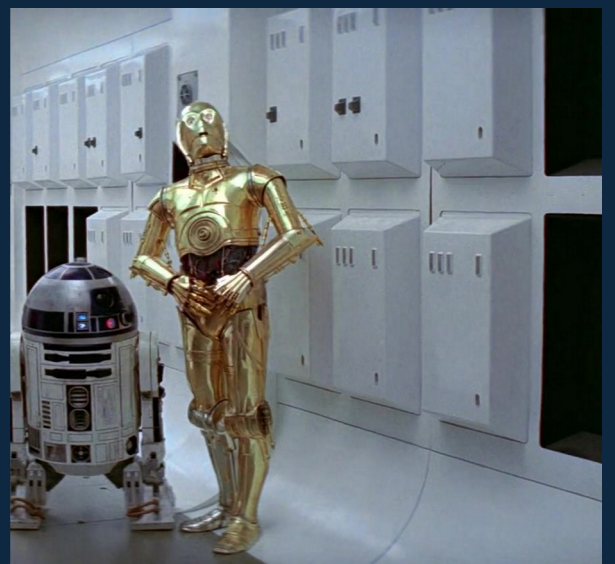
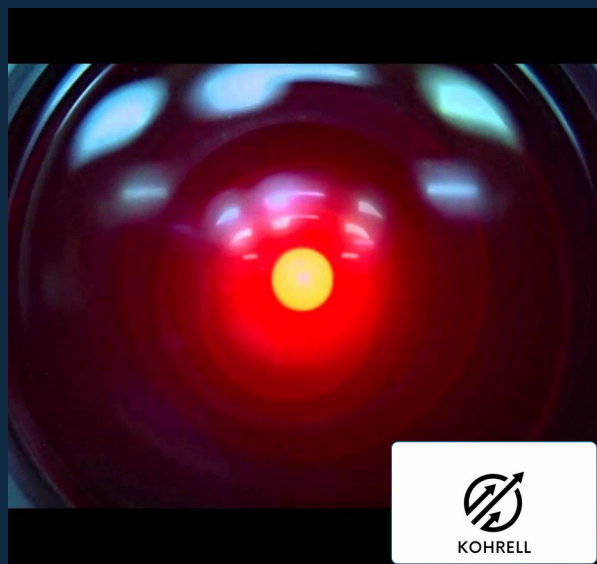
Advanced AI Risk - ISACA

It could be awesome, destroy all of humanity, or hype bigger than Y2k/dot.com; it won't be boring …

