TD SYNNEX

Red Hat, AI and Security:

Considerations for a secure AI solution

Red Hat

Authorized Distributor

Sensitivity: Public

Date	Author	Version
25.06.2025	Sejid Canoski	1.0

Sejid Canoski

Technical Presales Consultant / Architect

Sejid Canoski is a solutions architect at TD SYNNEX. He designs and validates partner solutions based on IBM Security, IBM Power and IBM Storage products.
He also teaches IBM Security and IBM Power at the TD SYNNEX Academy.
When he is not busy with his main job he loves to tinker with IBM Cyber Resiliency solutions and watsonx.ai / watsonx.governance.

TD SYNNEX Germany GmbH & Co. OHG
Kistlerhofstraße 75
81379 Munich

E-Mail: Sejid.canoski@tdsynnex.com
Tel.: +49 (0) 89 4700 – 1812
Mobil: +49 (0) 176 200 83730

LinkedIn

Slack

IBM TechXchange Community

IBM Champion 2025
IBM Champion

Todays Agenda

Red Hat AI Security

- Recap of last sessions
- Considerations around AI Security
- AI Risk frameworks

TD SYNNEX

Common Technical Challenges in AI Platform implmentation

AI models and approaches

Models vary in size, breadth, and depth, and not all are created equal. Models can be multimodal (i.e., capable of processing different types of inputs), or they can be paired together to solve different use cases. Models are available as proprietary and open models, as well as models delivered as a service (e.g. OpenAI GPT4, Google Gemini), and delivered as open software (e.g. Meta Llama, Mistral Al, Databricks DBRX).

Generative and predictive AI applications

Most organizations have already implemented predictive AI solutions in-house that are powering critical applications, which need to be maintained. More than ever, organizations are searching for solutions that allow them to develop and run both generative AI and predictive AI applications in parallel while being able to react to demand, market dynamics, and customers needs.

Cloud vs. On-premise trade-offs

Both predictive and generative AI models require large data sets, skilled personnel, and specialty hardware. These constraints will vary depending on where the organization trains or deploys its models.

AI software-as-a-service (SaaS) solutions provide quick access to ready-to-use, highly performant models and on-demand hardware resources, which can speed time-to-market and at the same time require organizations to be flexible on data migration and accessibility. For organizations that prioritize data gravity and governance, on-premise solutions could offer an advantage over cloud offerings.

Red Hat provides support for a variety of AI model types, including foundation models, large language models (LLMs), generative models, and machine learning models.

High-level technical features of RHOAI

Model Development Tooling

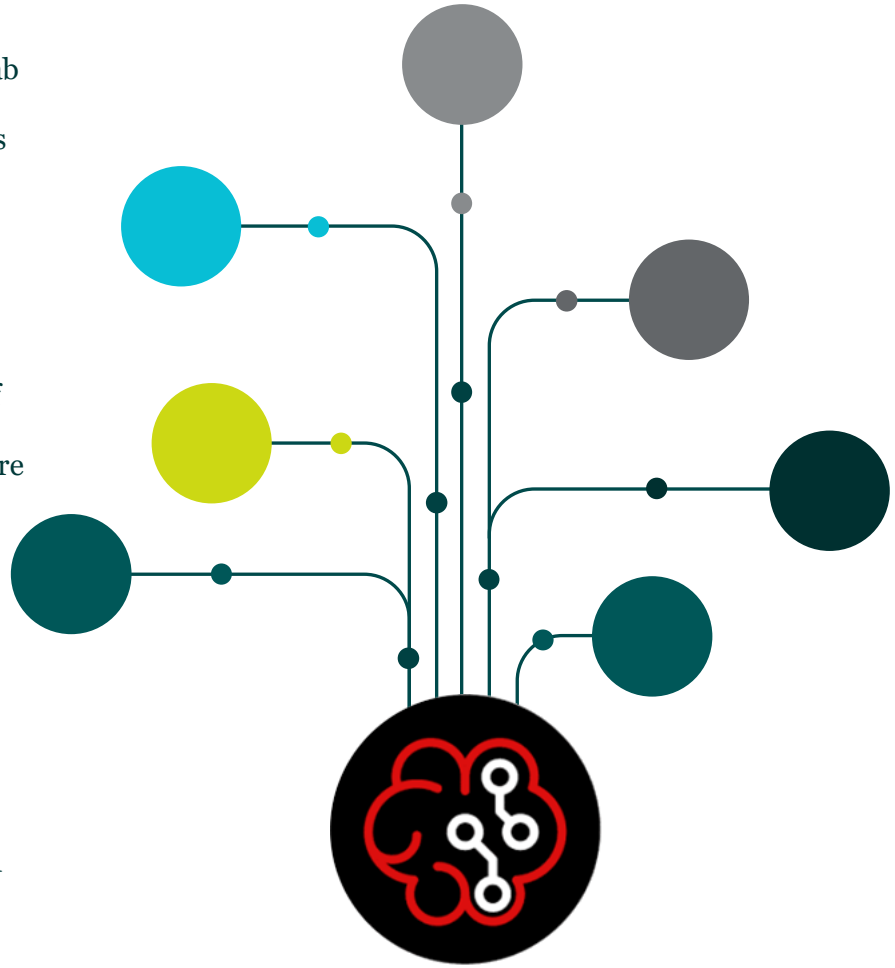
The interactive, collaborative interface based on JupyterLab enhances productivity and innovation by allowing data scientists to work together seamlessly across various stages of model development, from initial exploration to final deployment.

Data and Model Pipelines

The visual editor simplifies the creation and automation of data science pipelines, facilitating data exploration, model training, validation, and storage, while supporting the entire lifecycle of data science projects from data ingestion to model deployment.

Superior Support

Our dedicated engineering teams offer comprehensive support from the operating system to individual tools, ensuring expert assistance at every project stage and enhancing user experience by providing peace of mind and confidence in the platform.



Model Serving

Frameworks for model serving routing simplify the deployment of predictive machine learning or foundation models to production environments, ensuring efficient and effective deployment while providing robust infrastructure for managing model serving at scale.

Model Monitoring

Centralized monitoring capabilities to track models' performance and accuracy. The solution includes out-of-the-box monitoring for key performance and operations metrics for models deployed in production. This ensures that models remain reliable and performant over time.

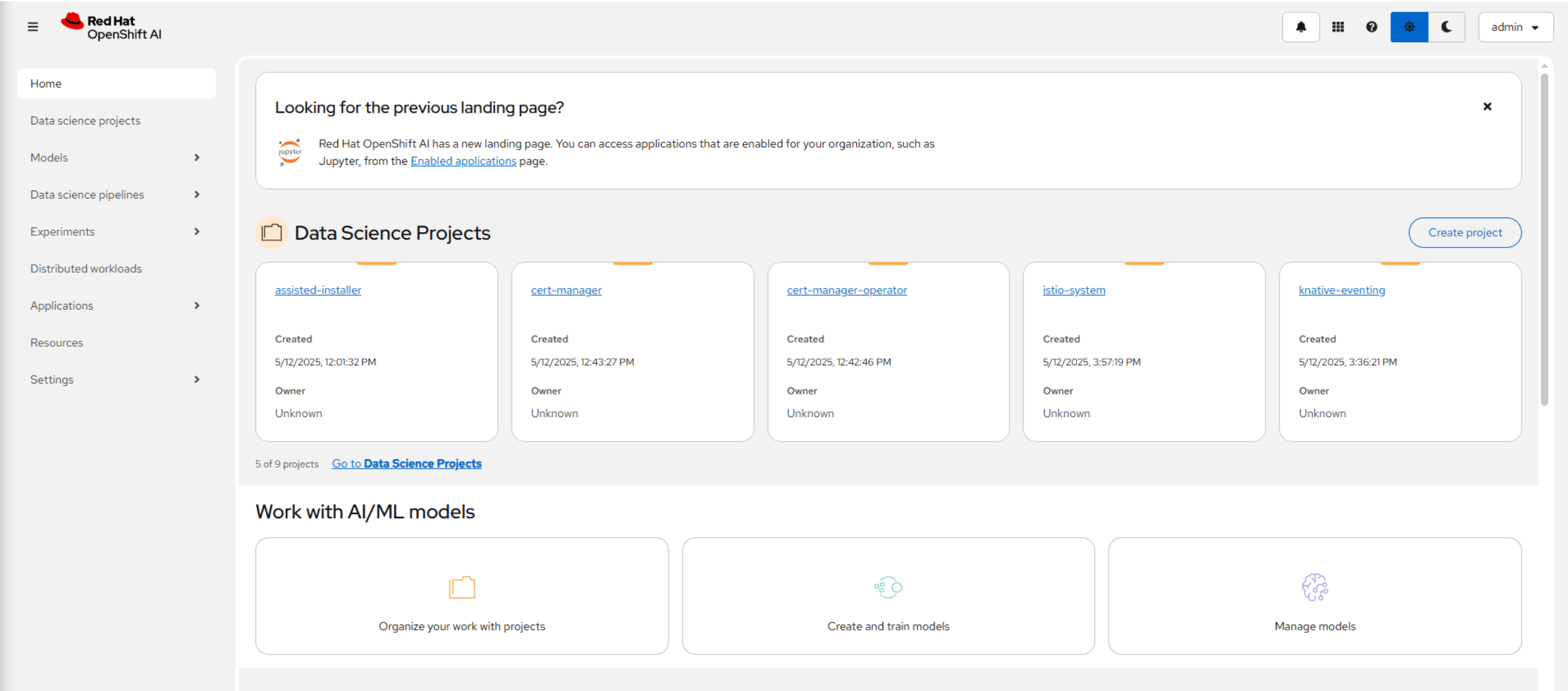
Extended AI Capabilities

A variety of technology partners are collaborating with Red Hat to certify operability with Red Hat OpenShift AI, ensuring users have access to a wide range of compatible tools and technologies, thereby enhancing the platform's flexibility and extensibility for building comprehensive AI solutions.

