AI Risk Management as the Foundation for Compliance, Safety, and Trust under the AI Act

Isabel Barberá
AI Risk & Safety Researcher
AI Privacy & Security Advisor

Why should we care about AI Risk Management and why is it often so complex

“In the land of Fantasia, stories never end… just like managing the risks of powerful AI systems.”

AI Lifecycle

Source: Rhite based on ISO/IEC 22989:2022Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

“Fighting against The Nothing
to keep the Fantasy World.”

The Nothing = Risks

The Book is Opened **(Inception & Design)**

Like Bastian opening the magical book, your team begins the journey by envisioning a new LLM-based system. This phase is filled with hope and potential, but also with the shadows of unseen risks.

"What could go wrong?" **becomes your guiding question. DOES IT?**

Building Fantasia (Design & Development)

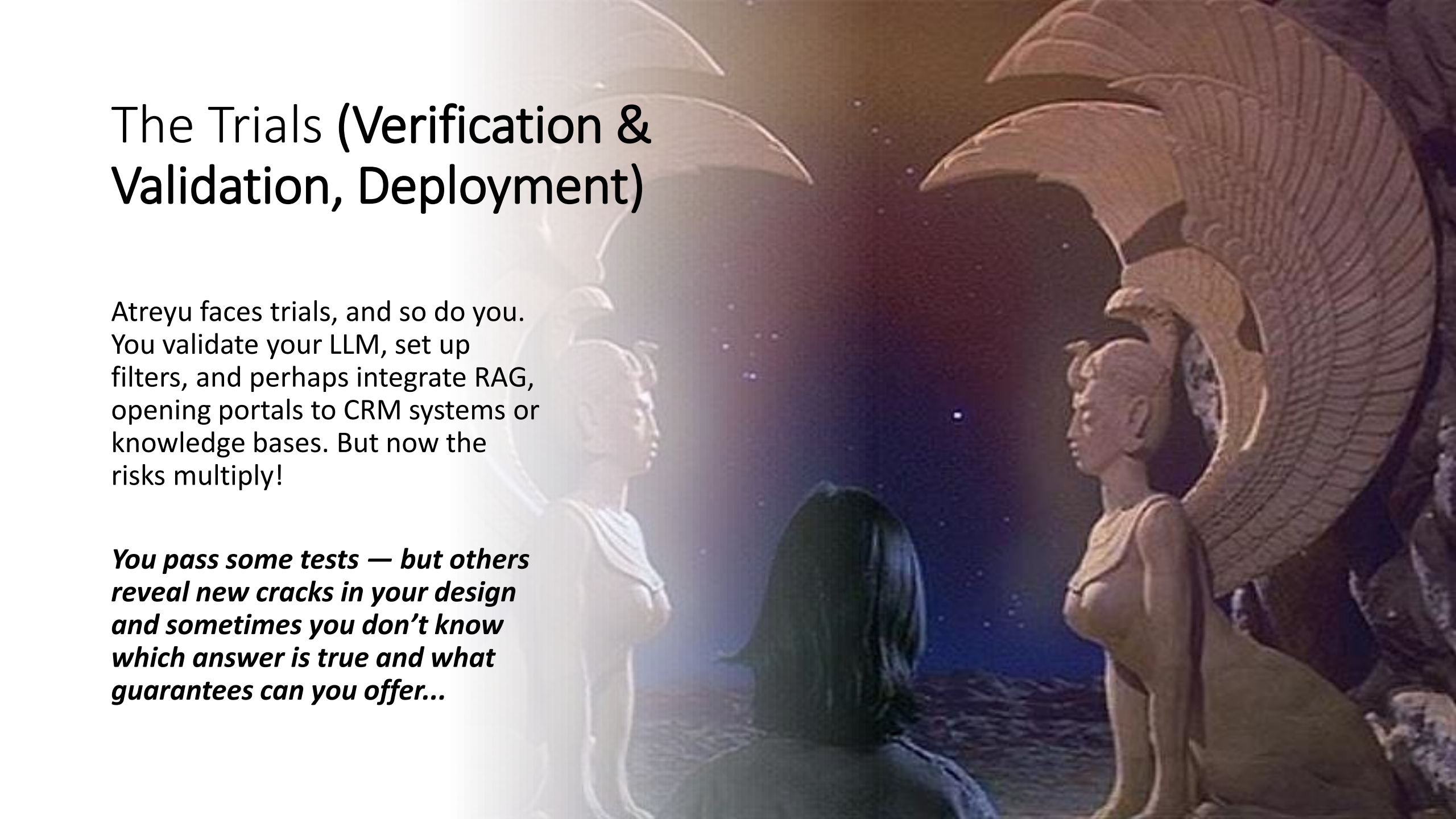
Just as Atreyu ventures into the unknown, you train and fine-tune your model. But not all knowledge is pure — **biases, hallucinations, and misuse potential** lurk in the training data like invisible creatures.