- David Kohrell, NEbraskaCERT 18FEB2026
- https://www.necert.org/CSF/





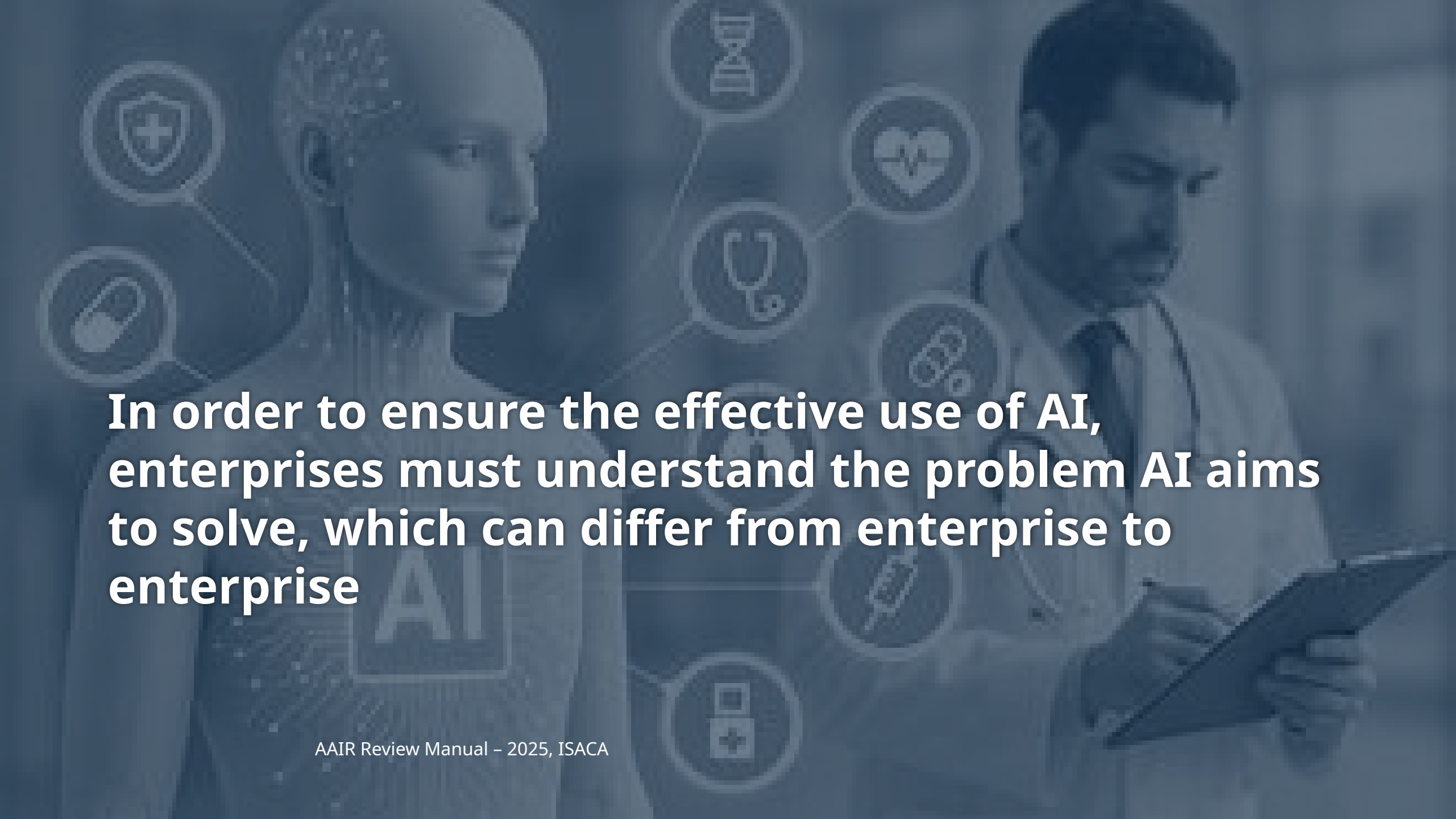
KOHRELL

Definitions and References

Risk = Threat x Vulnerability x Asset

ISSA, Managing Risk in Information Systems, 2022
Gibson & Igonor

AI risk refers to the composite measure of an event’s probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018). NIST RMF, January 2023



In order to ensure the effective use of AI, enterprises must understand the problem AI aims to solve, which can differ from enterprise to enterprise

AAIR Review Manual – 2025, ISACA

AAIR (Advanced AI Risk)

- AAIR is an AI-focused IT risk management credential crafted to enhance the expertise of certified IT risk professionals. It provides the strategic skills to guide management in addressing AI-related risks, safeguarding organizations from potential financial and reputational harm. AAIR's key practice areas include:
  - AI Risk Governance and Framework Integration
  - AI Life Cycle Risk Management
  - AI Risk Program Management