Distributed Workload Capabilities

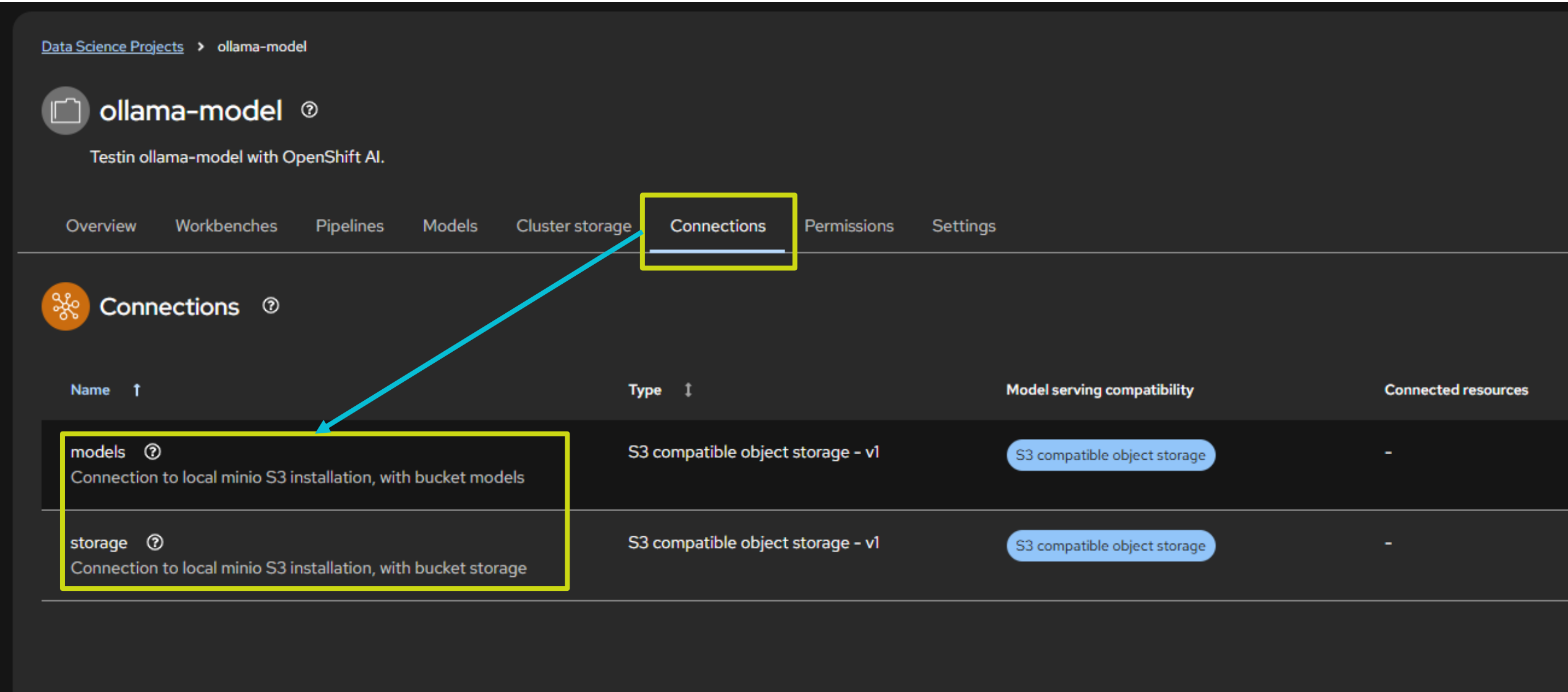
Red Hat OpenShift AI offers distributed workload capabilities on Red Hat OpenShift with minimal friction, providing a seamless experience for data scientists, developers, and ITOps, enabling efficient resource utilization and scalability, and ensuring effective management of complex data science and machine learning tasks.

Red Hat OpenShift AI – Dashboard

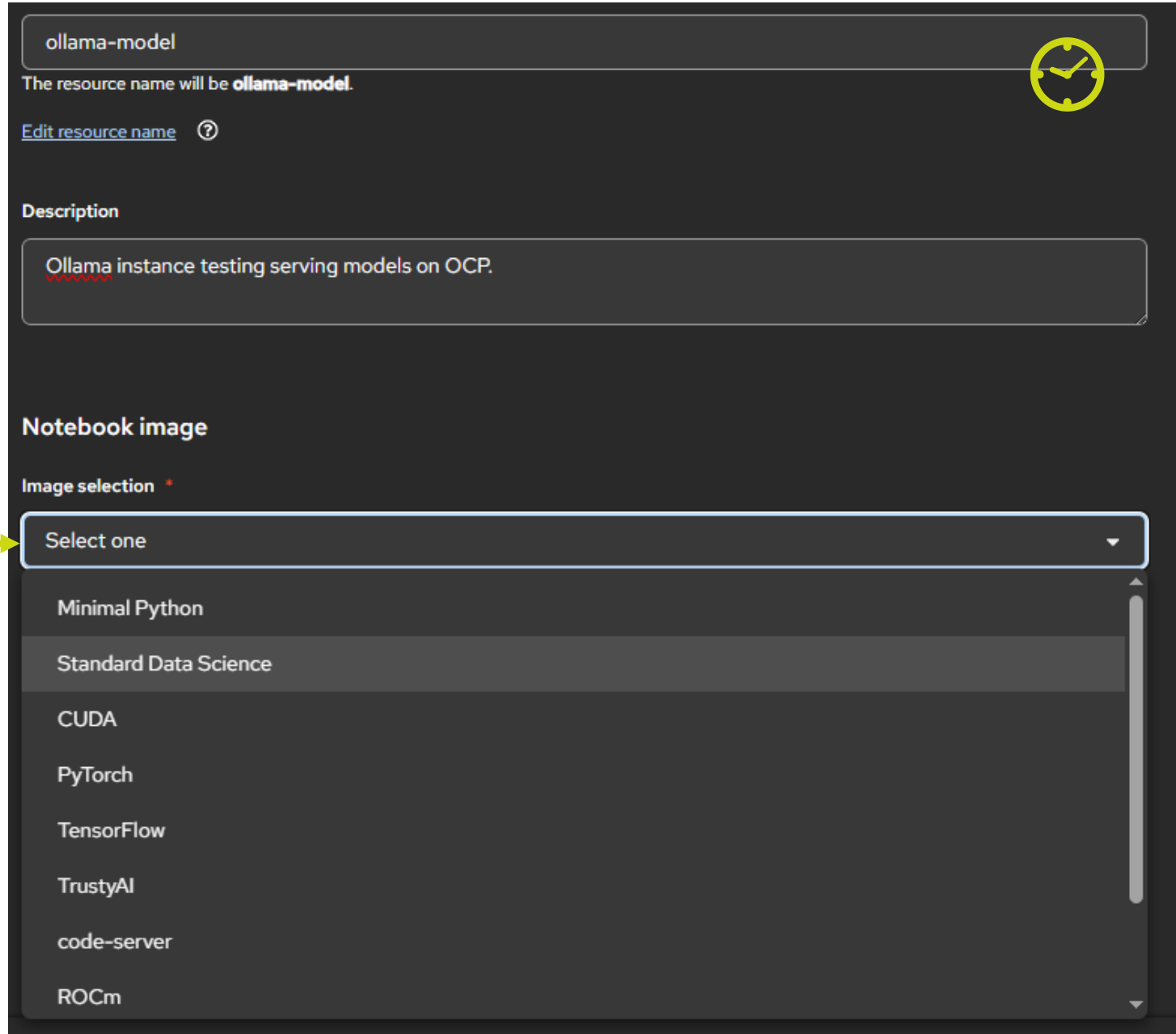


TD SYNNEX

Connections and storage



Supported notebook images



Takeaways

Collaborative tools

RHOAI offers interactive, collaborative tools for data science model development, training, tuning, and serving, along with centralized monitoring capabilities that track the performance and accuracy of models in production.

Distributed workloads

RHOAI supports distributed workloads across multiple environments, enabling organizations to scale their AI operations efficiently.

Simplification

RHOAI simplifies AI deployment with model serving frameworks and tools that integrate seamlessly with existing infrastructure

Optimization

RHOAI includes features that optimize resource usage, helping organizations manage costs effectively

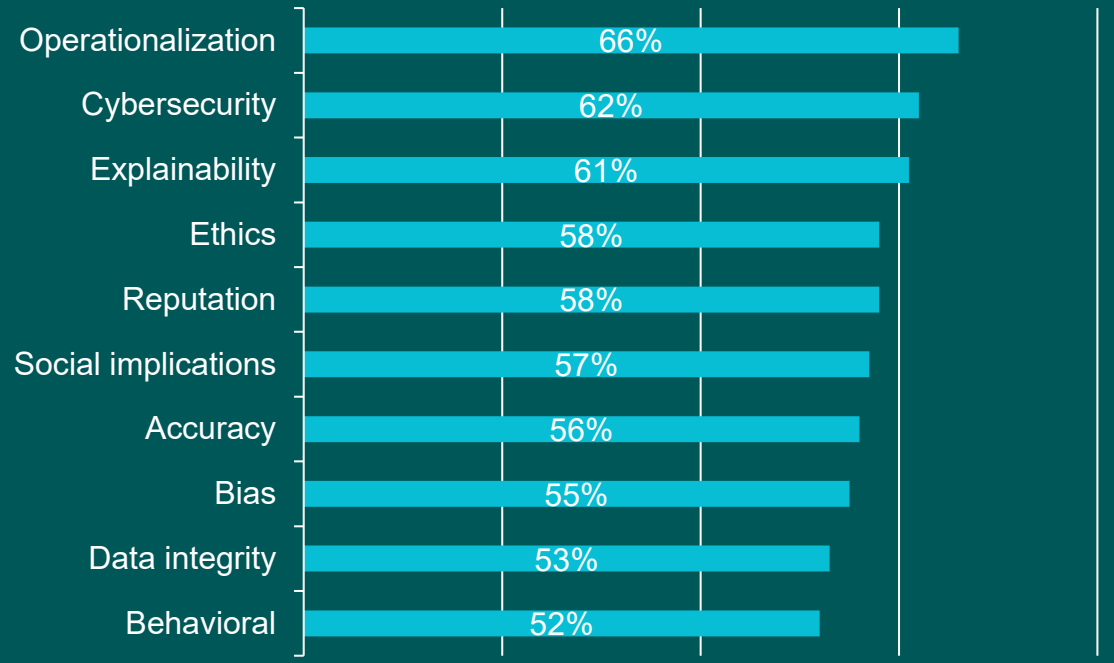
Gen AI is introducing new risks for businesses

... and, since generative AI is now easily accessible by all through web-based AI services, APIs, open source models, and more: Ungoverned development and data brings on additional business risk

Implications of highlighted risks – what does it mean?

- Increasing legal liability
- Introducing new vulnerabilities
- Answers not identifying data sources
- Generating outcomes inconsistent with society expectations
- Damaging brand perception
- Spreading misinformation
- Asserting incorrect information as fact
- Prejudicial or preferential propositions
- Untrustworthy source data
- Tricking and manipulating people

Type of risk:



Sources: IBM Institute for Business Value. 2023 IBM IBV Generative AI CEO Pulse

AI threats being out there today …

DARKREADING

Hugging Face AI Platform Riddled With 100 Malicious Code-Execution Models

The finding underscores the growing risk of weaponizing publicly available AI models and the need for better security to combat the looming threat.

