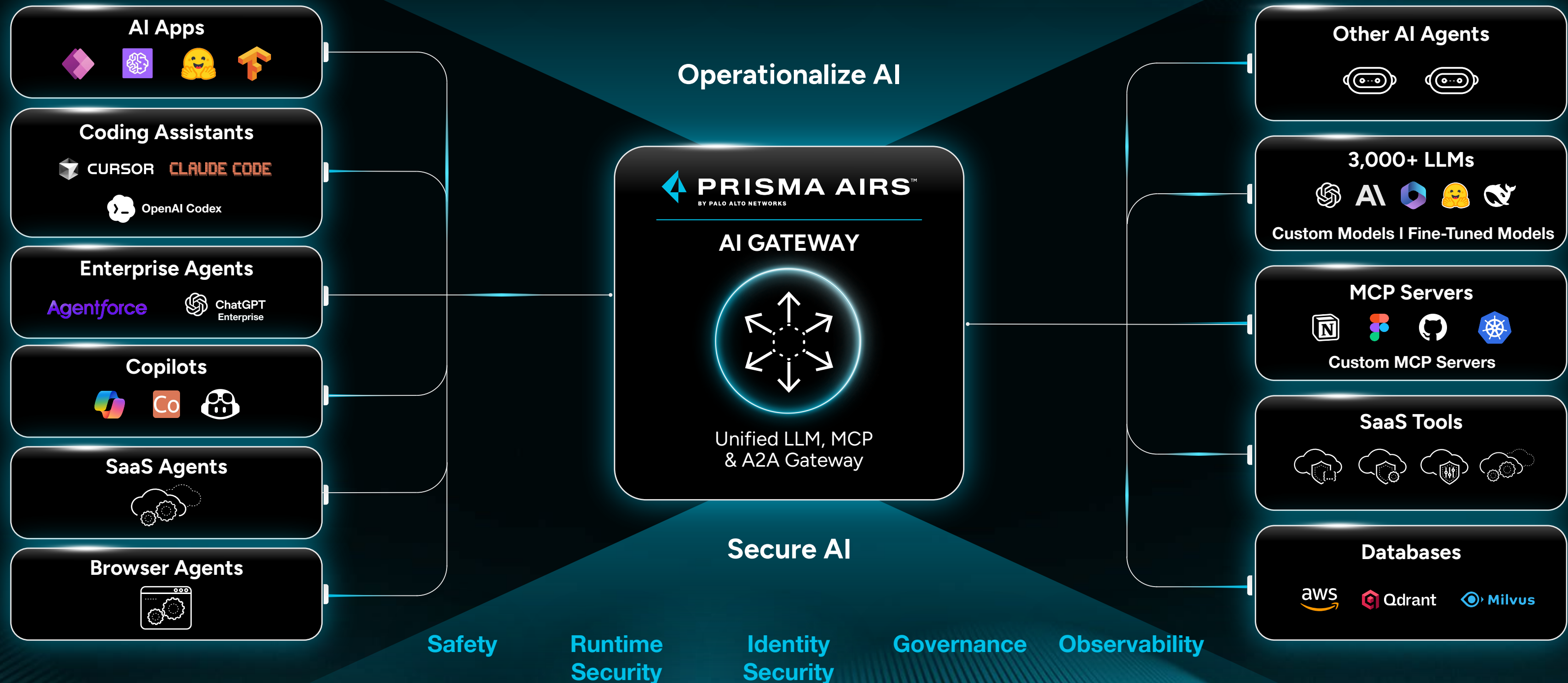
Prompt
Management

Universal API to
3,000+ LLMs

Load
Balancing

Semantic
Routing

FinOps and
Usage Mgmt.