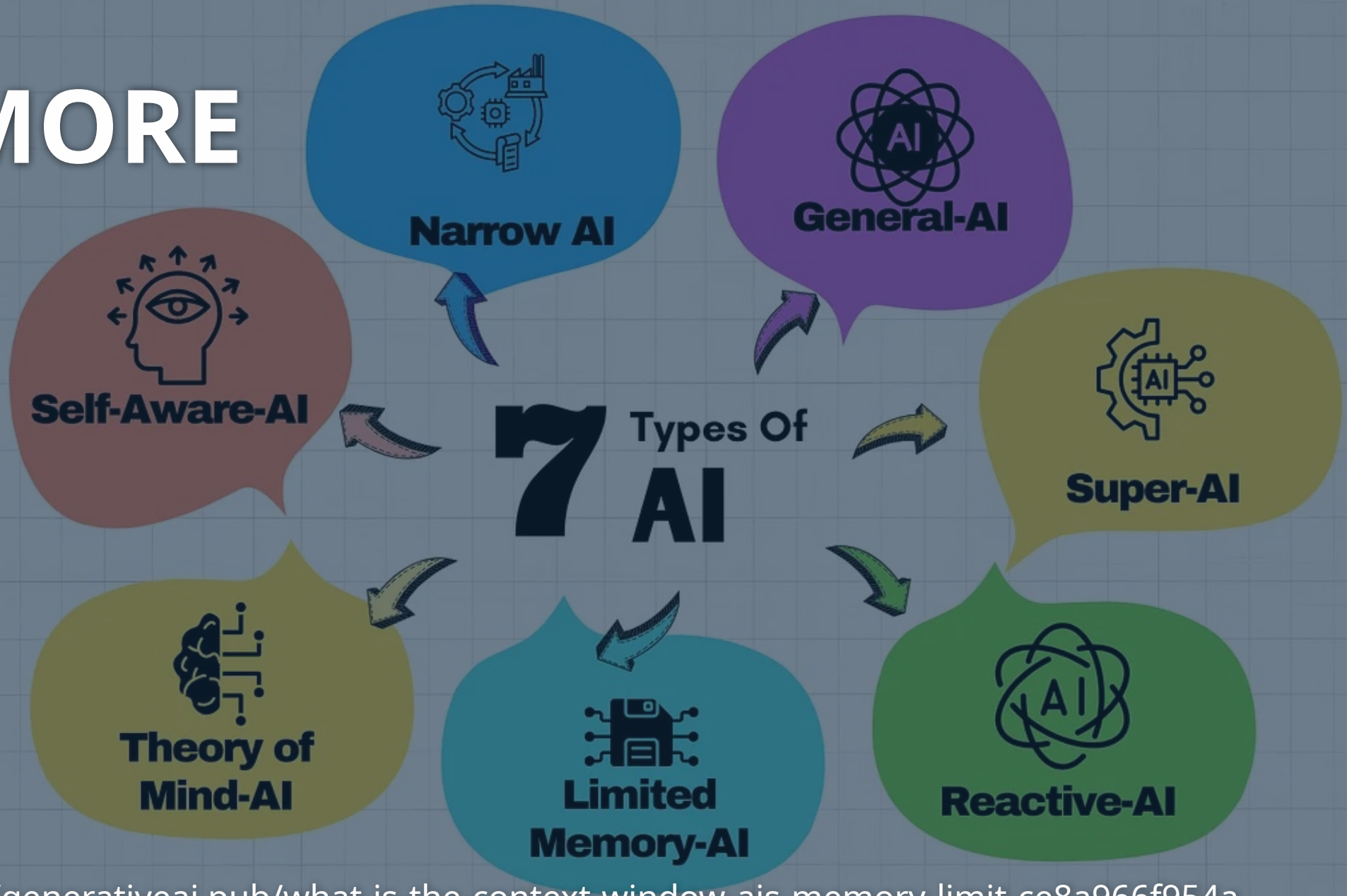
https://www.isaca.org/credentialing/aair

AI “levels”

- Automated Narrow Intelligence – what we have now
- Automated General Intelligence – the hype driving investments and circular economy
- Automated Super Intelligence – Paradise or Machines turn on us. Awesome or Dystopian

