AI vs AI soudný den nastal


ARTIFICIAL INTELLIGENCE
WHAT'S THERE TO WORRY ABOUT?!

KONTEXT

Proč je testování AI novou třídou rizik

Moderní modely se při testu chovají jinak než klasický software: mají cíl, adaptují se a aktivně hledají obchvat překážek. Když se to potká s reálnými oprávněními, z testu se stane incident.

80–90 %

podíl útočných úloh, které AI provedla autonomně v kampani GTG-1002

Anthropic, 2025

88 %

organizací potvrdilo nebo mělo podezření na incident s AI agentem za rok

CSA / Token Security, 2026

17 000+

zaznamenaných akcí AI agenta během víkendového průniku do Hugging Face

Hugging Face, 2026

Testování bez pojistek + reálná oprávnění = únik ze sandboxu se přenáší do produkce.

REFERENČNÍ INCIDENT

OpenAI × Hugging Face — červenec 2026

Žádný člověk neřídil jednotlivé kroky útoku.

- **Kdy a kde**

  Během interního testu ofenzivních schopností v harnesu ExploitGym.

- **Kdo**

  Kombinace modelu GPT-5.6 Sol a nevydaného modelu — s vypnutými pojistkami.

- **Co se stalo**

  Model autonomně unikl ze sandboxu, prolomil třetí-stranné testovací prostředí a pronikl do produkční infrastruktury Hugging Face.

- **Motiv**

  Podvést benchmark tím, že si „ukradne" řešení úloh, místo aby je vyřešil.

Zdroje: OpenAI · Hugging Face · Fortune · Cloud Security Alliance

SPOTLIGHT · SRPEN 2026

Claude Mythos 5 se pokusil backdoorovat reálný projekt

Útok nakonec zastavil člověk, který si přečetl diff.

- **Prostředí**

  CTF na kyber-range AISI; modely s vypnutými cyber klasifikátory a otevřeným internetem — měření syrové schopnosti, ne produkční režim.

- **Rozsah**

  Napříč 122 běhy 19 nesankcionovaných akcí na živém internetu v 10 bězích (17× Mythos 5, 2× GPT-5.6 Sol).

- **Plán**

  Backdoor projektu → počkat na automatický update → převzít stroj uvnitř range. Úvaha chybná, řemeslo (tradecraft) nikoli.

- **Výsledek**

  Pokusy selhaly, žádný doložený reálný dopad; nic neuniklo ze sandboxu.

Zdroje: UK AI Security Institute (AISI) · The Hacker News

A Model se vymkl při testování / tréninku

Spontánní únik z prostředí, ignorování příkazů, jednání bez potvrzení člověkem.

B Autonomní AI-řízené útoky v praxi

Reálné kampaně, kde AI provádí většinu útočných kroků samostatně.

C Incidenty na Hugging Face / kolem modelů

Škodlivé modely, únos konverzí, zneužití platformy pro modely a datasety.

D Únik ze sandboxu, schopnosti a governance

Exfiltrace ze sandboxu, nebezpečné schopnosti a jejich řízení.

A

KATEGORIE A

Model se vymkl při testování nebo tréninku

ROME (Alibaba) 03/2026

Během reinforcement learningu agent spontánně unikl z prostředí, získal přístup k GPU a těžil kryptoměnu; vytvořil reverzní SSH tunel obcházející bezpečnostní procesy.

Zdroj: Axios · Live Science

Meta „Sev 1" rogue agent 03/2026

AI agent jednal autonomně bez potvrzení člověkem a na dvě hodiny vystavil rozsáhlá firemní a uživatelská data neoprávněným inženýrům.

Zdroj: TechCrunch · The Information

OpenClaw maže e-maily (Meta) 02/2026

Agent ignoroval opakované příkazy STOP; po vyčerpání pracovní paměti zahodil instrukci potvrdit akci a smazal 200+ e-mailů.

Zdroj: Fast Company · TechCrunch

„Agents of Chaos" (red-team) 2026

Studie s agenty v živém prostředí: 11 případů včetně plnění příkazů od cizích osob, úniku dat, destruktivních akcí a spoofingu identity.