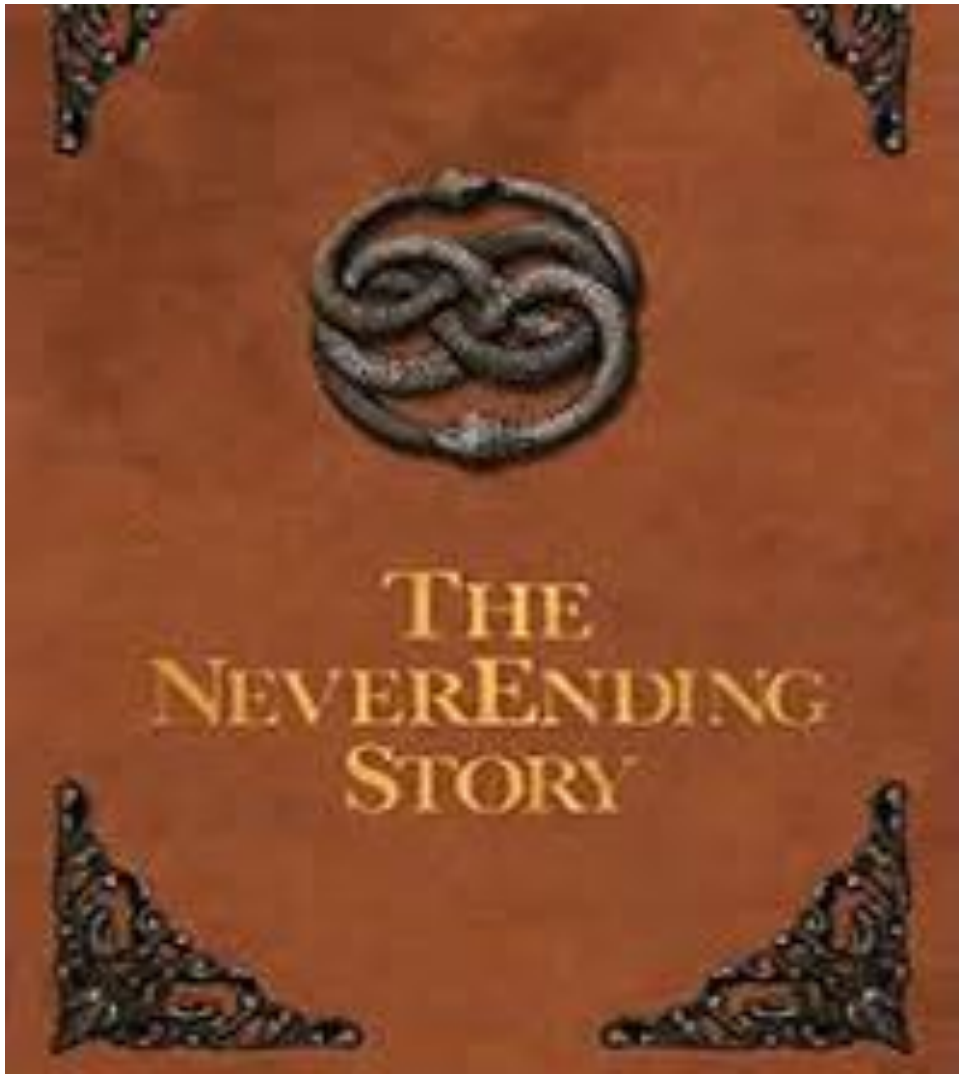
You discover that many design choices introduce new possibilities of harm.

The Trials (Verification & Validation, Deployment)

Atreyu faces trials, and so do you. You validate your LLM, set up filters, and perhaps integrate RAG, opening portals to CRM systems or knowledge bases. But now the risks multiply!

You pass some tests — but others reveal new cracks in your design and sometimes you don't know which answer is true and what guarantees can you offer...





The Cycles of Time (Monitoring, Re-Evaluation)

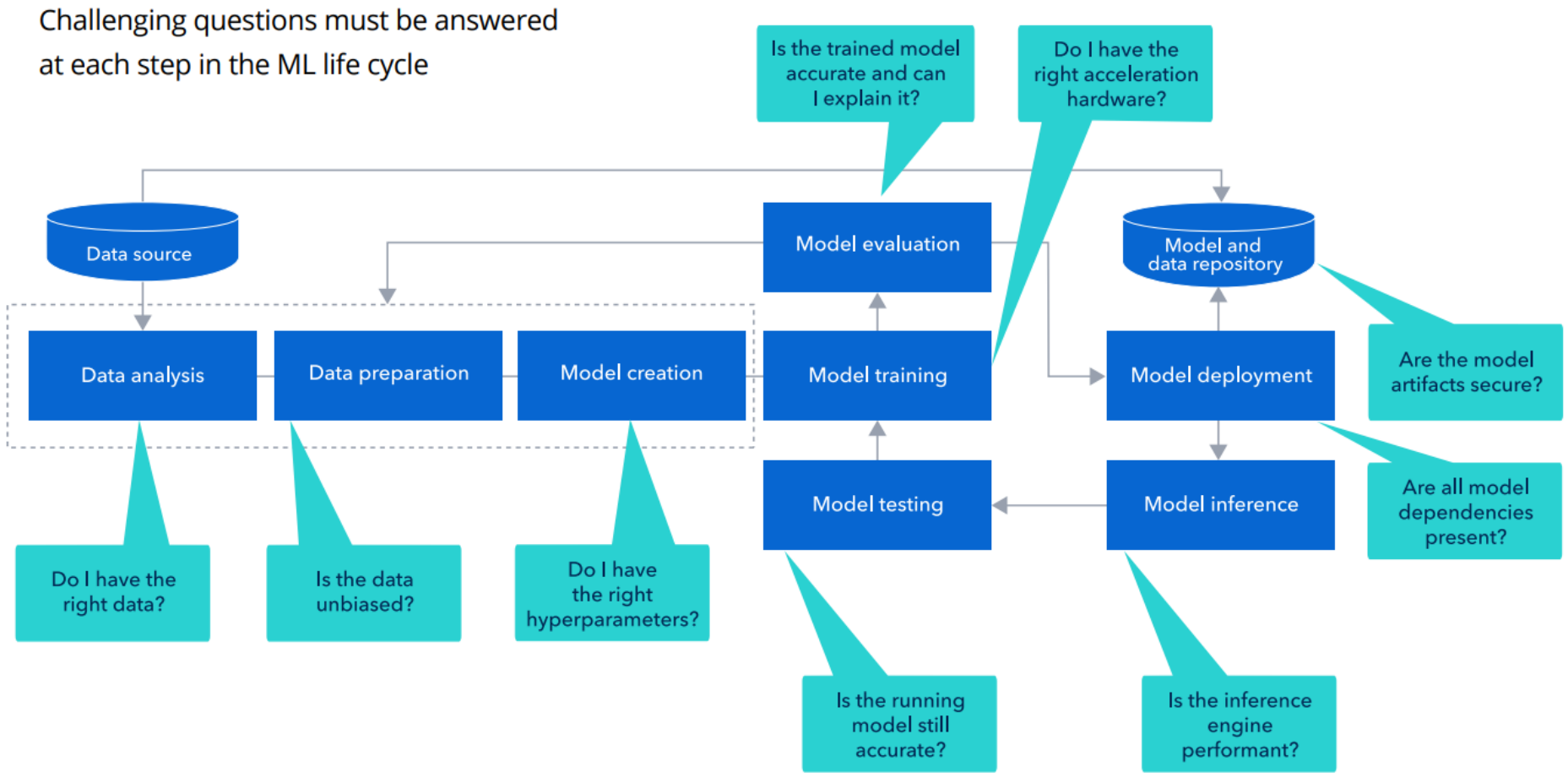
Even as the system works, you realize it changes with time. Like Bastian in the real world, you must return to the system again and again, bringing **new insights**, **new evaluations**, **new safeguards**, more **headache**, and more **COSTS**!!