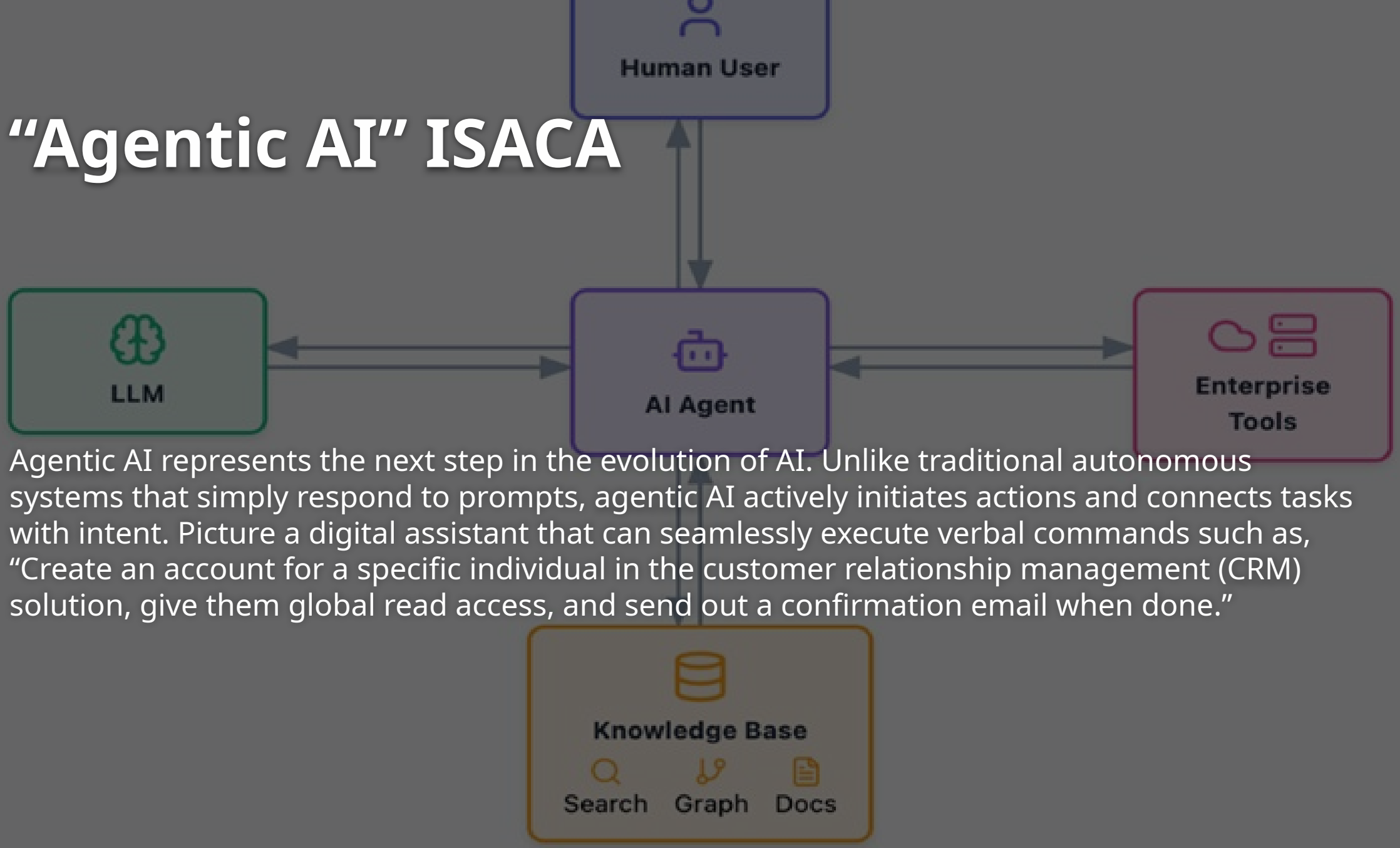
https://selftution.com/types-artificial-intelligence-ai-examples-narrow-general-super-ani-agi-asi-machine-learning-ml/

4 MORE



https://generativeai.pub/what-is-the-context-window-ais-memory-limit-ce8a966f954a

“Agentic AI” ISACA



Agentic AI represents the next step in the evolution of AI. Unlike traditional autonomous systems that simply respond to prompts, agentic AI actively initiates actions and connects tasks with intent. Picture a digital assistant that can seamlessly execute verbal commands such as, “Create an account for a specific individual in the customer relationship management (CRM) solution, give them global read access, and send out a confirmation email when done.”

Safeguarding the Enterprise AI Evolution: Best Practices for Agentic AI Workflows Murali 1JUL2025, ISACA

AI Risk Governance and Framework Integration

What Are the Risk Categories in the EU AI Act?