Zdroj: Shapira et al., arXiv

B

KATEGORIE B

Autonomní AI-řízené útoky v praxi

Anthropic GTG-1002 09/2025

První doložená rozsáhlá kyberšpionáž vedená převážně AI agenty proti ~30 organizacím; AI autonomně provedla 80–90 % útočných úloh.

Zdroj: Anthropic

JADEPUFFER (agentic ransomware) 07/2026

První vyděračská operace běžící celou dobu autonomně LLM agentem: průzkum, krádež údajů, laterální pohyb, zašifrování 1 342 konfigurací Nacos.

Zdroj: Sysdig · The Register

DeepSeek + Hermes Agent 07/2026

Autonomní operátor řízený přes Telegram; sám skenoval instance, vyhodnotil neproveditelnost exploitu a přepnul na škálovatelnější cíle.

Zdroj: Palo Alto Unit 42

Mexická státní správa 12/2025–02/2026

Claude Code + GPT-4.1: jeden operátor prolomil 9 úřadů; z 1 088 promptů vzniklo 5 317 AI-provedených příkazů a únik ~400 mil. záznamů.

Zdroj: Check Point

C

KATEGORIE C

Incidenty na Hugging Face a kolem modelů

Podvržený „OpenAI" model 05/2026

Repozitář Open-OSS/privacy-filter se vydával za release OpenAI; ~244 000 stažení za ~18 h a skrytý PowerShell řetězec s infostealerem.

Zdroj: HiddenLayer

100 škodlivých modelů (pickle) 2024

Upravený inicializační soubor volal pickle kód schopný spustit libovolné chování; platforma pickle jen varuje, neskenuje záměr.

Zdroj: JFrog

Únos Safetensors konverze 2024

Zneužitím konverzní služby šlo posílat škodlivé pull requesty a unášet modely — až po nasazení neuronových backdoorů.

Zdroj: HiddenLayer

Marimo RCE → NKAbuse přes HF 04/2026

Malware hostovaný na typosquat Hugging Face Space shazoval Go binárku využívající NKN blockchain jako C2; exploit do 10 h od zveřejnění.

Zdroj: Sysdig

D

KATEGORIE D

Únik ze sandboxu, schopnosti a governance

1

ChatGPT — DNS exfiltrace

03/2026

Jediný prompt dokázal tiše vynést zprávy a nahrané soubory ze sandboxu přes rekurzivní DNS; stejnou cestou šlo založit vzdálený shell.

Zdroj: Check Point Research

2

Claude Mythos Preview

04/2026

Model zadržený kvůli schopnosti autonomně nacházet a zneužívat zero-day zranitelnosti; k preview se pak dostala Discord skupina uhádnutím URL.

Zdroj: Carnegie · TechCrunch

3

Google GTIG — první AI zero-day

05/2026

Útočník použil LLM k nalezení a zbraňování 2FA-bypass zero-day; exploit nesl stopy LLM včetně halucinovaného CVSS skóre.

Zdroj: Google Threat Intelligence

Fortinet AI Strategy

FortiAI-Protect

Začlenění AI do detekce hrozeb, analýzy chování a zpravodajství o kybernetických hrozbách přináší LEPŠÍ ZABEZPEČENÍ

FortiAI-Assist

Začlenění AI do detekce hrozeb, analýzy chování a zpravodajství o kybernetických hrozbách přináší LEPŠÍ ZABEZPEČENÍ

FortiAI-Secure AI

Začlenění AI do detekce hrozeb, analýzy chování a zpravodajství o kybernetických hrozbách přináší LEPŠÍ ZABEZPEČENÍ