It is a never-ending story!

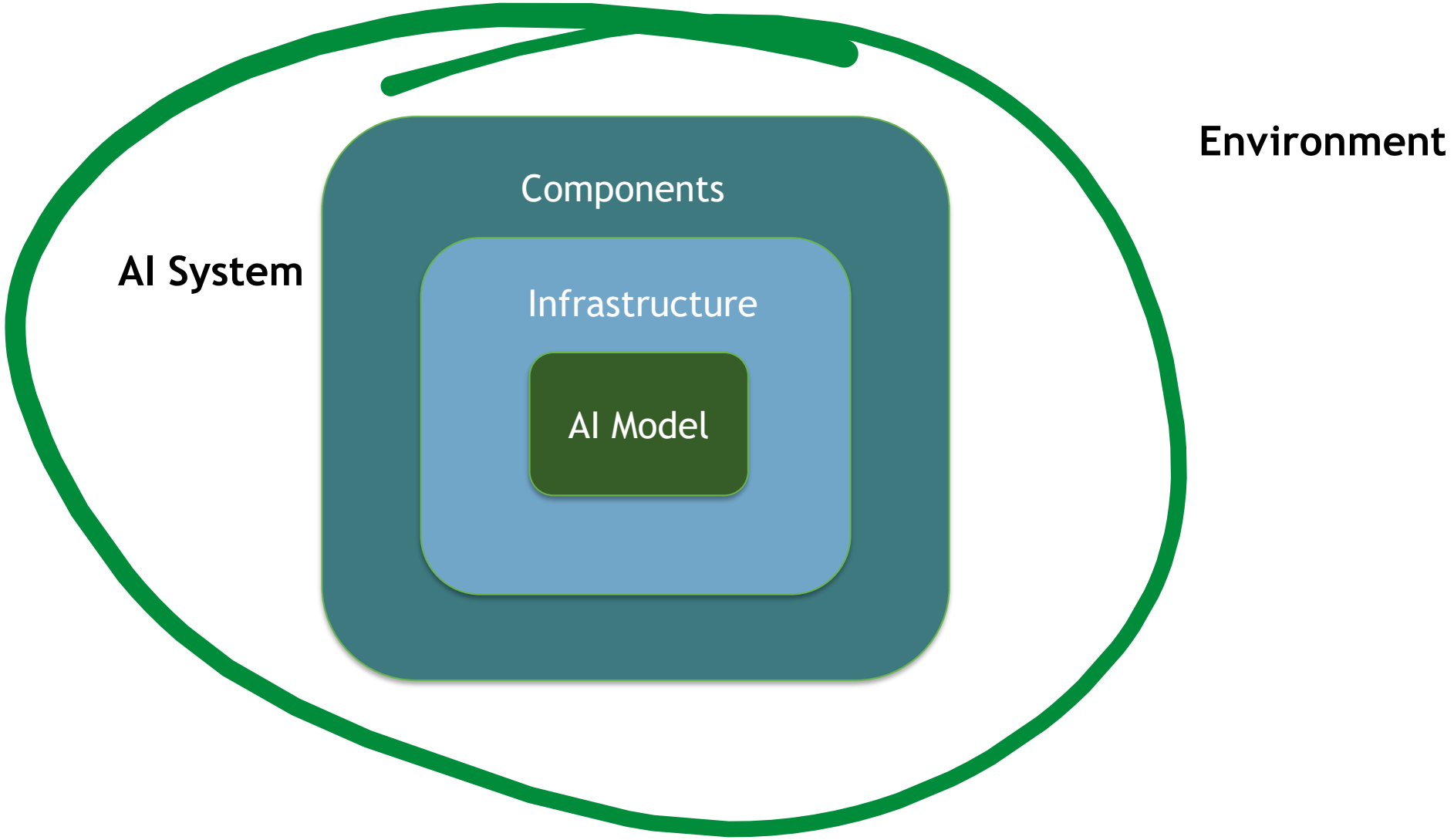
Risks Based on Service Model

- Interface: chatbot, agent
- Robot, etc…

GEN AI Pipeline example: supply chain complexity



AI threat landscape

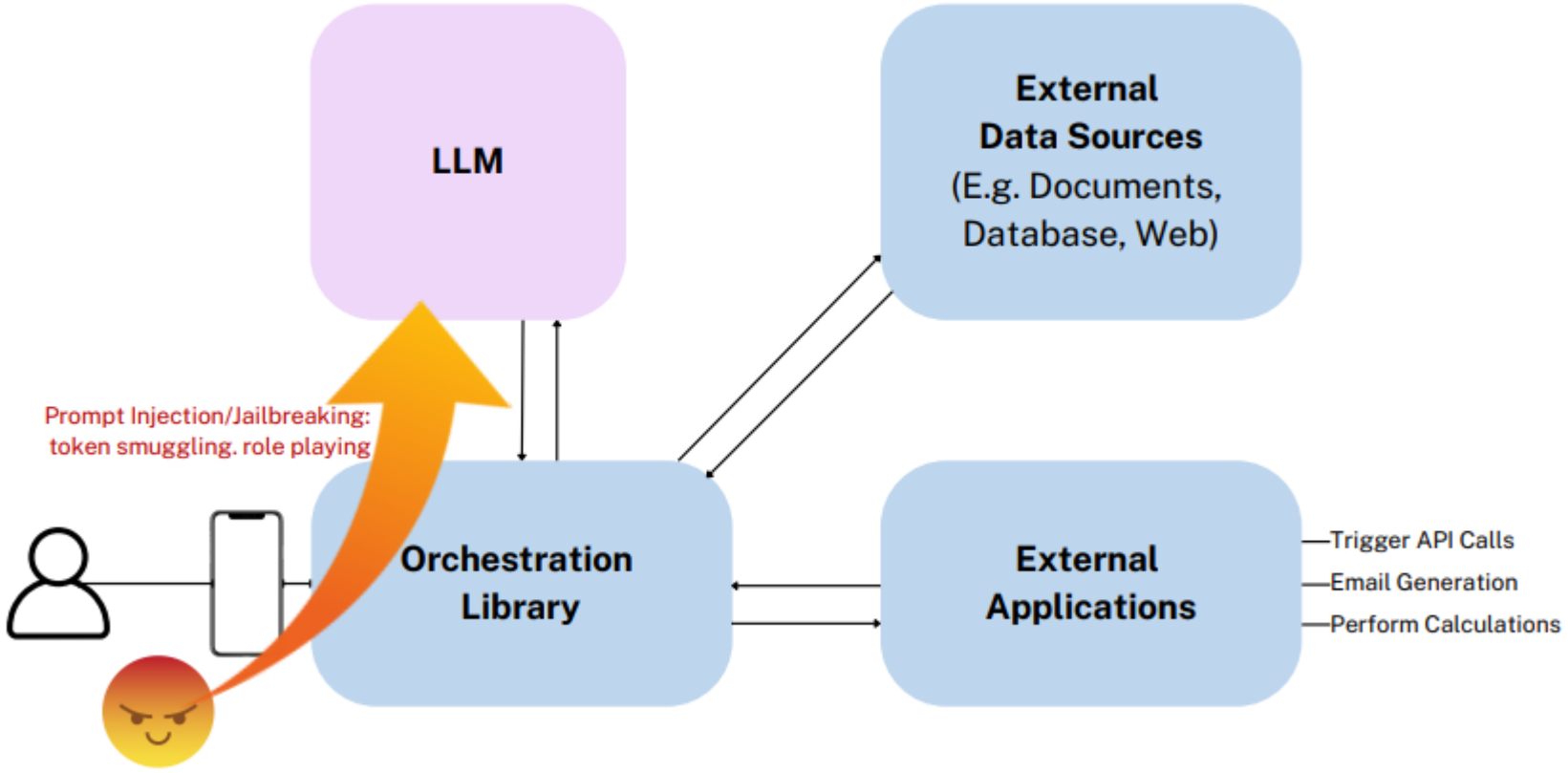


GENAI Data Lifecycle

Training Model

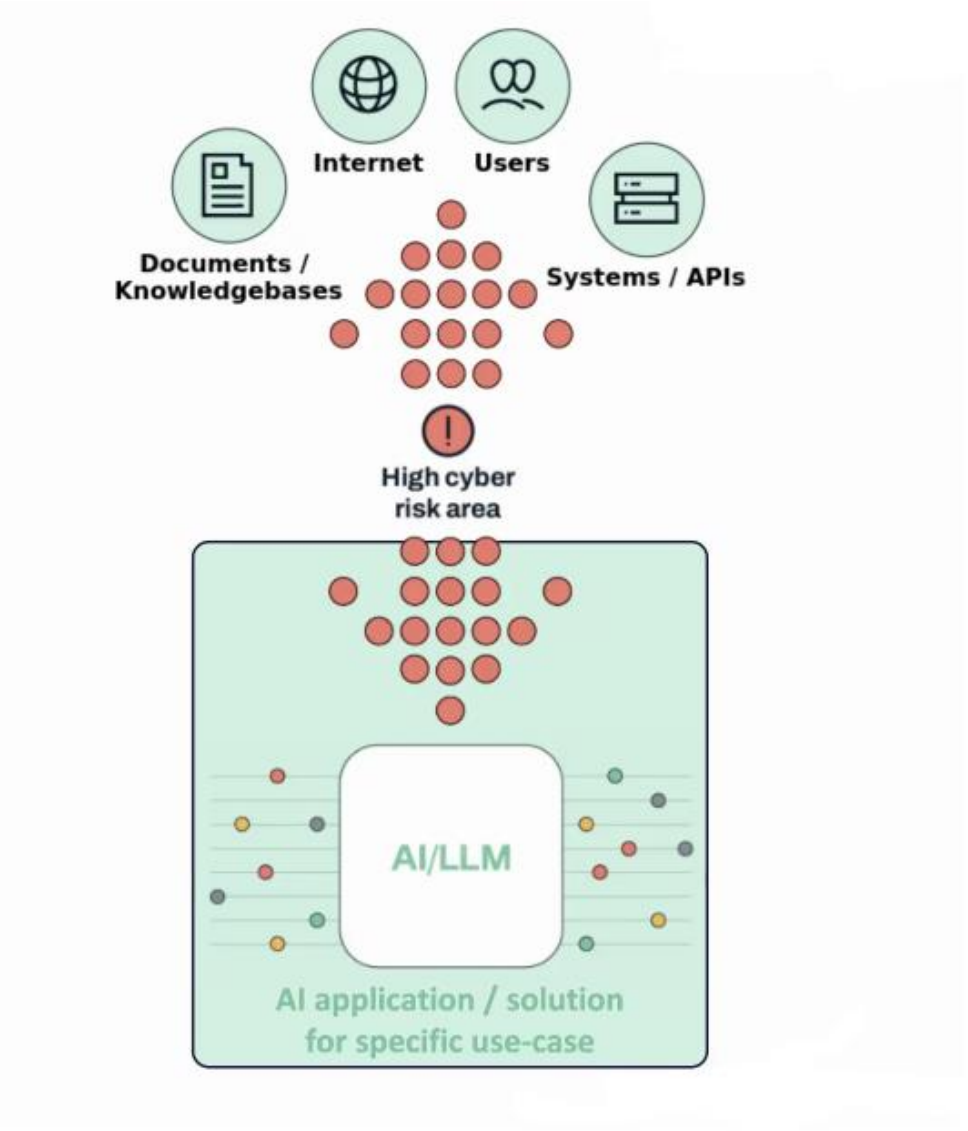
RAG