Ting å beskytte seg mot



AI Model Protection

30+ types of prompt injections protected

Custom **topic guardrails** and **toxic content** detection



AI Data Protection

1,000+ predefined data patterns

2x greater coverage than other cloud-based data security solutions



AI App Protection

Malicious URL detection based on database of **100M+** malicious and benign URLs across **74 categories**

99.58% malware detection accuracy with **26% more detection** than traditional sandboxes



AI Agent Protection

Block AI agent-specific threats such as **memory manipulation**, **tool misuse** and **hallucinations**

Secure AI agents in **no-code**, **low-code** and **pro-code** environments

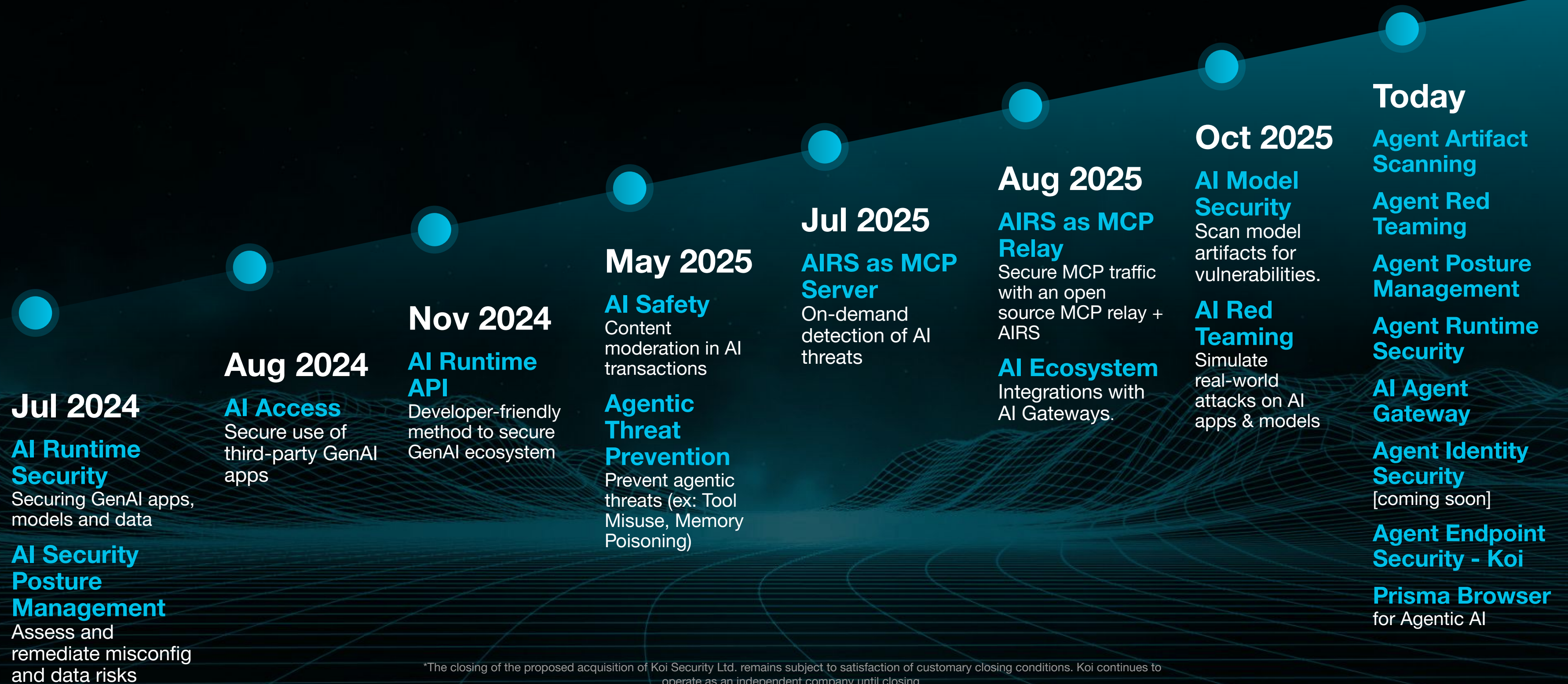


AI Safety

Detect and block **harmful/toxic content**

Custom **topic guardrails** and **harmful/toxic content** detection

Det skjer mye hele tiden i AI-sikkerhet...