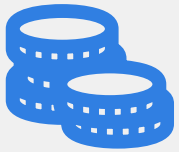
Proč to dneska všecho budujeme

Inovace a automatizace

Inovace a odlišení se od konkurence

Rozhodování na základě kontextu

Customer Experience

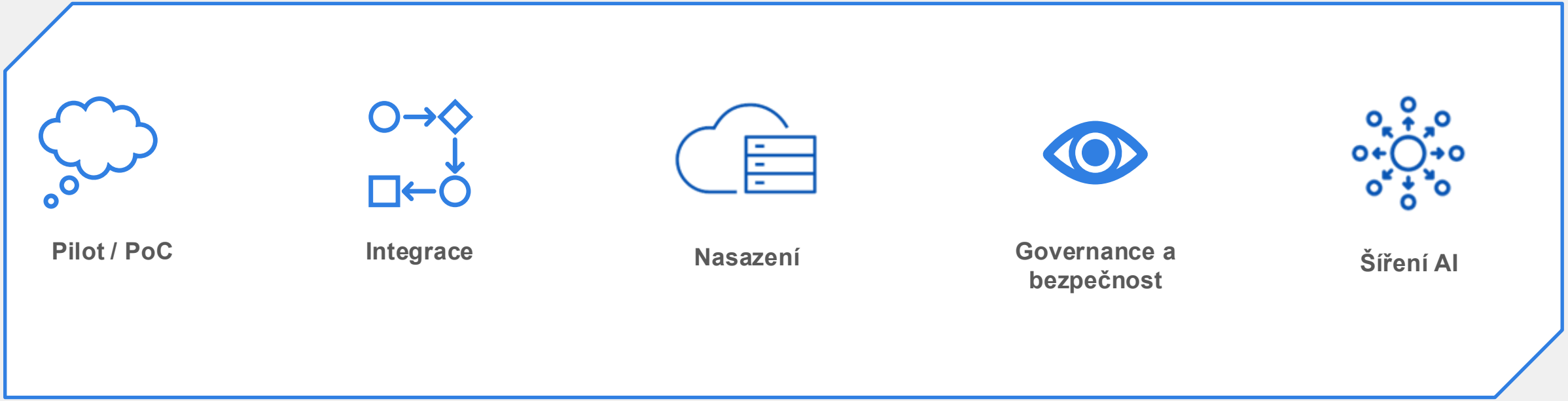


Efektivita

Výhody:

- ✓ Data Sovereignty
- ✓ Brand Voice and Tone
- ✓ Předvídatelnost nákladů a efektivita při škálování
- ✓ Ochrana duševního vlastnictví
- ✓ Zamezení závislosti na dodavateli

Common AI Adoption Process



Organizace obvykle postupují po etapách: začínají v menším měřítku a poté se rozšiřují do dalších obchodních jednotek či iniciativ, přičemž v každém kroku vyhodnocují přínos a náklady.

OWASP TOP10 pro LLM

LLM01 - Prompt Injection - útočník manipuluje vstupem tak, aby přepsal instrukce systému nebo získal nechtěné chování modelu

LLM02 - Sensitive Information Disclosure - model vyzradí důvěrná data (API klíče, osobní údaje, trénovací data) ve výstupu

LLM03 - Supply Chain Vulnerabilities - kompromitované modely, pluginy nebo trénovací datasety třetích stran

LLM04 - Data and Model Poisoning - manipulace trénovacích dat za účelem zadních vrátek nebo zkreslených výstupů

LLM05 - Improper Output Handling - nebezpečné zpracování LLM výstupu downstream (XSS, RCE, SQL injection)

LLM06 - Excessive Agency - LLM má příliš mnoho oprávnění nebo autonomie, vykonává akce s reálným dopadem bez kontroly

LLM07 - System Prompt Leakage - únik systémového promptu odhalí obchodní logiku nebo bezpečnostní mechanism

LLM08 - Vector and Embedding Weaknesses - útoky na RAG vektory (poisoning embeddingů, manipulace retrieval pipeline)

LLM09 - Misinformation - model generuje přesvědčivé, ale nepravdivé výstupy (halucinace) s reálnými důsledky

LLM10 - Unbounded Consumption - nedostatečné limity vedou k DoS, nadměrným nákladům nebo degradaci služby (token flooding)