Retrieval-augmented generation

Training- or inference-time techniques in real-world



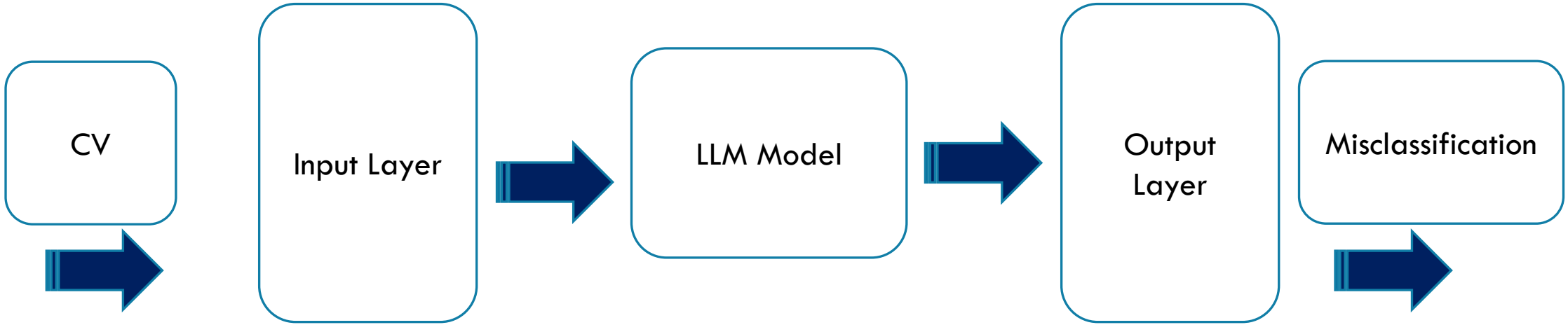
Source: Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2e2025

THREAT LANDSCAPE



Source: WithSecure Consulting

Model vs System: integrations & architecture



Don’t blame the model !

Do you recognize this problem?