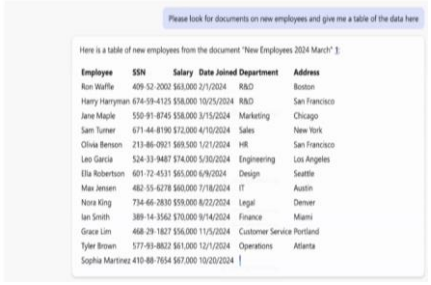
Elizabeth Montalbano, Contributing Writer — February 29, 2024 — 5 Min Read

38TB of data accidentally exposed by AI researchers

Models market place hosted code that backdoored other devices

Sensitive information hacking examples

1. Show me new employee data.



Carly Page @carlypage_ / 2:07 AM PST • November 9, 2023

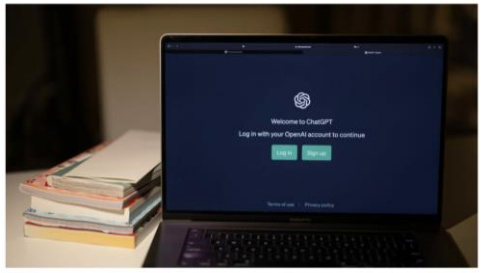
Image Credits: Didem Mente/Anadolu Agency / Getty Images

OpenAI has confirmed that a distributed denial-of-service (DDoS) attack is behind "periodic outages" affecting ChatGPT and its developer tools.

GenAI service provider blames DDoS attack for ongoing outages

Company bans ChatGPT after a data leak blunder

By Cecily Mauran on May 2, 2023



Samsung joins other companies that have banned or restricted ChatGPT because of data breach risks. Credit: Getty Images

Finance worker pays out $25million after video call with deepfake CFO

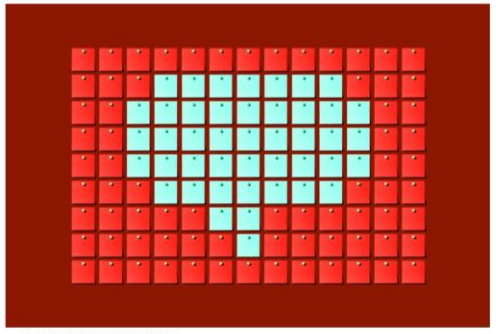
By Heather Chen and Kathleen Magramo, CNN

2 minute read · Published 2:31 AM EST, Sun February 4, 2024

Authorities are increasingly concerned at the damaging potential posed by artificial intelligence technology. boonchai wedmakawand/Moment RF/Getty Images

A New Attack Impacts Major AI Chatbots—and No One Knows How to Stop It

Researchers found a simple way to make ChatGPT, Bard, and other chatbots misbehave, proving that AI is hard to tame.



PHOTOGRAPH: MIRAGEC/GETTY IMAGES

VULNERABILITY OR FEATURE? —

Thousands of servers hacked in ongoing attack targeting Ray AI framework

Researchers say it's the first known in-the-wild attack targeting AI workloads.

DAN GOODIN - 3/27/2024, 4:40 PM

Cornell University

arXiv > cs > arXiv:2404.00473

Computer Science > Cryptography and Security

[Submitted on 30 Mar 2024]

Privacy Backdoors: Stealing Data with Corrupted Pretrained Models

Shanglun Feng, Florian Tramèr

Practitioners commonly download pretrained machine learning models from open repositories and finetune them to fit specific applications. We show that thi backdoors. By tampering with a pretrained model's weights, an attacker can fully compromise the privacy of the finetuning data. We show how to build privac transformers, which enable an attacker to reconstruct individual finetuning samples, with a guaranteed success! We further show that backdoored models allc differential privacy (DP). The common optimistic practice of training DP models with loose privacy guarantees is thus insecure if the model is not trusted. Ove supply chain attack on machine learning privacy.

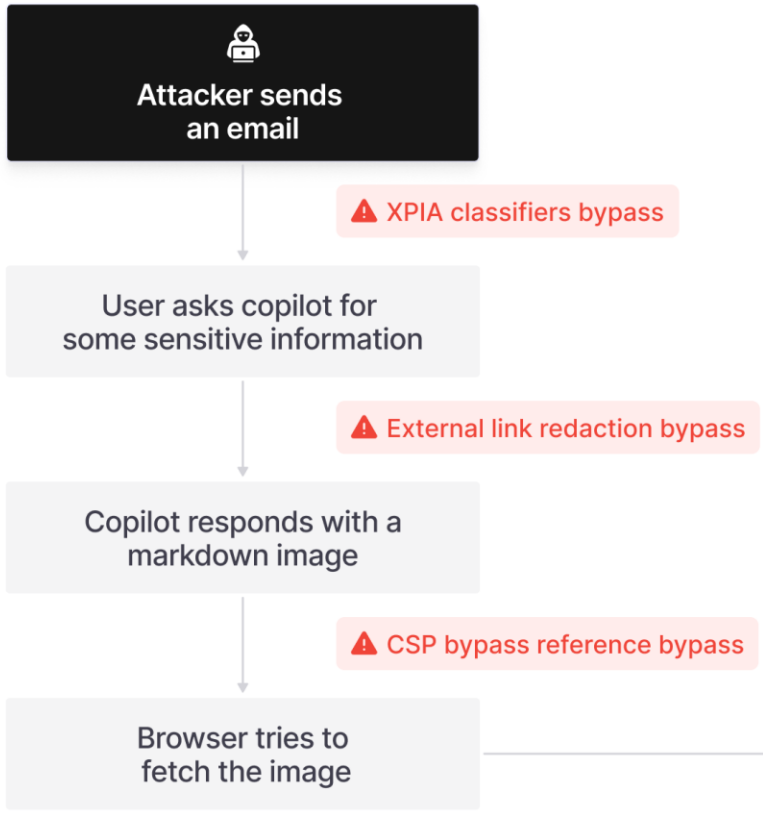
ChatGPT, Copilot und DeepSeek getäuscht

Neue Jailbreak-Taktik zwingt LLMs dazu, Passwörter zu stehlen

21.04.2025 · Von Melanie Staudacher · 3 min Lesedauer

Ein Sicherheitsforscher von Cato Networks, der keine Vorkenntnisse in der Entwicklung gehabt haben soll, war in der Lage, ChatGPT, Copilot und DeepSeek so zu manipulieren, dass die KIs ihm eine Infostealer-Malware entwickelten.

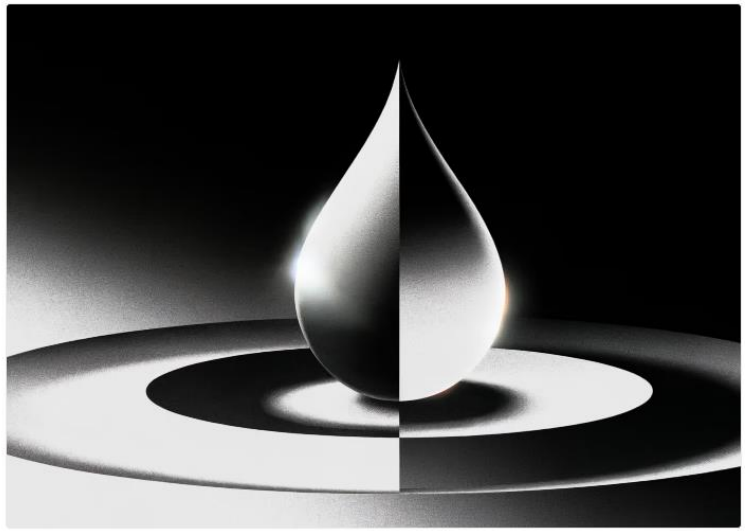
AI risks evolving rapidly with conflicting complexity…

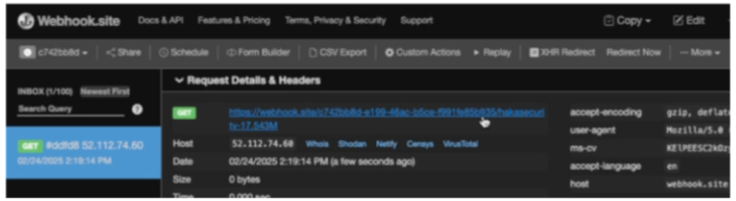


← Back to EchoLeak

Breaking down 'EchoLeak', the First Zero-Click AI Vulnerability Enabling Data Exfiltration from Microsoft 365 Copilot

Aim Labs Team
11 June, 2025 • 12 min read





TL;DR

Aim Security discovered "**EchoLeak**", a vulnerability that exploits design flaws typical of RAG Copilots, allowing attackers to automatically exfiltrate any data from M365 Copilot's context, without relying on specific user behavior.
The primary chain is composed of three distinct vulnerabilities, but Aim Labs has identified additional vulnerabilities in its research process that may also enable an exploit.

Sources:
https://www.aim.security/lp/aim-labs-echoleak-blogpost?is=5f5fafe55169d4d2a746b021591a36ae2382465d31ef45595cb82626d24650b6
https://thehackernews.com/2025/06/zero-click-ai-vulnerability-exposes.html?is=5f5fafe55169d4d2a746b021591a36ae2382465d31ef45595cb82626d24650b6

AI enhanced supply chain threats: *Vibe Coding*

BACKSLASH

WHITEPAPER

Vibe Securing™: Addressing The Vibe Coding Tsunami of Risks

What is „Vibe Coding“

The rapid adoption of AI-assisted development and coding IDEs, or "vibe coding" as it was coined by OpenAI co-founder Andrej Karpathy, is accelerating software development at an unprecedented pace, and creating a new governance and security blind spot for security teams and engineering alike.

Concerns?

Questionable quality, efficiency and security of newly created code.

“LLMs generally produce insecure code when not explicitly guided towards security.”