Unacceptable Risk:

- Prohibited due to significant ethical or safety concerns.
- Examples: Social scoring by governments, manipulative AI, biometric surveillance

High Risk:

- AI systems that significantly impact people's rights or safety.
- Requires conformity assessments, risk management, human oversight.
- Examples: AI used in critical infrastructure, education, employment, law enforcement, and healthcare.

General Purpose AI (GPAI):

- Broad AI systems that can be used across sectors and are subject to specific transparency and risk management requirements.
- Examples: Large AI models by companies like OpenAI. These models must comply with extra obligations when used in high-risk contexts.

Limited or Minimal Risk:

- Minimal risk to safety or fundamental rights, subject to limited transparency obligations (e.g., disclosing that users are interacting with an AI).
- Examples: Customer service chatbots, AI-based recommendations in e-commerce.

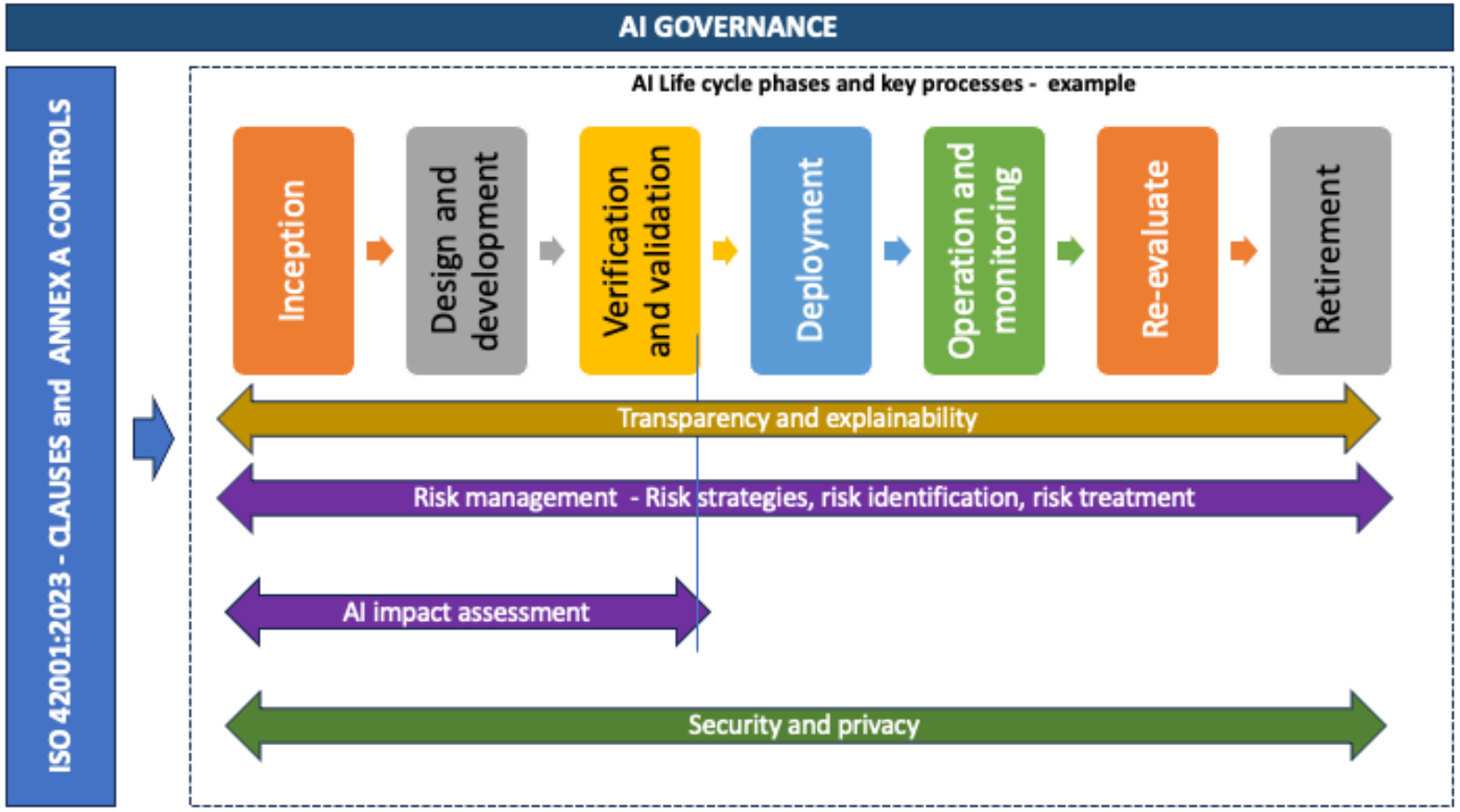
SecurityCompass

EU AI Act

https://www.securitycompass.com/blog/understanding-eu-ai-act-risk-categories/

NIST RMF

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf



ISO/IEC 42001:2023 for AI

AWS Security Blog, Javid & Welch, May 2025

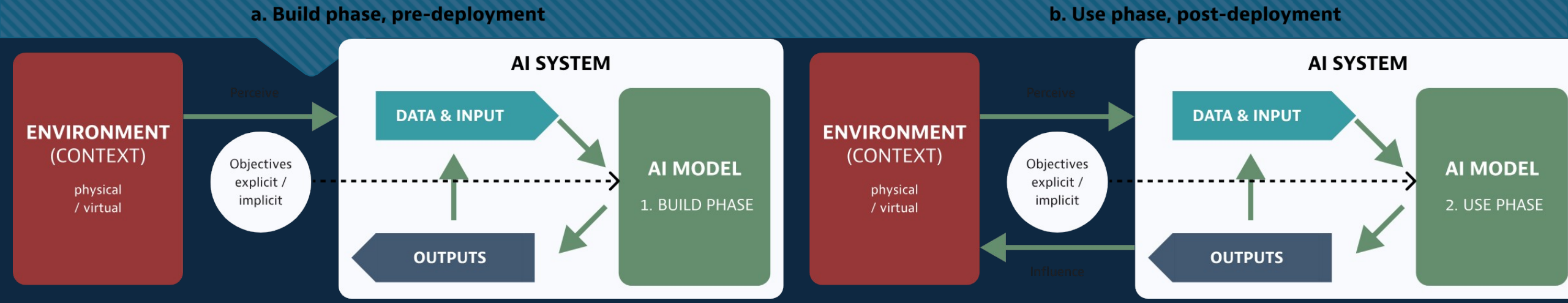
https://aws.amazon.com/blogs/security/ai-lifecycle-risk-management-iso-iec-420012023-for-ai-governance/