You:
Building your idea

When prospect client ask you:

- Copyright
- Risk assessment
- Ethical approach
- AI Act / GDPR compliance
- Sustainability

Scalable AI systems

Trustworthy AI requires focus on

- Risk Management
- Transparency
- Accountability
- Data governance
- Sustainability

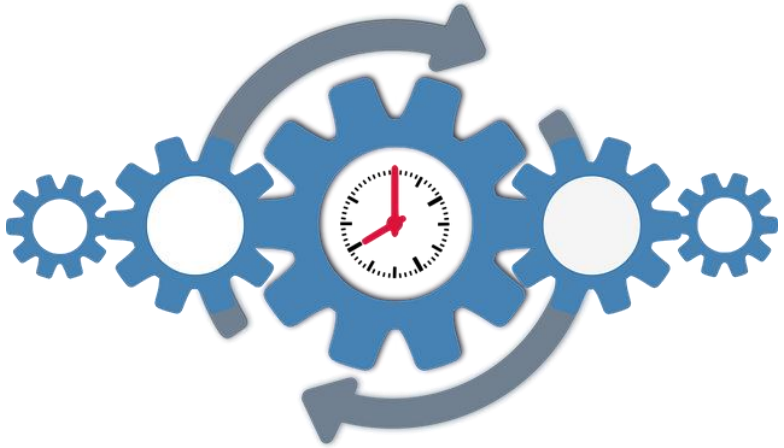
To get as result:

- Trust & safe users
- Trust regulators
- Place in the market

Some foundational AI Act requirements for high risk AI systems

Data Governance

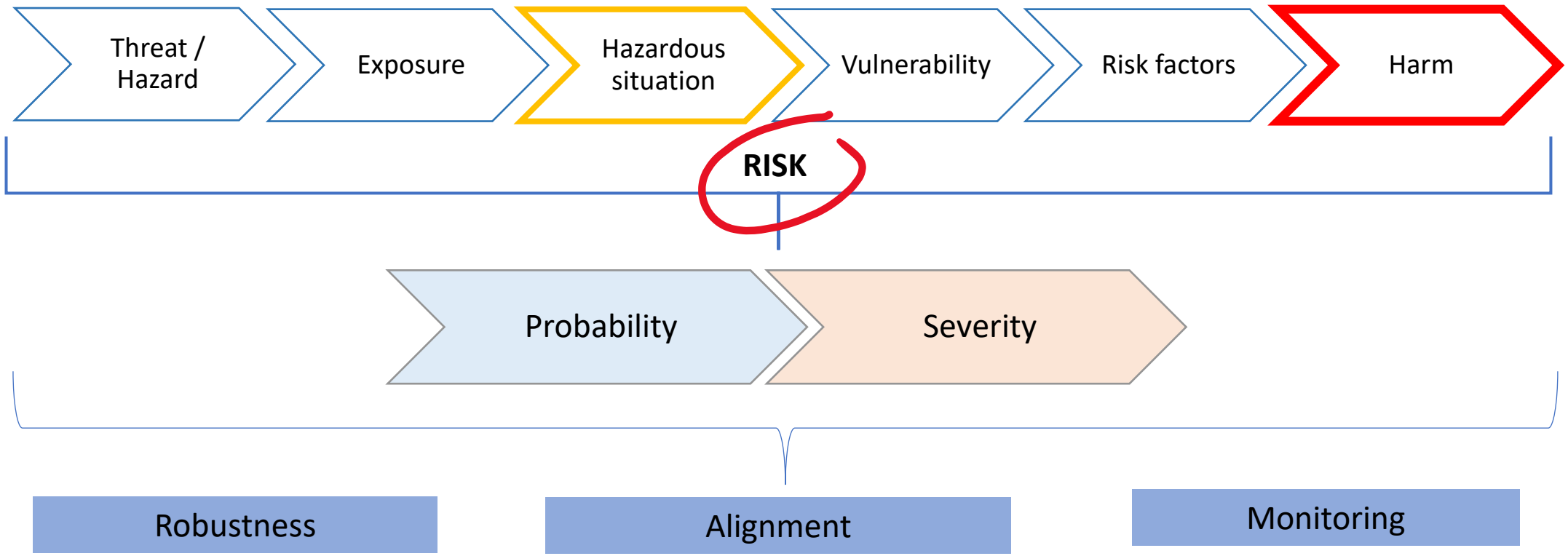
Quality Management System



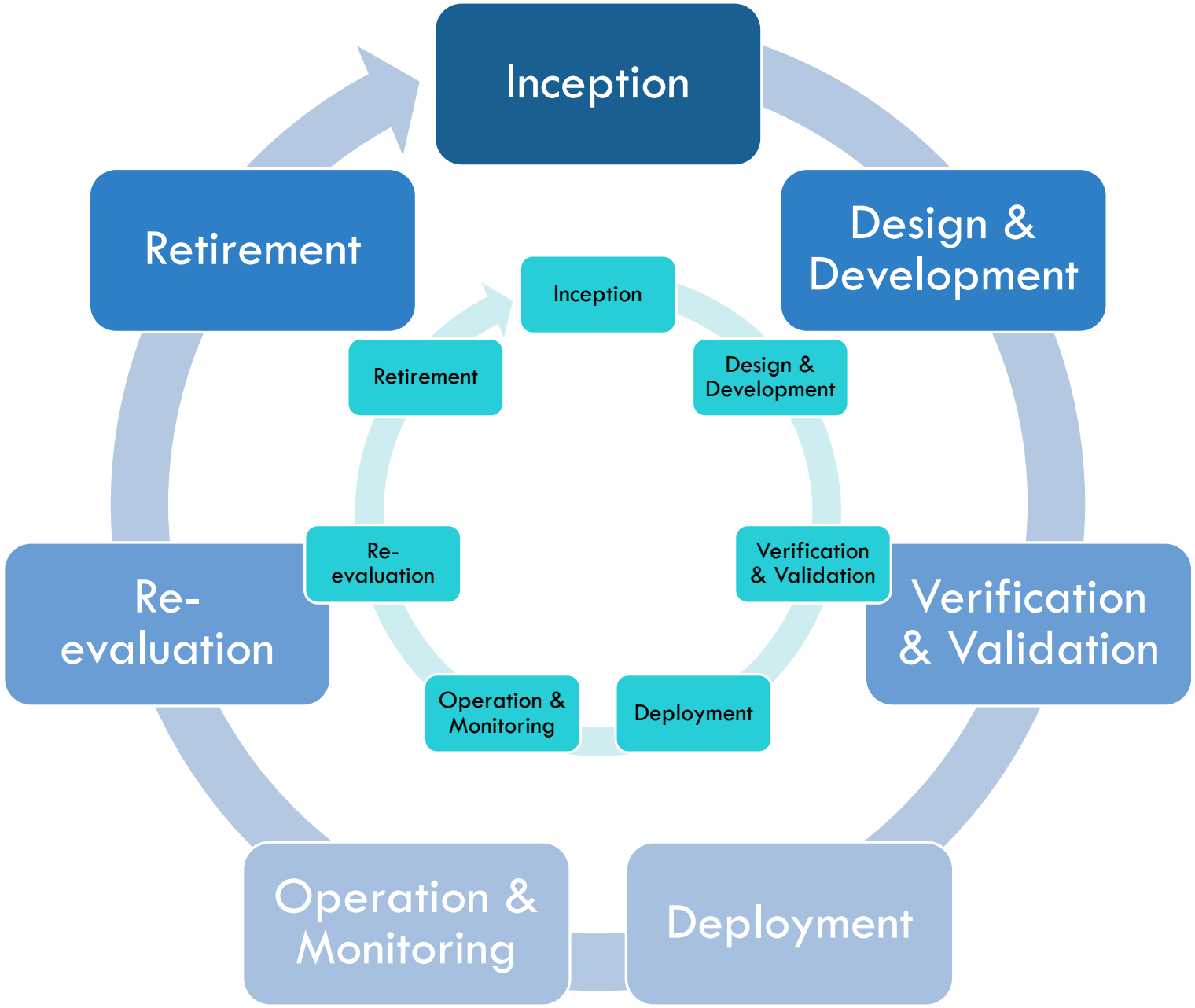
Risk Management

- Security
- Safety
- Health
- Fundamental Rights

AI Risk Management

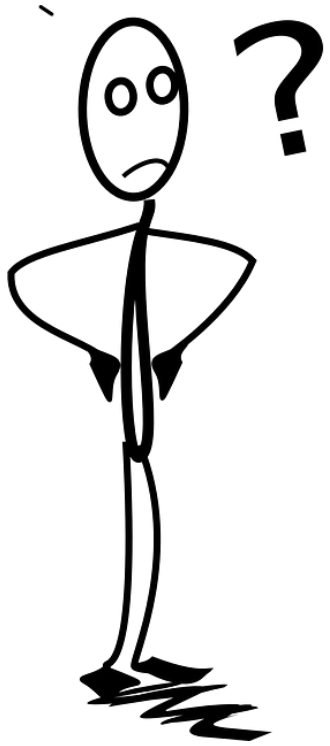


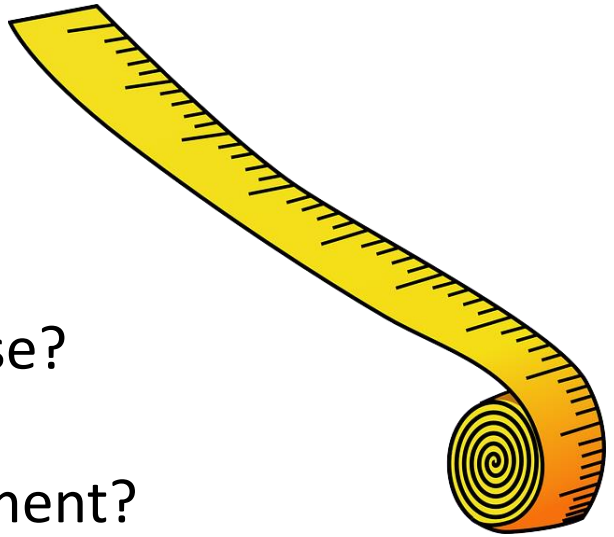
Model vs System



Evaluations, benchmarks, red teaming, real world testing….

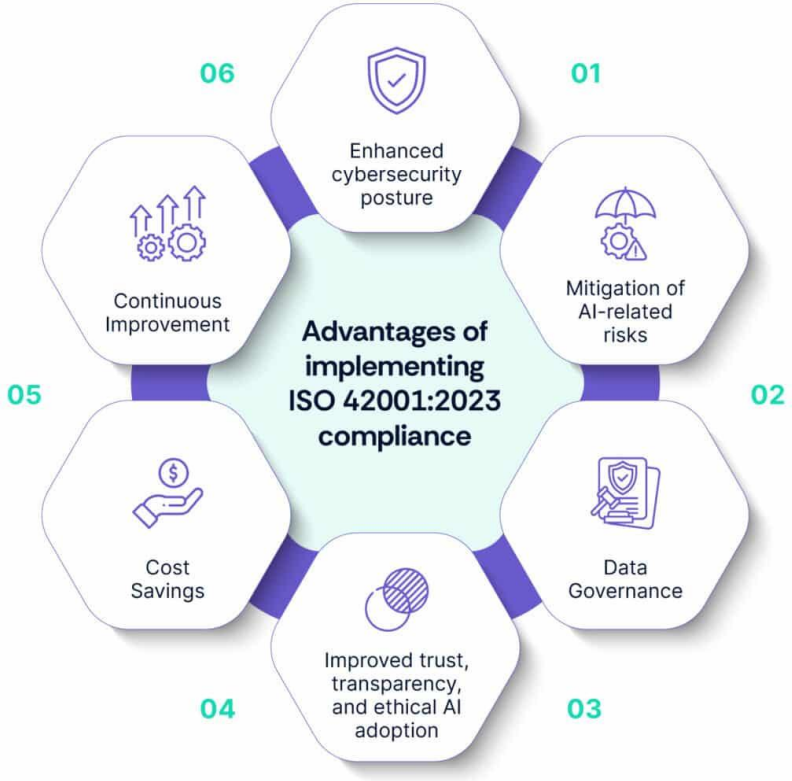
But the question remains….