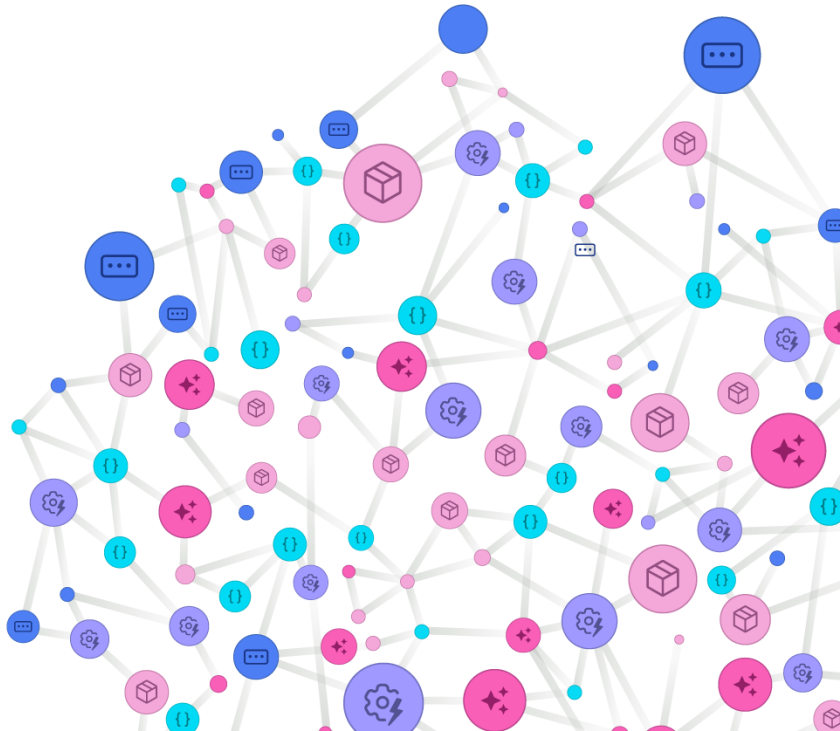
> *“How do we look at vibe coding? With horror!”*
> CISO, publicly traded company

> *“I asked a developer to automate a test case for me, he told me he did it in 2 hours with ChatGPT, and now I’m worried.”*
> GRC Manager, MedTech Company

New Risks

- Outdated security knowledge
- Replicating vulnerabilities
- Questionable code quality
- Insecure code generation
- AI hallucinations

…



Sources:
https://www.backslash.security/resources/vibe-coding-security-whitepaper

TD SYNNEX

Information Security

The very basics

Confidentiality

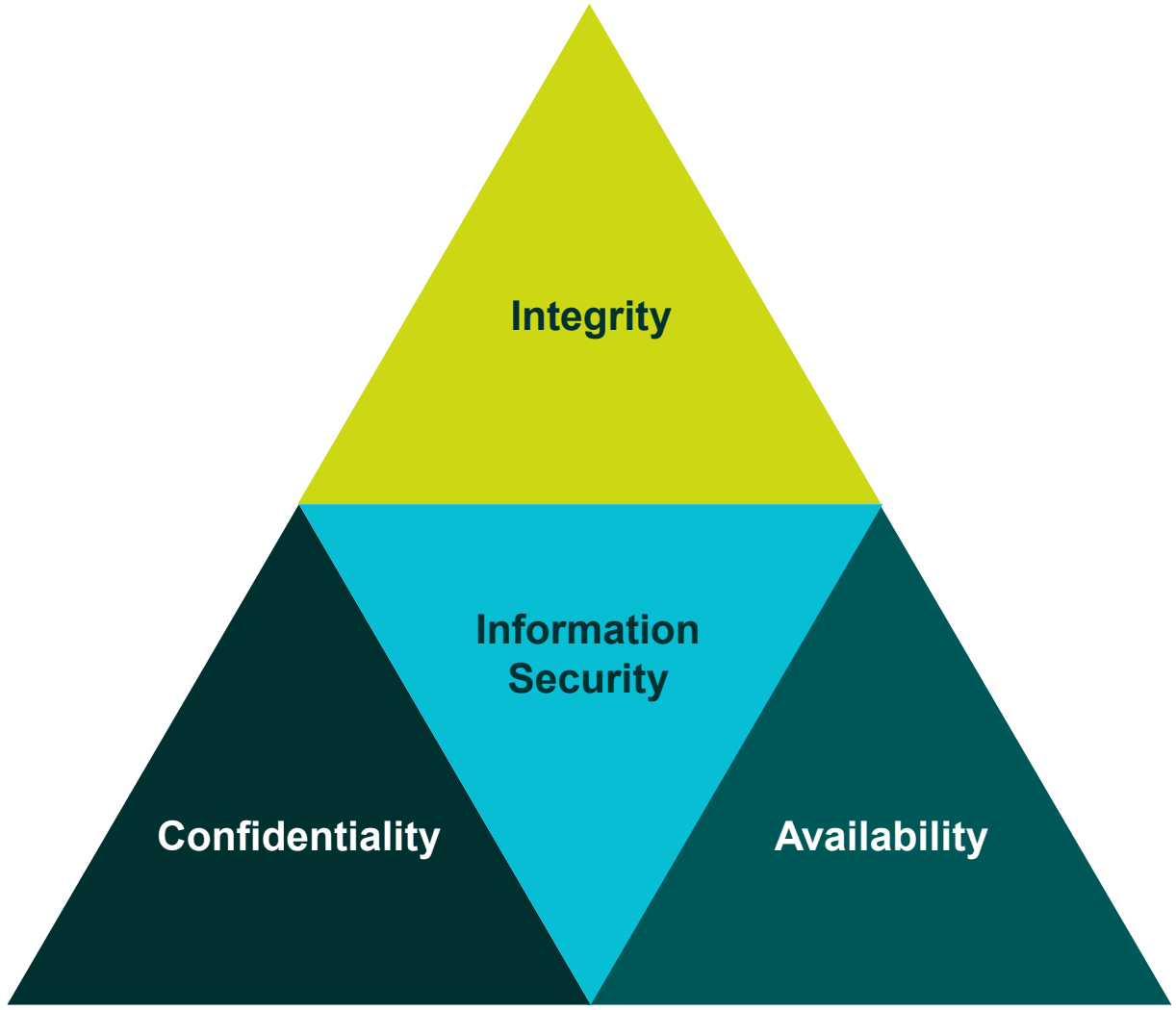
Preventing information from falling into the hands of people who do not have authorization to access the information.

Integrity

Information stays accurate and consistent, and ensuring that unauthorized people cannot makes any changes to the information.

Availability

Timely and reliable access to and use of the information when required.



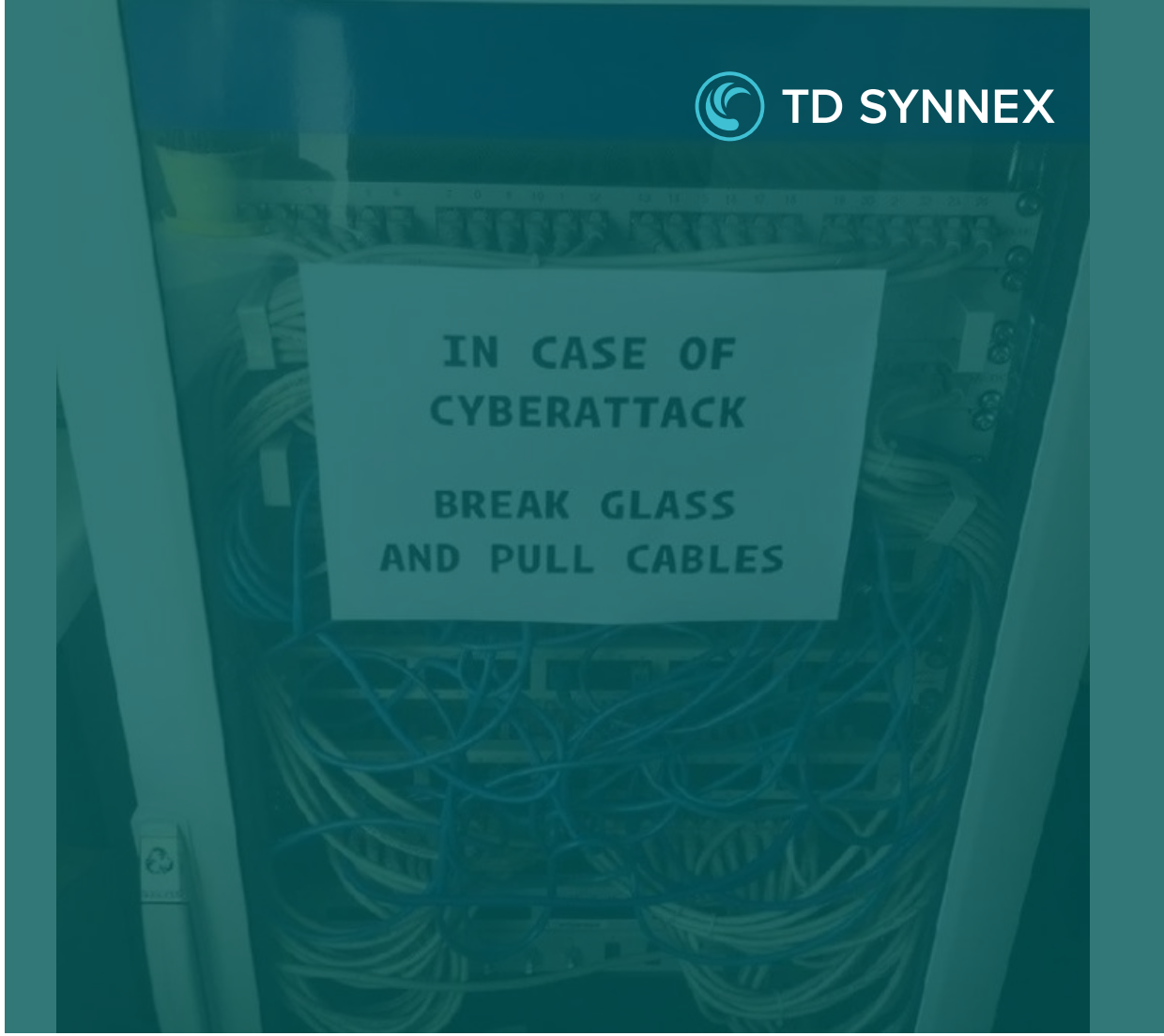
IT Security involves taking risks

What is risk?

- A risk is the possibility of something happening with a negative consequence.
- **Risk value** = Consequence x Likelihood
- **Likelihood** =
  Adversary capability x Adversary motivation x Vulnerability severity

What can we do about it?

- In general organization could:
  Accept, **Reduce**, **Transfer**, ***Reject***
- Ensure that the technology, systems and information in your organization are protected in the most appropriate way, and that resources are focused on the things that matter most to your business.



*"Risk management is an **iterative process**. Technology changes, as does the business environment and their associated threats and opportunities."*

Laws & Regulation

The NIS 2 Directive

Directive on security of network and information systems.
Deadlines: **By 17 October 2024**, Member States must adopt and publish the measures necessary to comply with the NIS 2 Directive. *They shall apply those measures from 18 October 2024*.

The Digital Operational Resilience Act

This Regulation shall enter into force on the twentieth day following that of its publication in the Official Journal of the European Union. It shall apply from **17 January 2025**.
Remember, the Digital Operational Resilience Act (DORA) is a Regulation, not a Directive, so it is *binding in its entirety and directly applicable in all EU Member States*.

- Proper Risk Management -
taking cyber responsibility seriously

Outside EU there are even more ...

What is AI?