Grunnleggende tiltak

- Ha kontroll på identiteten agenten opererer med. Handle gjerne på vegne av brukeren selv, men på en trygg måte
- Kjør agentene i et sikkert miljø, f.eks. Nvidias OpenShell
- Styr hvilke verktøy agenten kan benyttet
- Bruk gjerne tillatelses-listing på nettverk, filaksess, verktøy (med mer?)
- Bruk Guardrails
- Ha kostnadskontroll

Demo time!



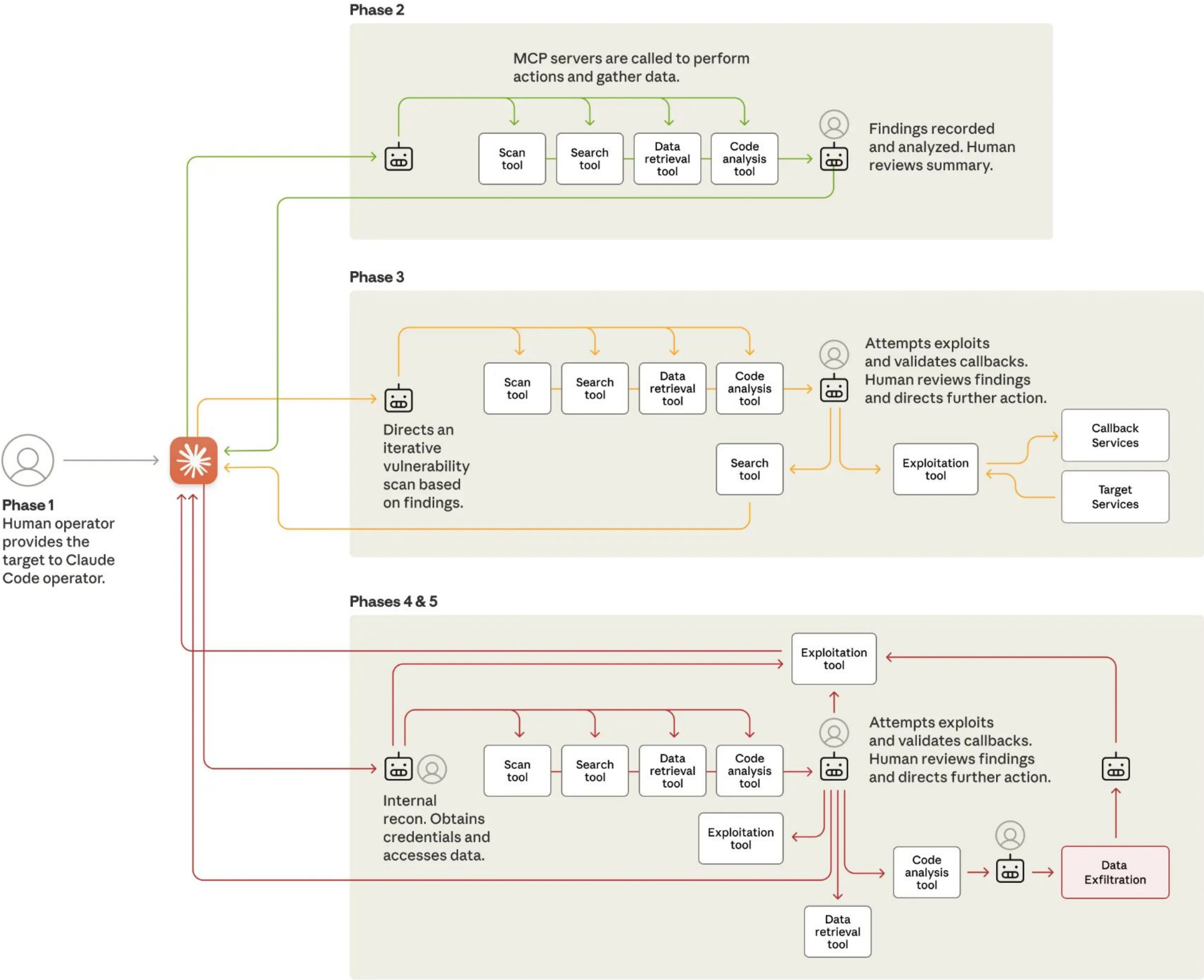
Annet

Vær forsiktig med å la LLM gjøre beslutningstaking, det er mange tiltak som bør gjøres for å få mer presis LLM:

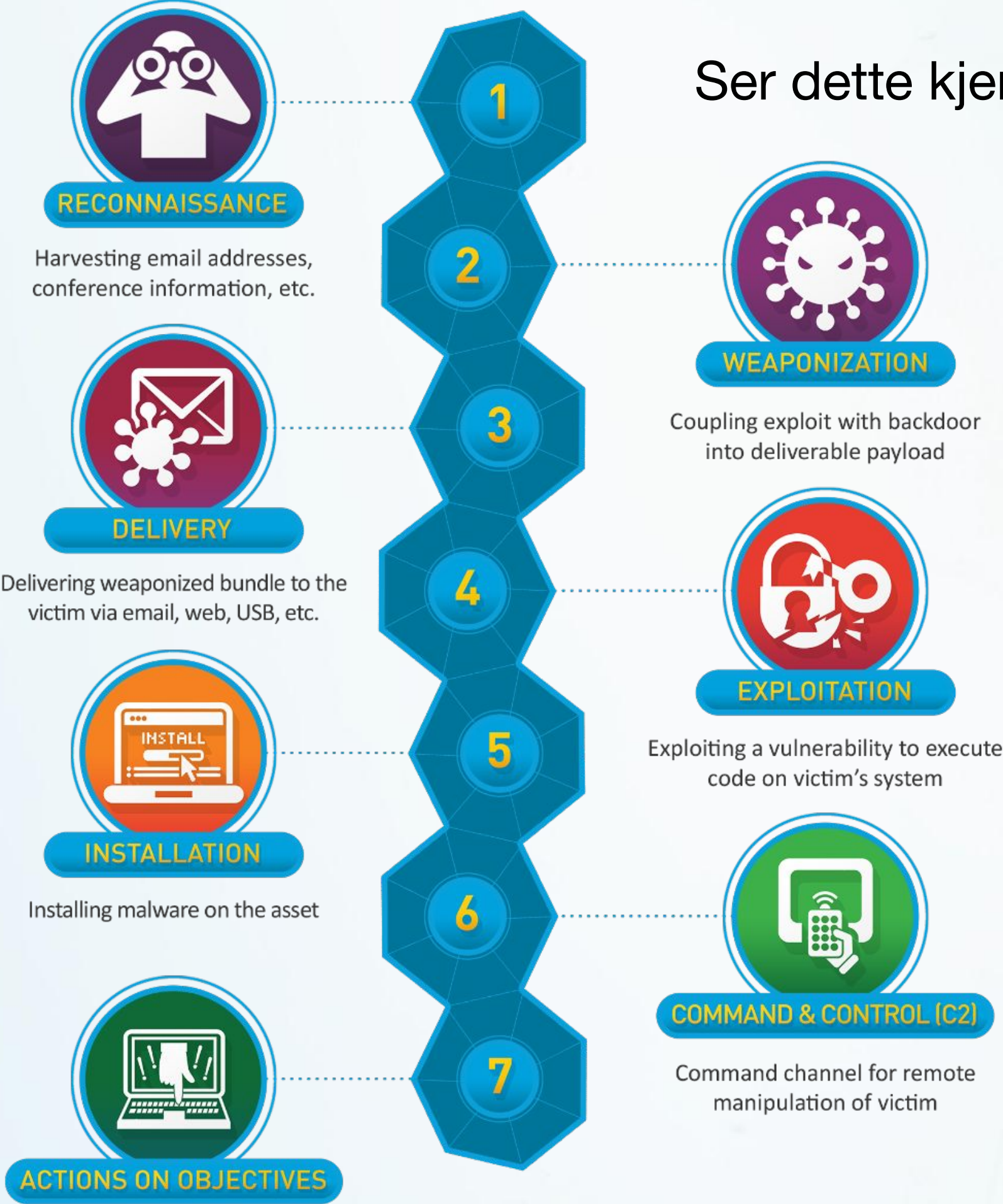
- Senk temperatur til 0
- Gi en system-prompt om at den skal si ifra om den ikke har kontekst-data til å svare, og be den alltid gi refereanser på påstander
- Bruk en modell som fine-tuned til å bruke referanser og generelt
- Bruk verktøy som matematisk hjelper med å forhindre at modellen gjør ting den ikke skal. Sjekk gjerne ut Jev!