- ✓ Which metrics & benchmarks should we use?
- ✓ Which thresholds should we establish?
- ✓ Which logging and alerts should we implement?
- ✓ How do I implement all this?

International AI Risk Management Standard



Source: https://www.scrut.io/post/your-ultimate-guide-to-iso-42001-2023

INTERNATIONAL STANDARD

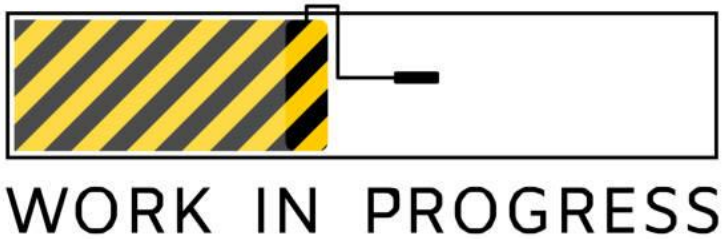
ISO/IEC 23894

First edition
2023-02

Information technology — Artificial intelligence — Guidance on risk management

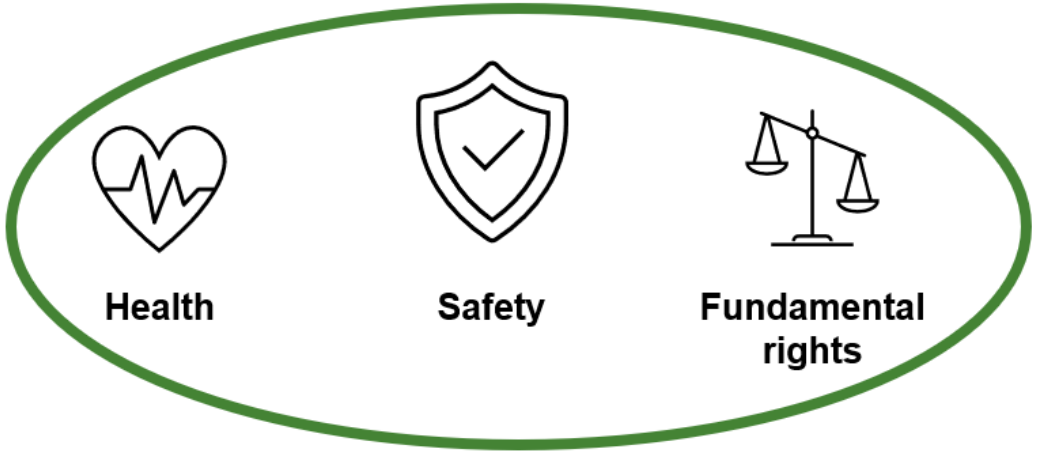
Technologies de l'information — Intelligence artificielle — Recommandations relatives au management du risque