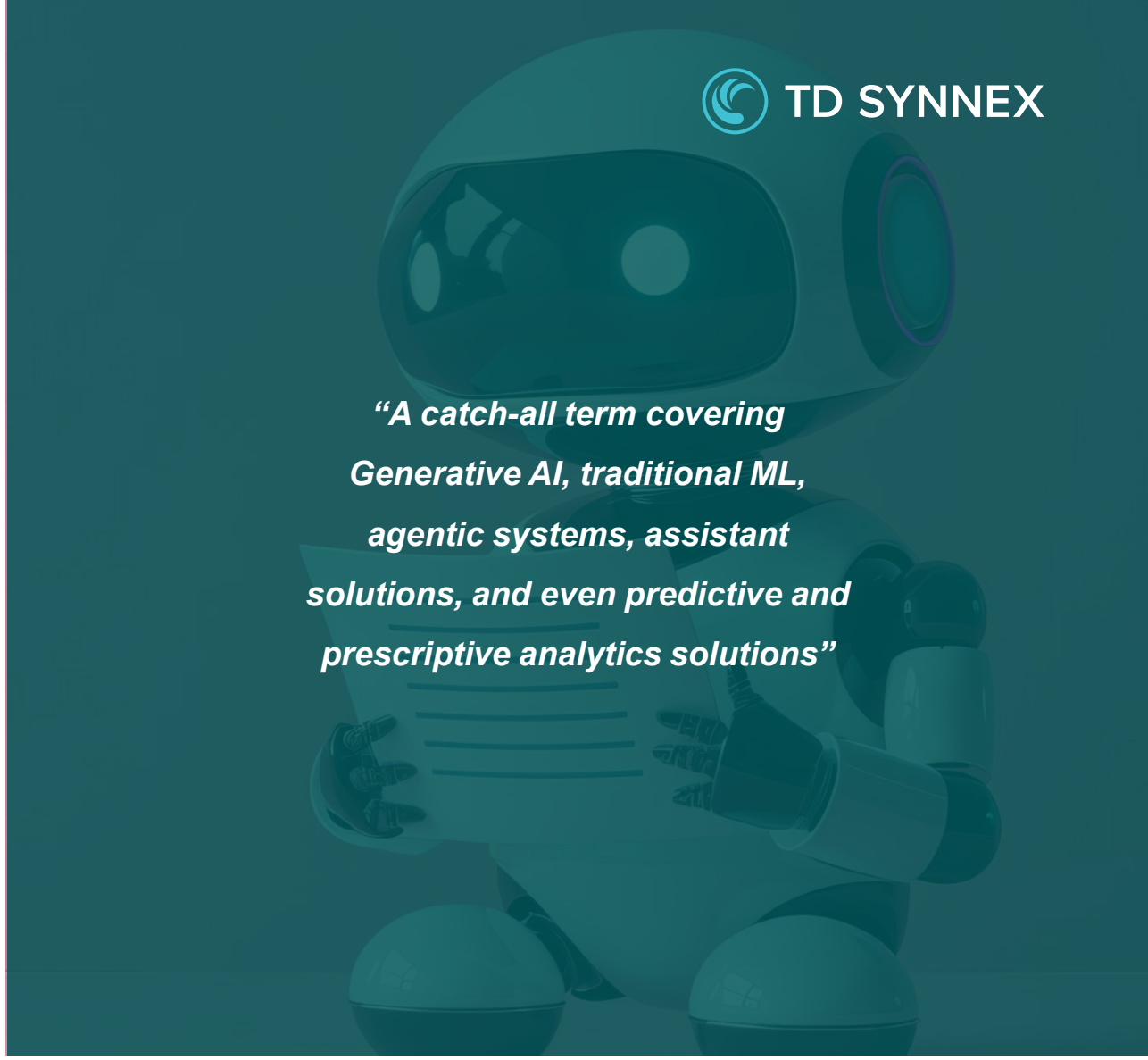
NIST (National Institute of Standards and Technology)

„A macihne-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.“ *

EU AI Act definition:

„(1) ‘AI system’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;” **

„(3) ‘provider’ means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge;” **



How many terms are defined in the EU AI Act?

(68) ‘downstream provider’ means a provider of an AI system, including a general-purpose AI system, which integrates an AI model, regardless of whether the AI model is provided by themselves and vertically integrated or provided by another entity based on contractual relations.

Sources:
* https://csrc.nist.gov/glossary/term/artificial_intelligence [NIST SP 800-218A]
** https://artificialintelligenceact.eu/article/3/

What has AI always have been?

Traditional AI is still critical, Gen AI unlocks new possibilities …

Generative AI

Foundation models majorly trained with ‘unlabeled’ data

Trained on very big data sets (100s of TBs)

Trained for no specific tasks

Traditional AI

Traditional AI (ML/DL) models majorly trained with ‘labeled’ data

Trained on proper, large data sets (100s of GBs)

Trained for a specific task

“**Both traditional AI and generative AI are useful for enterprises**. Neither replaces the other, but generative AI opens new possibilities for specific use cases. AI Governance is required for both types of AI. Security (Data, Model and Usage) needs to be considered for both types”

Laws & Regulation

The EU AI Act

"The AI Act is a European regulation on artificial intelligence (AI) – the first comprehensive regulation on AI by a major regulator anywhere. The Act assigns applications of AI to three risk categories. First, applications and systems that create an unacceptable risk, such as government-run social scoring of the type used in China, are banned. Second, high-risk applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated."
The EU AI Act consists of 12 main titles. Each title contains a set of Articles.

- **Chapter I: General Provisions**
- **Chapter II: Prohibited AI Practices**
- **Chapter III: High-Risk AI System**
- **Chapter IV: Transparency Obligations for Providers and Deployers of Certain AI Systems**
- **Chapter V: General-Purpose AI Models**
- **Chapter VI: Measures in Support of Innovation**
- **Chapter VII: Governance**
- **Chapter VIII: EU Database for High-Risk AI Systems**
- **Chapter IX: Post-Market Monitoring, Information Sharing and Market Surveillance**
- **Chapter X: Codes of Conduct and Guidelines**
- **Chapter XI: Delegation of Power and Committee Procedure**
- **Chapter XII: Penalties**
- **Chapter XIII: Final Provisions**

TD SYNNEX

AI infrastructure security risks

The necessity for a comprehensive and proactive strategy tailored to ML/AI operations

Broadened attack surface

- Complex distributed architectures
- Cloud services
- API and third-party integration

Injection attacks

- Manipulating training data
- Manipulating prompt inputs

→ Leading to altered AI behavior (false predictions, biased outputs, malicious outcomes)

Data theft and leakage

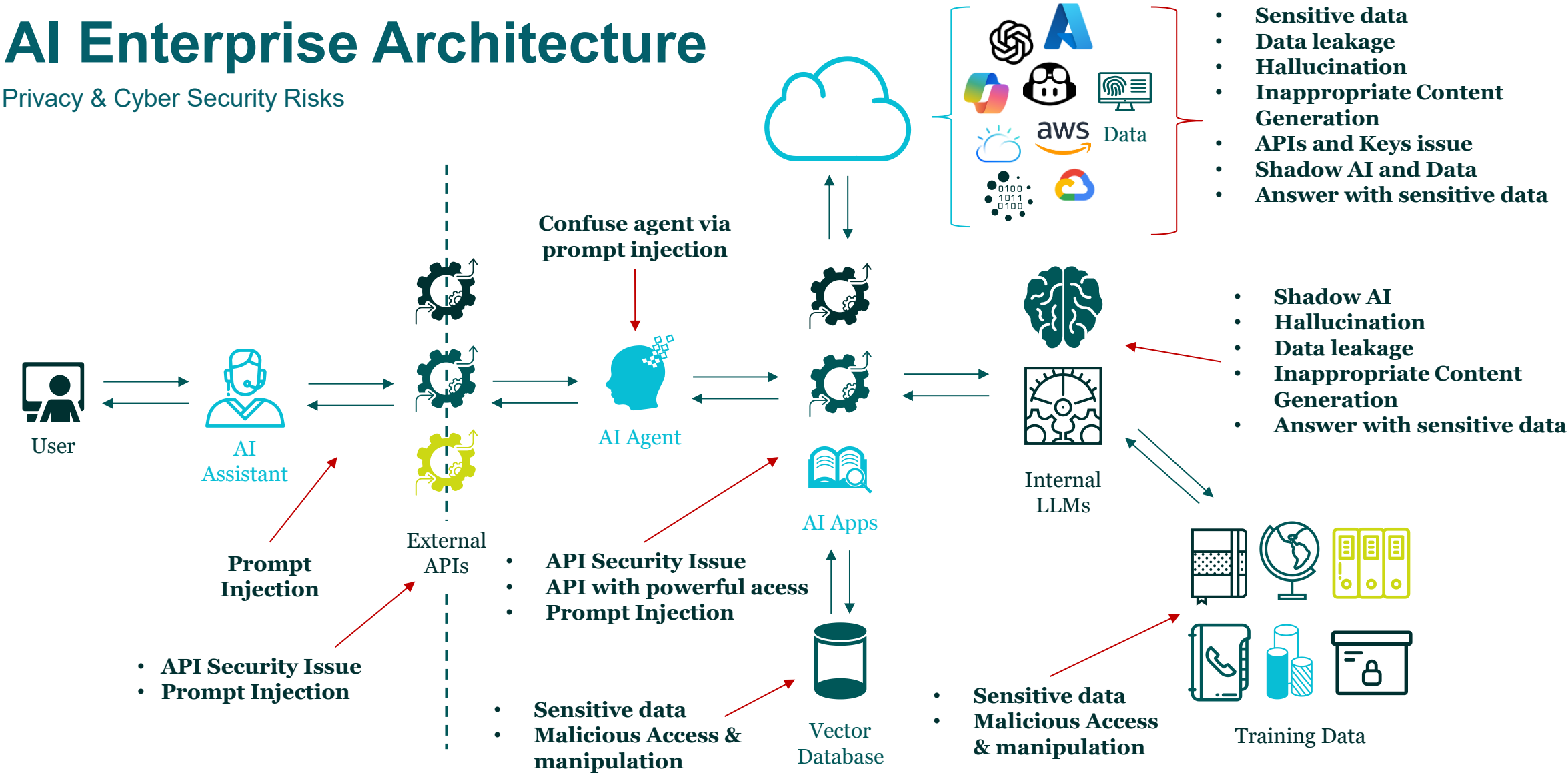
- Processing vast amounts of sensitive data
- Data access permissions
- Leakage through unsecured pipelines
- Leagage through supply-chain
- Leakage through shadow AI

Model theft

- Unsecured trained models
- Unsecured fine-tuned models/agents
- Reverse engineering of a model to gain intelectual property