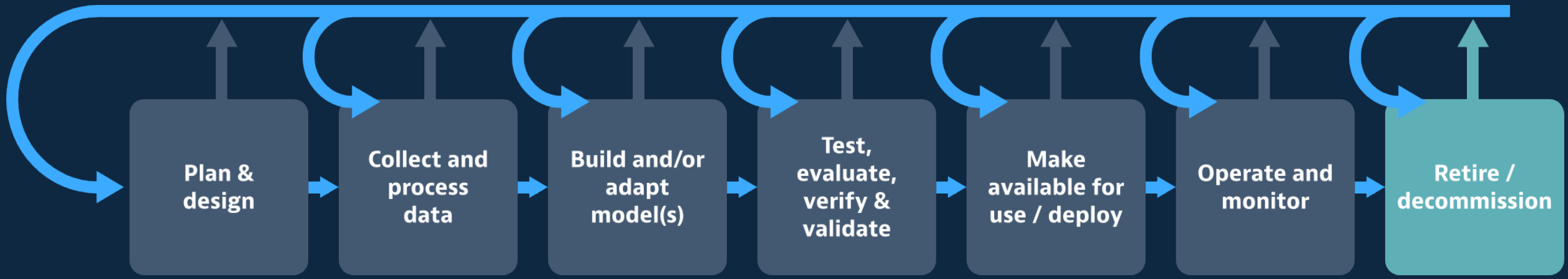
AI Life Cycle Risk Management

Four Governance Key’s

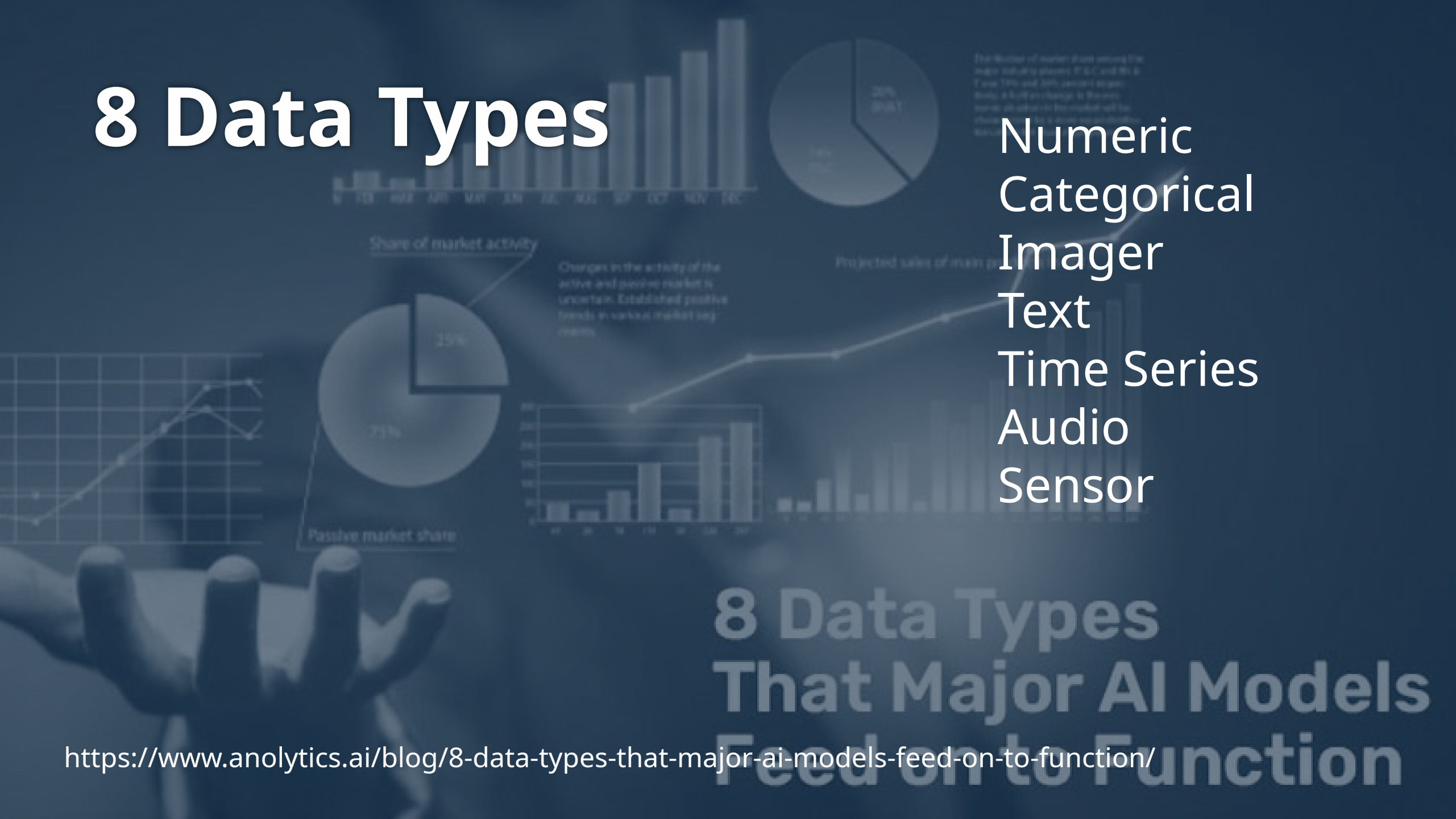
- Human in the Loop
- Transparency
- Traceability
- Explainability

OECD https://oecd.ai/en/ai-principles (2019 and 2024)