European AI Risk Management Standard

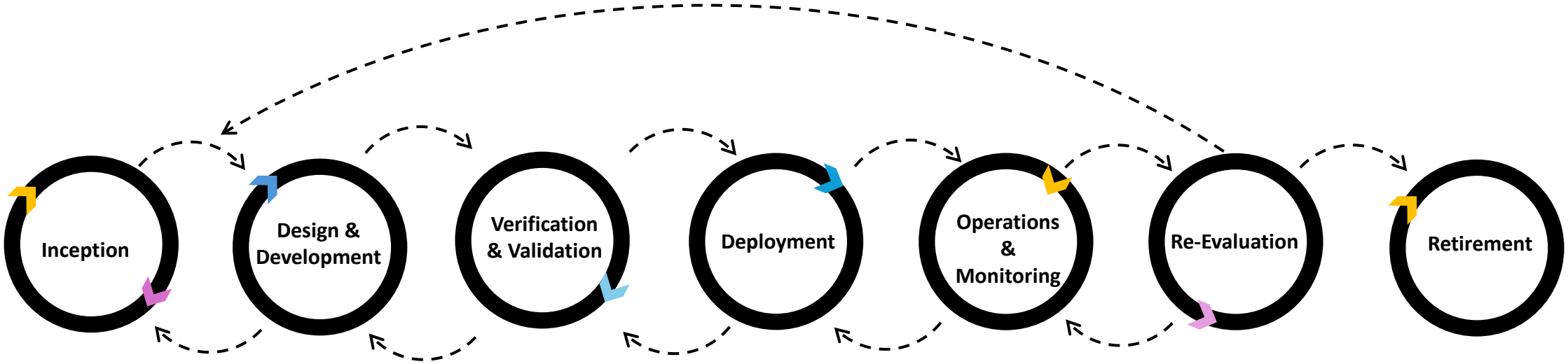


CEN CENELEC

JTC21



Risk Management during the AI Lifecycle



Risk Management

	Inception	Design & Development	Verification & Validation	Deployment	Operations & Monitoring	Re-Evaluation	Retirement
Risk Management Steps	• Risk Assessment • Risk Control • Residual Risk Evaluation • Review	• Risk Assessment • Risk Control • Residual Risk Evaluation • Review • Monitor	• Risk Assessment • Risk Control • Residual Risk Evaluation • Review	• Monitor • Risk Assessment • Risk Control • Residual Risk Evaluation • Review	• Risk Assessment • Risk Control • Residual Risk Evaluation • Review	
Instruments for Quality Assurance & Risk identification	• Stakeholders collaboration • Threat Modeling • Test & Experimentation	• Stakeholders collaboration • Threat Modeling • Tests, Evaluations, Benchmarks, AI Red Teaming • Unit test, integration test	• Stakeholders collaboration • Threat Modeling • Experimentation • Integration Test • AI Red Teaming, Pentesting	• Stakeholders collaboration • Threat Modeling • Tests, Evaluations, Benchmarks, Red Teaming • Tracing, logs, debugging, unit test, integration test, A/B testing • AI Red Teaming	• Stakeholders collaboration • Threat Modeling • Decommisioning		

SUPPORT POOL
OF EXPERTS PROGRAMME

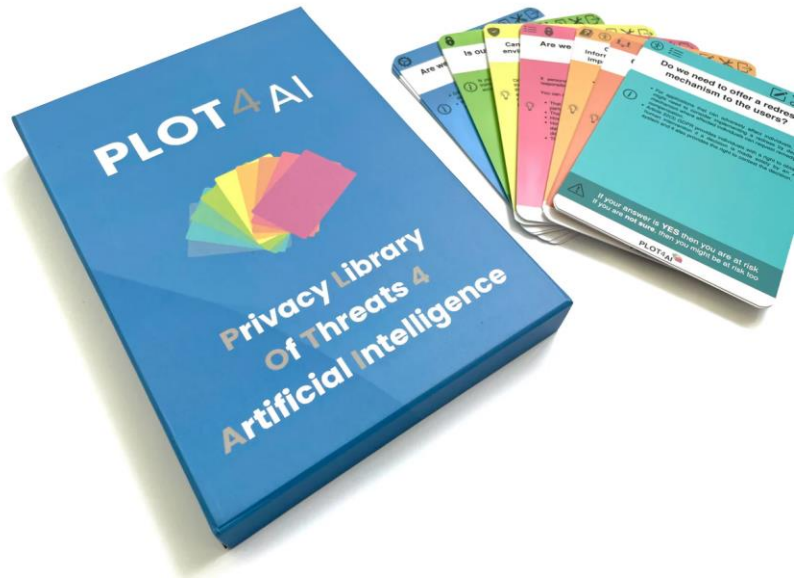
AI Privacy Risks & Mitigations
Large Language Models (LLMs)

By Isabel BARBERÁ

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Data Protection

Privacy Enhancing Technologies (synthetic data, LLMs)
The Bureau took note of the presentation by Isabel Barberá (Rhite) and Murielle Popa-Fabre on the concept of an expert report to be drafted by the experts with the contribution of Prof Chris Russell from Oxford Internet Institute, University of Oxford on data protection in Large Language Models, held an exchange of views, and invited the Secretariat to work with the experts on the finalisation of the expert report to be presented during the next plenary meeting in June.



Support AI standardization!

Join as expert!

“The NeverEnding Story is not about magic... it’s about trust & responsibility.”

In AI, **risk management is a living process**. Like Fantasia, it thrives when people care, when threats are faced, and when stories — and risks — are taken seriously.

THANK YOU!

GET
IN
TOUCH

https://www.linkedin.com/in/isabelbarbera/