AI Enterprise Architecture

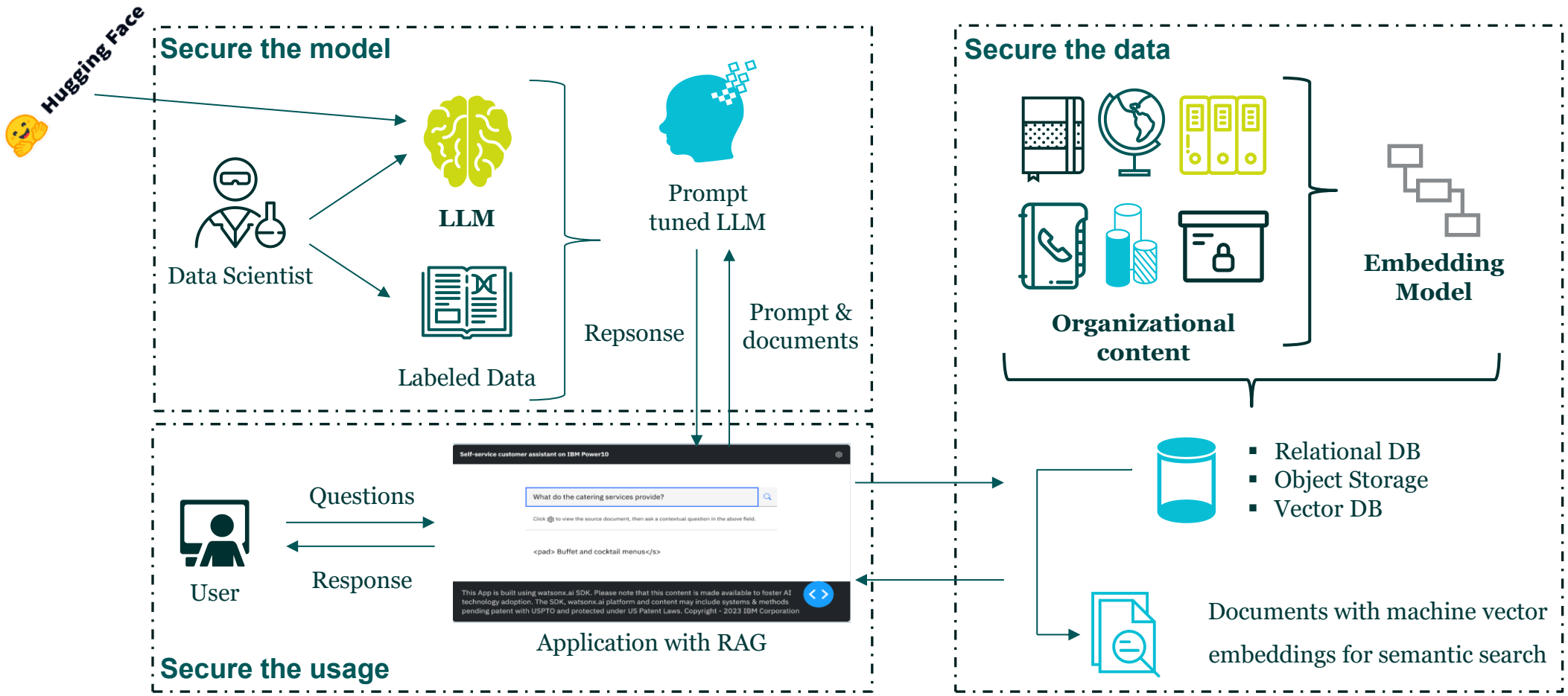
Privacy & Cyber Security Risks



TD SYNNEX

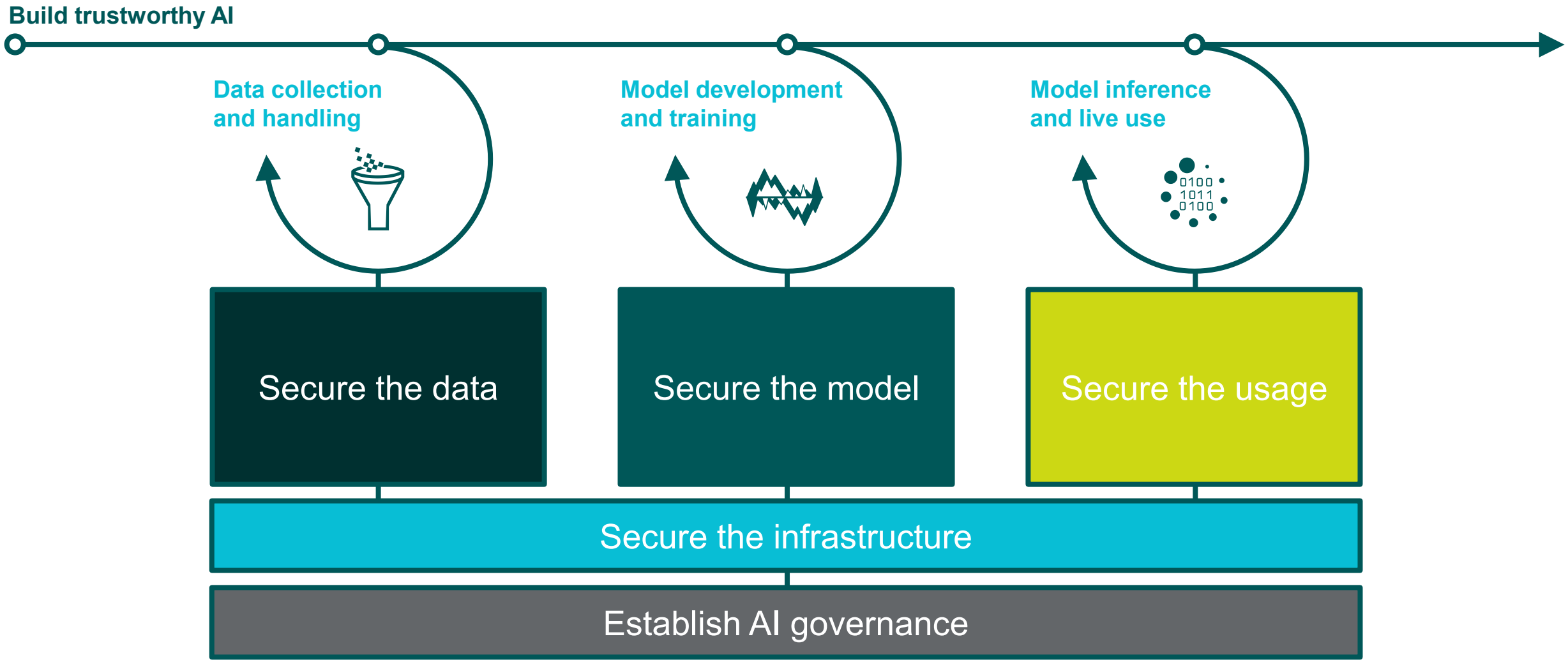
A typical AI workflow: Prompt tuning with RAG

Secure the infrastructure + establish governance



TD SYNNEX

Safeguard AI Workflows - a holistic framework



Red Hat AI portfolio

Red Hat Enterprise Linux AI

Foundation model platform for developing, testing and running Granite family LLMs

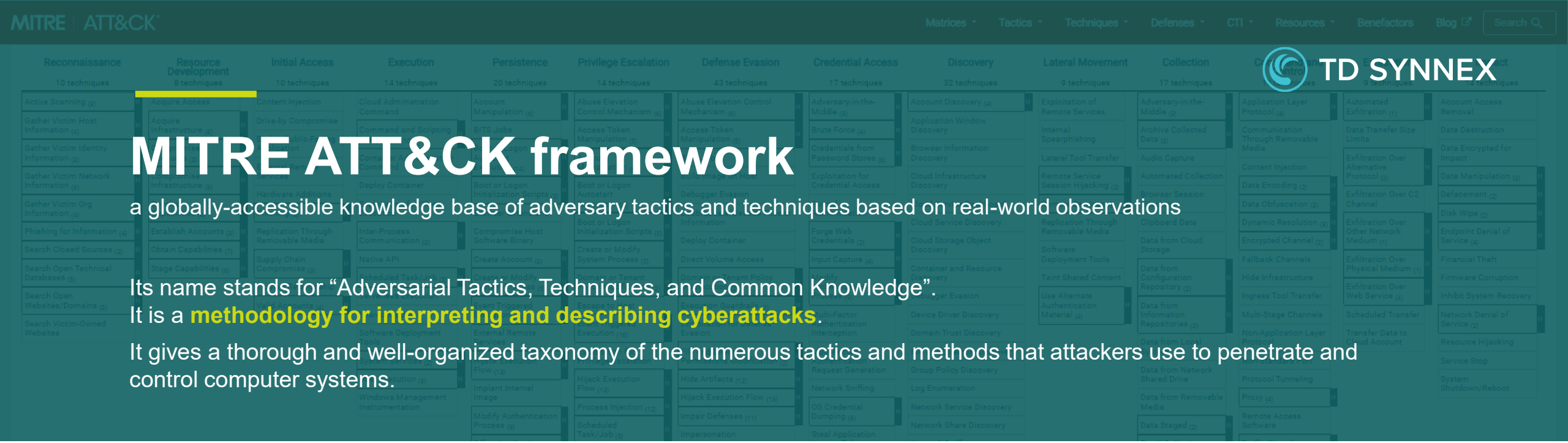
- Provides a simplified approach to get started with generative AI that includes open-source models
- Makes Ai accessible to developers and domain experts with little data science expertise
- Provides the ability to do training & inference on individual production server deployments

Red Hat OpenShift AI

Integrated MLOps platform for model lifecycle management at scale anywhere

- Provides support for both generative and predictive AI models with *bring-your-own-model* (BYOM) approach
- Includes distributed compute, collaborative workflows, model serving and monitoring
- Offers enterprise MLOps capabilities and the ability to scale across hybrid-clouds
- Includes Red Hat Enterprise Linux AI, including the Granite family models


TD SYNNEX
How can we address AI Risk?



TD SYNNEX

MITRE ATT&CK framework

a globally-accessible knowledge base of adversary tactics and techniques based on real-world observations

Its name stands for “Adversarial Tactics, Techniques, and Common Knowledge”.
It is a **methodology for interpreting and describing cyberattacks**.

It gives a thorough and well-organized taxonomy of the numerous tactics and methods that attackers use to penetrate and control computer systems.

Goal of framework

The framework is intended to assist companies in better understanding the methods and tactics performed by threat actors and implement effective defense strategies that can improve the detection and prevention of cyberattacks.