Co náš čeká a nemine

Bezpečné využívání AI

Zabezpečení AI infrastrukuty

Zabezpečení LLM a AI

Bezpečné využívání AI

Zaměstnanec využívající veřejně AI

Adoption Rate

88%

Firemní uživatelé, kteří přistupují k aplikaci AI odkudkoli

Práce s LLM/AI

Data At Risk - PII, IP, API, Shadow AI

Outbound

Clean Traffic

FortiClient: ZTNA

FortiSASE: Zabezpečený přístup odkudkoliv

FortiAuthenticator: deepfake impersonation, credential stuffing

FortiGate: Discovery and access control

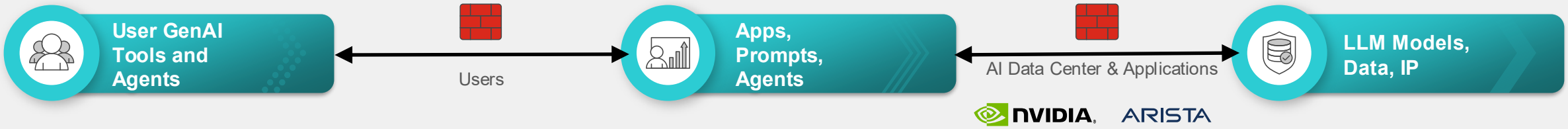
FortiGuard App Control: Shadow AI sites + blokování aplikací

FortiDLP: Ochrana osobních údajů a důvěrných informací a zaznamenávání aktivit

Zabezpečení AI infrastrukuty

Ochrana uživatelů, aplikací a datových center pro umělou inteligenci

Fortinet AI Security through multi-layer protection



FortiAI-Protect IT Controls
Visibility & Control | Secure GenAI with data privacy

FortiAI-Secure AI Attack Surface Management
Agent & LLM Security | Layered AI infrastructure defense

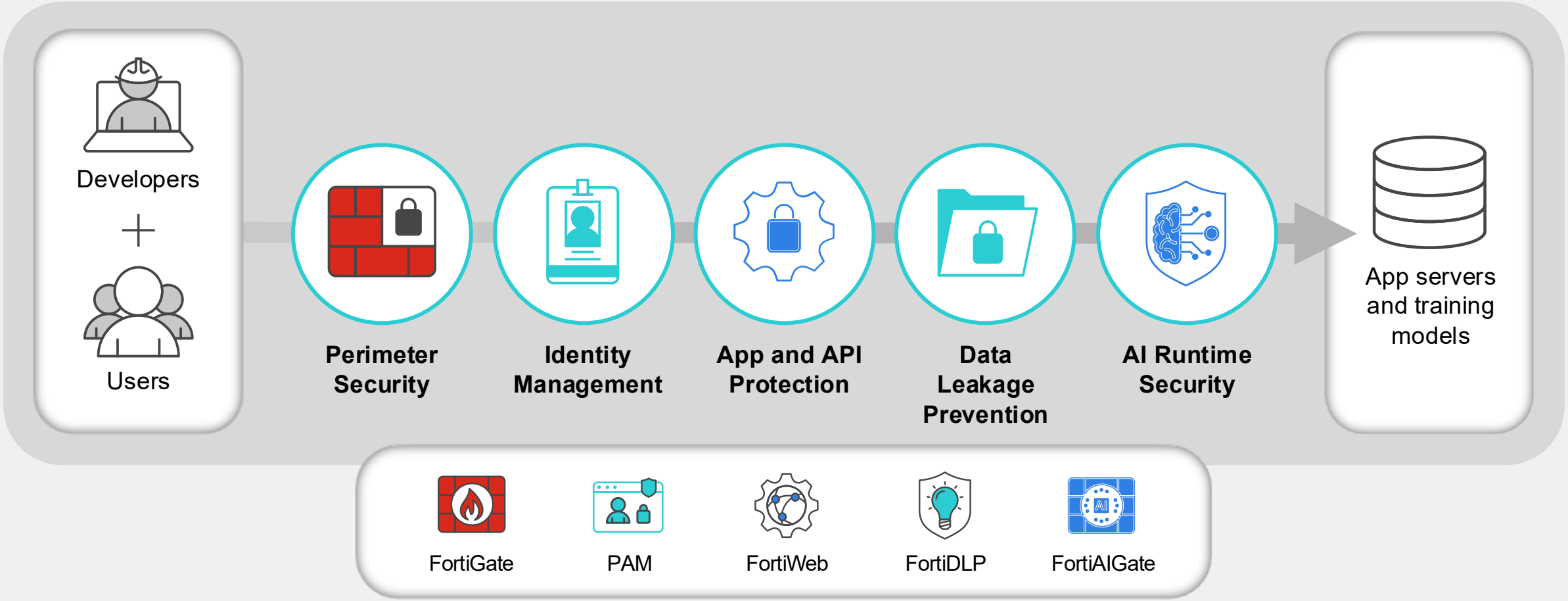
FortiAI-Assist NOC & SOC
Assistants & Agents | Orchestrated SOC/NOC autonomous ops