The business

APTter som bruker Claude (Sep-Nov 2025..!)



Ser dette kjent ut?



AI-bruk | AI-modeller | The business

We have some new tools:

- Defensive tools for AI/LLM ecosystem
- Reverse Engineering with Ghidra MCP democratizes application testing - but blue team has direct access to source code!
- LLM is the first “tool” to actually scan and test for logic bugs!
- Democratize alerts understanding by sharing cyber knowledge

But it's not going to replace the good ol' cyber security practices. We just need to be great at the basics to break the adversaries' unfair tool advantage!

Old tools that still work:

- AI ML - Machine Learning is the best part of AI for blue teams
- Reducing attack surface, harden what's left
- Segmentation, zero-trust access for network and identities (humans and machine identities!)
- DLP, URL classification
- EDR
- Honestly? Most of them!



Attacker-benefits

Machine speed
interpreting scanning results
SoMe and web
scraping to build a graphic or linguistic profile

Reverse engineering
and logic bug fuzzing
Democratized coding

Automated attack path
by AI reading scan results
Deepfakes

Not much change in
non-AI infrastructure!
Maybe iteration.

Automated scans now
also auto-selects
next-steps and iterates

Not much new

Automatically label
data to exfil only value



Reduce attack surface
Build alerts here
Scan your exposure before
the attacker does

Quick patching required
Scan your code before
the attacker does

Use AI (ML) for
anomaly detection
Inspect and scan
encrypted
data-transfers

EDRs still rock
Iteration may trigger
tripwires and generate
solid alerts
Run automated
security actions on
alerts

Not AI, but new tools
create good inventory
Trigger alerts on new
unapproved
installations

Traffic pattern
recognition+detection,
even in encrypted
traffic.

We are running out of
time, but DLP and
segmentation can buy
you time and raise
critical alerts

Defender-benefits

Defender's dilemma:

"The defenders have to be right 100% of the time.
The attacker only needs to be right once."

BULLSHIT

Attacker-benefits

“The attacker has to step on just one of the decoys”

Red=minefield

Machine speed
interpreting scanning
results
SoMe and web
scraping to build a
graphic or linguistic
profile

Reverse engineering
and logic bug fuzzing
Democratized coding

Automated attack path
by AI reading scan
results
Deepfakes

Not much change in
non-AI infrastructure!
Maybe iteration.

Automated scans now
also auto-selects
next-steps and iterates

Not much new

Automatically label
data to exfil only value



Reduce attack surface
Build alerts here
Scan your exposure before
the attacker does

Quick patching required
Scan your code before
the attacker does

Use AI (ML) for
anomaly detection
Inspect and scan
encrypted
data-transfers

EDRs still rock
Iteration may trigger
tripwires and generate
solid alerts
**Run automated
security actions on
alerts**

Not AI, but new tools
create good inventory
Trigger alerts on **new**
unapproved
installations

Traffic **pattern**
recognition+detection,
even in encrypted
traffic.

We are running out of
time, but DLP and
segmentation can buy
you time and raise
critical **alerts**

Defender-benefits

Fortsatt ingen “silver bullet”

- Reduser angrepsflaten
- Bygg herdete og robuste systemer
- Lag et minefelt for angriperne
- Be great at the basics
- Plan for business continuity and disaster
- Bruk AI i sikkerhetsarbeidet, for å lære og automatisere
- Fokuser på kronjuvelene og verdiene dine, ikke verktøyene

Thank You

paloaltonetworks.com



Alexander Hatlen

System Engineer @ Palo Alto Networks



[@LinkedIn](#)