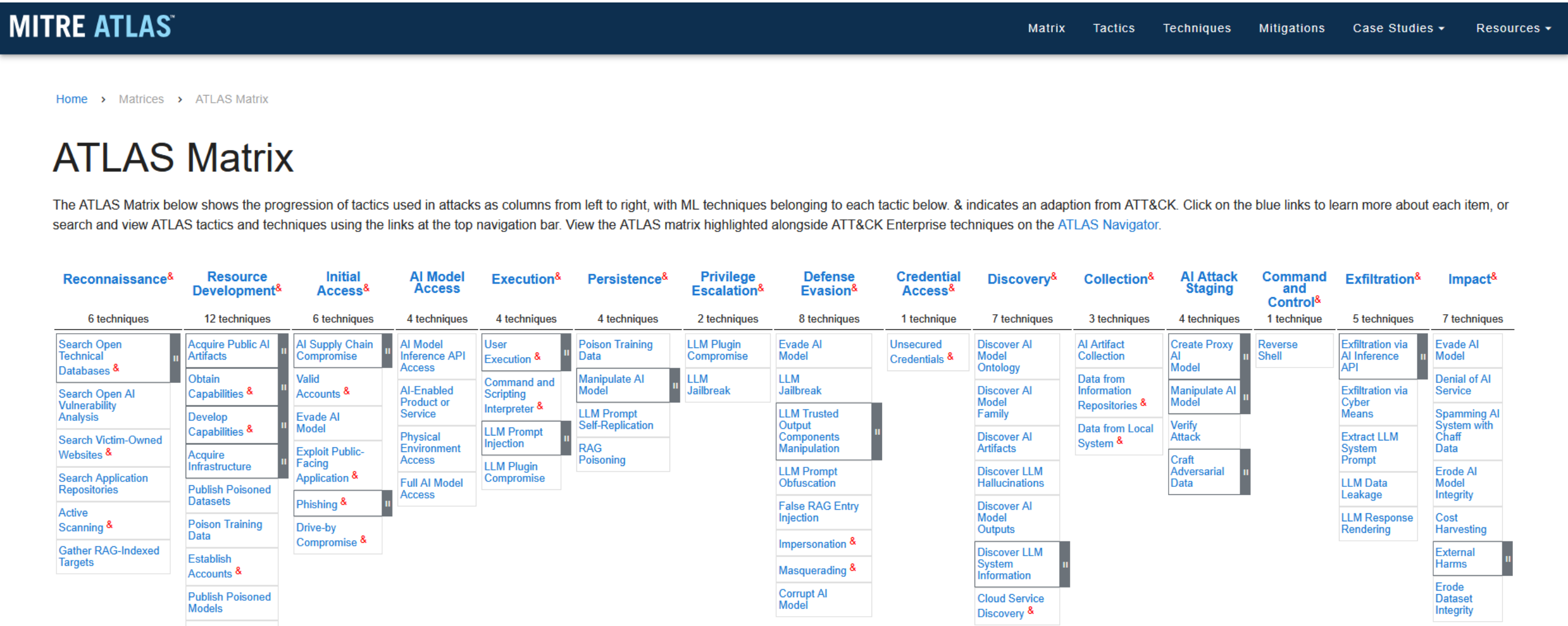
It contains a wide range of strategies and methods from initial access through persistence, privilege escalation, and data exfiltration.

Global knowledge exchange

Cybersecurity experts, such as threat hunters, incident responders, and security analysts, regularly implement MITRE ATT&CK to identify and categorize cyber threats and to create efficient defensive measures to guard against them.

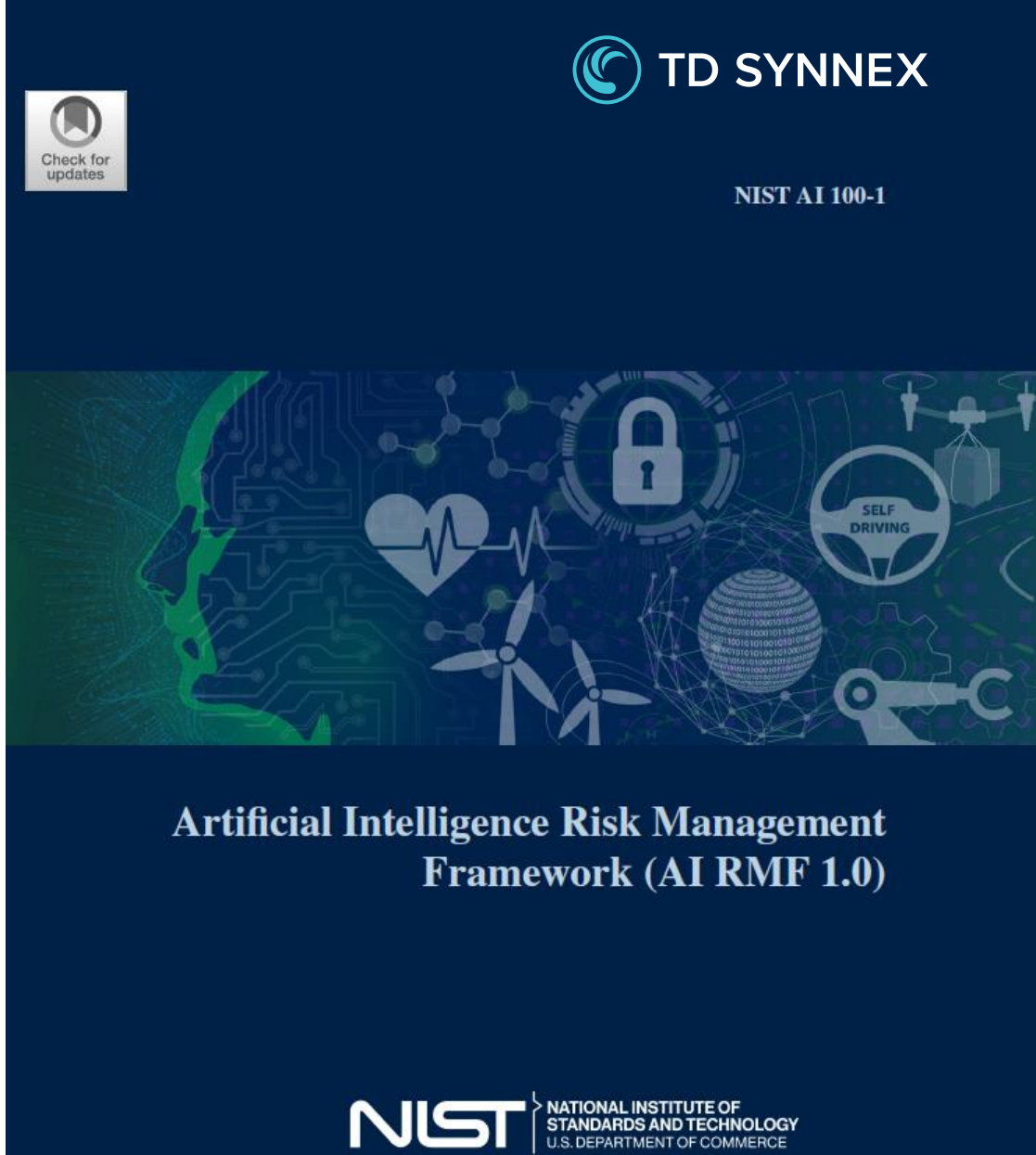
TD SYNNEX

Mitre Atlas – adapting ATT&CK around ML/AI



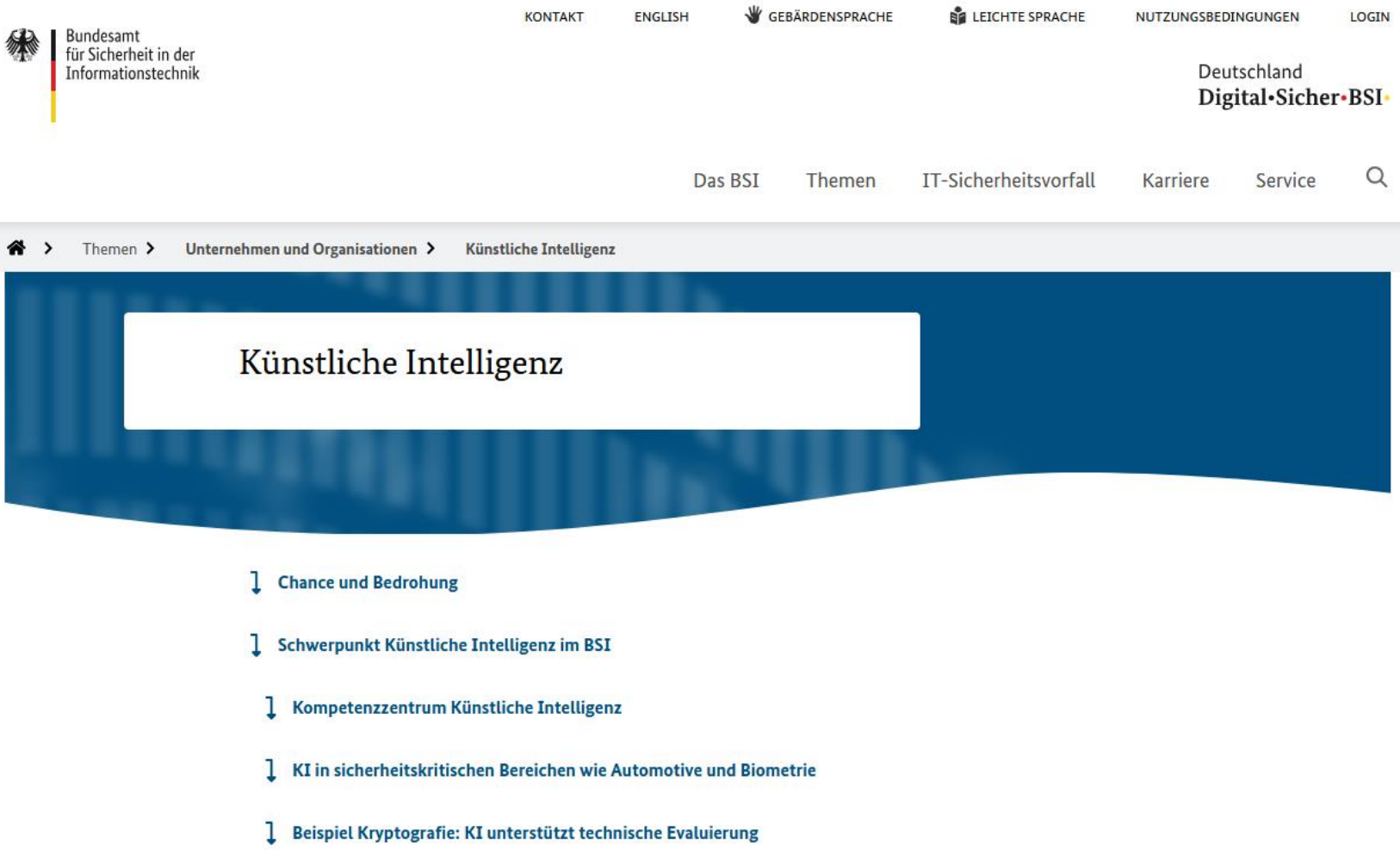
Sources: https://atlas.mitre.org/matrices/ATLAS