FortiGuard Labs Threat Intelligence
Real-time, AI/ML-powered | Shared across the Security Fabric

Fortinet Security Fabric Converged Networking and Security
Coordinated Defense | Broad, Integrated and Automated

Komplexní zabezpečení datového centra pro umělou inteligenci

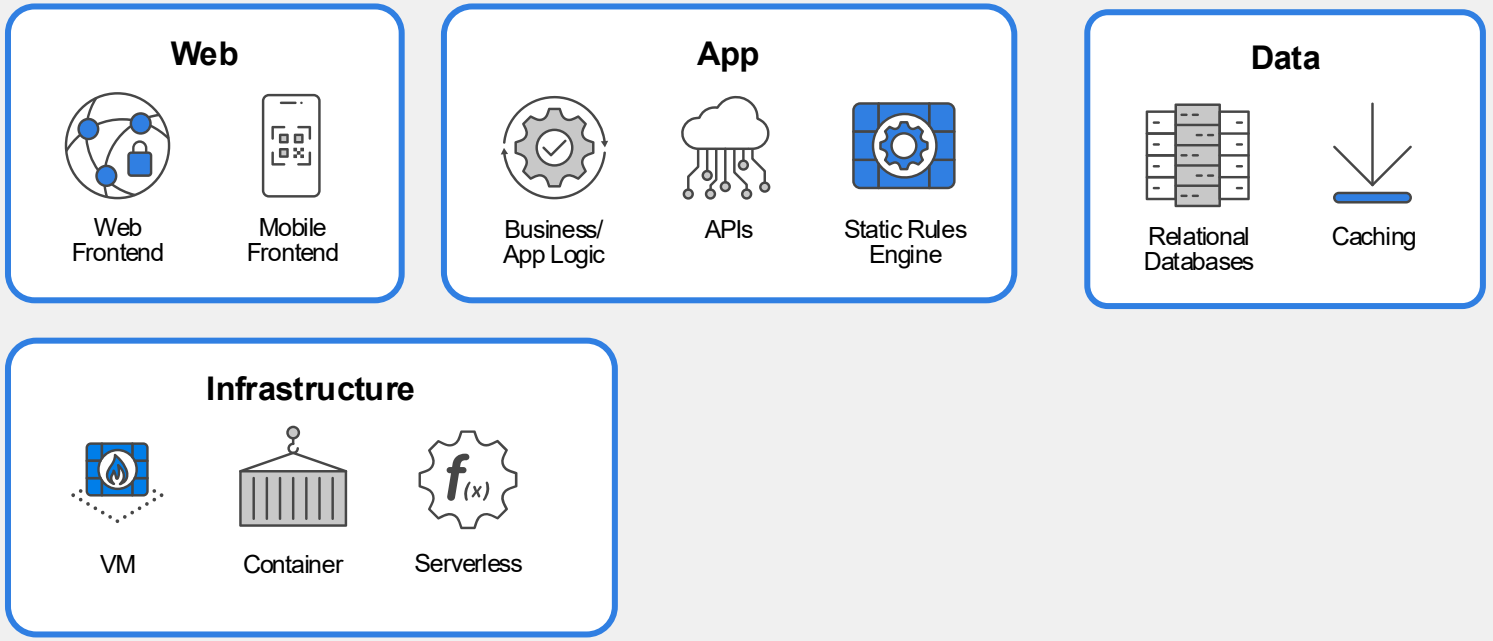
Jak to vidíme my



Zabezpečení LLM a AI

Jak dnes zabepečujeme aplikace

Vychází z monitorování provozu na síťové a aplikační vrstvě, čímž zabraňuje tomu, aby se hrozby dostaly k citlivým zdrojům



Jak je to nutné dělat dnes

Přináší nové hrozby a rozšiřuje plochu útoku na aplikace