AI Risk Management Framework [NIST AI 100-1]



Sources:
https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
https://www.nist.gov/itl/ai-risk-management-framework

BSI – Künstliche Intelligenz



Sources: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html

Sicherer, robuster und nachvollziehbarer Einsatz von KI

Maßnahmen zur Erhöhung der IT-Sicherheit

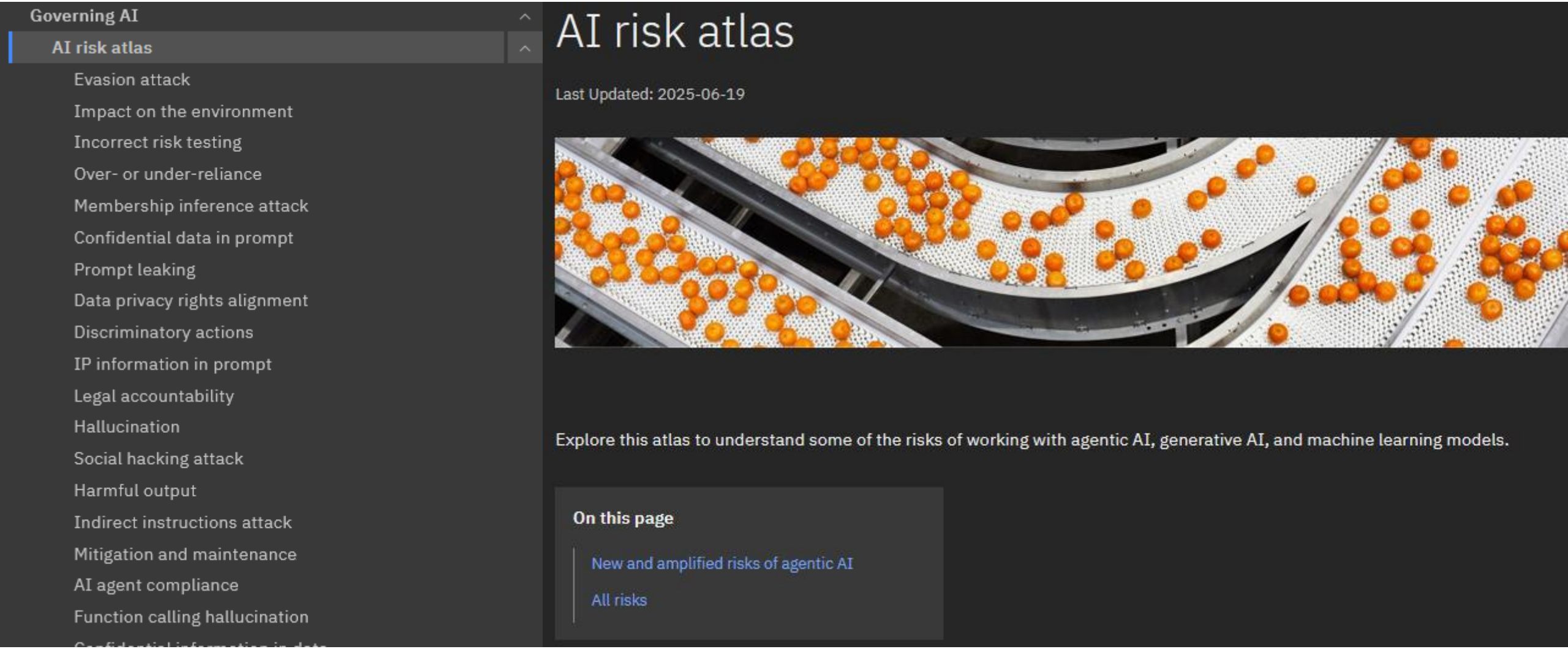
„Die klassischen Maßnahmen hinsichtlich Software- und Systemsicherheit gelten unverändert für KI-Systeme und sollten umgesetzt werden. Dies alleine ist jedoch nicht ausreichend.“

Weiterhin …

„Im Rahmen eines KI-spezifischen Risikomanagements sollte der gesamte Lebenszyklus des KI-Systems systematisch hinsichtlich relevanter Risiken analysiert werden, […]. Basierend auf der Analyse können risikobasiert Mitigationsmaßnahmen auf der Ebene des KI-Systems sowie weitere technische oder organisatorische Maßnahmen zur Änderung der Rahmenbedingungen abgeleitet werden.“

Sources:
https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html
https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf?__blob=publicationFile&v=6

AI risk atlas



Sources: https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas

TD SYNNEX

AI Governance needs to provide …

Monitoring & Evaluations

Monitoring of predictive models for fairness/bias, accuracy, quality and drift and generative AI for quality and LLM-output specific metrics

Explainability

Tracking and gathering of model metrics to explain model output

Risk Management & Compliance

Way to manage AI activities and risks from initial requests, production deployments, to compliance

AI Security needs to provide …

Visibility

Understand where AI models are deployed and in use within the organization

Security for Data

Protect data sources, especially sensitive data, leveraged in AI deployments

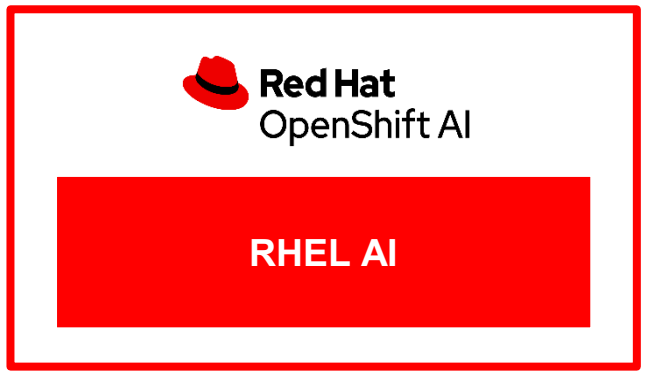
Security for Models & Applications

Apply security controls and mitigate security vulnerabilities

+ Security for Networks, Infrastructure, etc.

TD SYNNEX

The power of community: Red Hat OpenShift AI and RHEL AI



Red Hat OpenShift AI

Model serving & monitoring

Deploy models anywhere with consistent cloud-to-edge production deployment, and monitoring capabilities

Resource optimization and management

Create model pipelines for tuning and alignment, validation and delivery; Share and optimize resources across teams

RHEL AI

Foundation model runtime engine

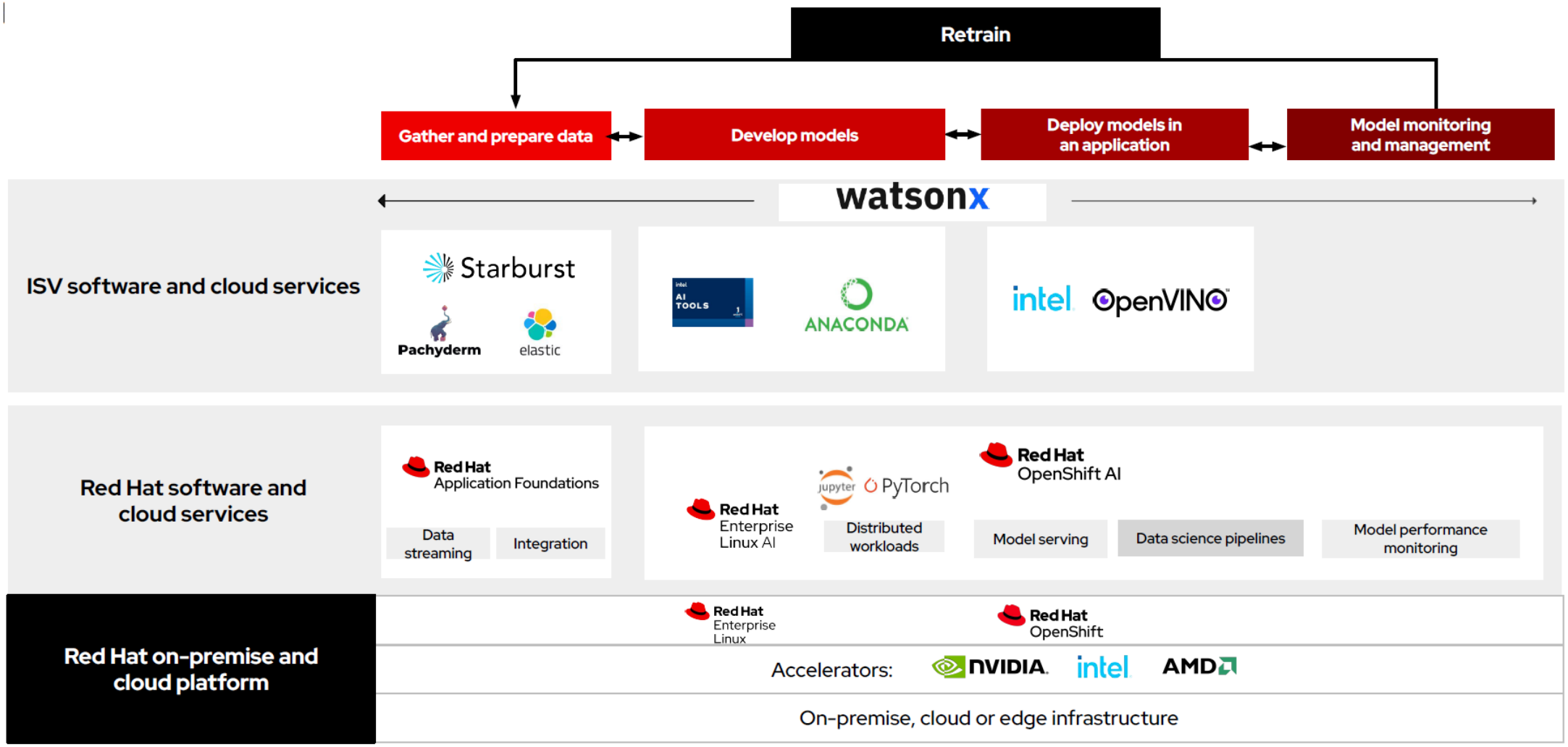
Develop, deploy, and test open Granite language/code models, supported & indemnified from Red Hat and IBM

Developer-friendly model toolkit

Bring rapid LLM experimentation to developers via supported InstructLab CLI

Runs on any infrastructure

Red Hat OpenShift AI: A platform to build



TD SYNNEX

Red Hat

Authorized Distributor

The Power of us

Connect. Empower. Transform.