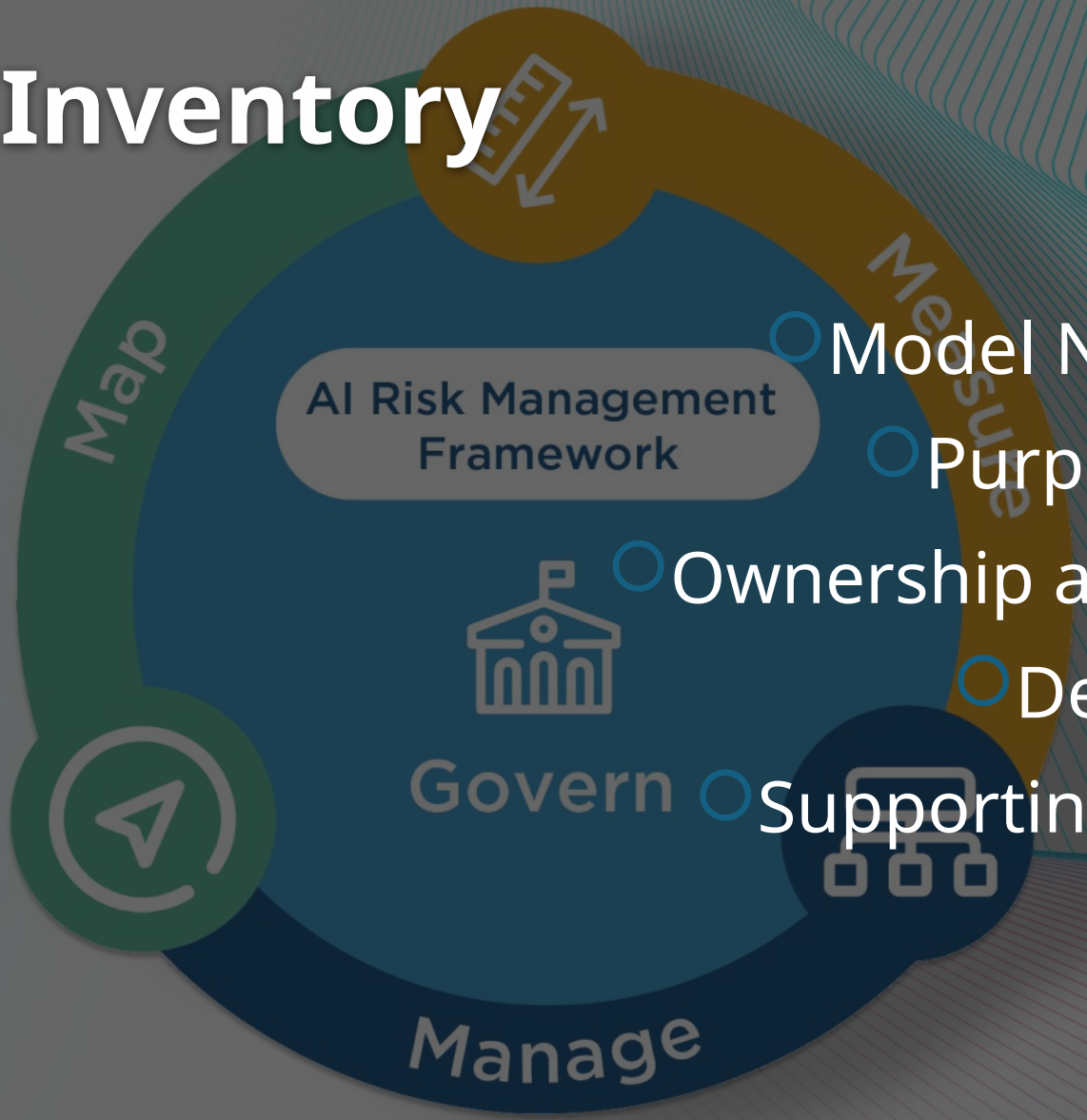
8 Data Types



- Numeric
- Categorical
- Imager
- Text
- Time Series
- Audio
- Sensor

https://www.anolytics.ai/blog/8-data-types-that-major-ai-models-feed-on-to-function/

AI Model Inventory



- Model Name and Version
- Purpose and Use Case
- Ownership and Accountability
- Deployment Details
- Supporting Documentation

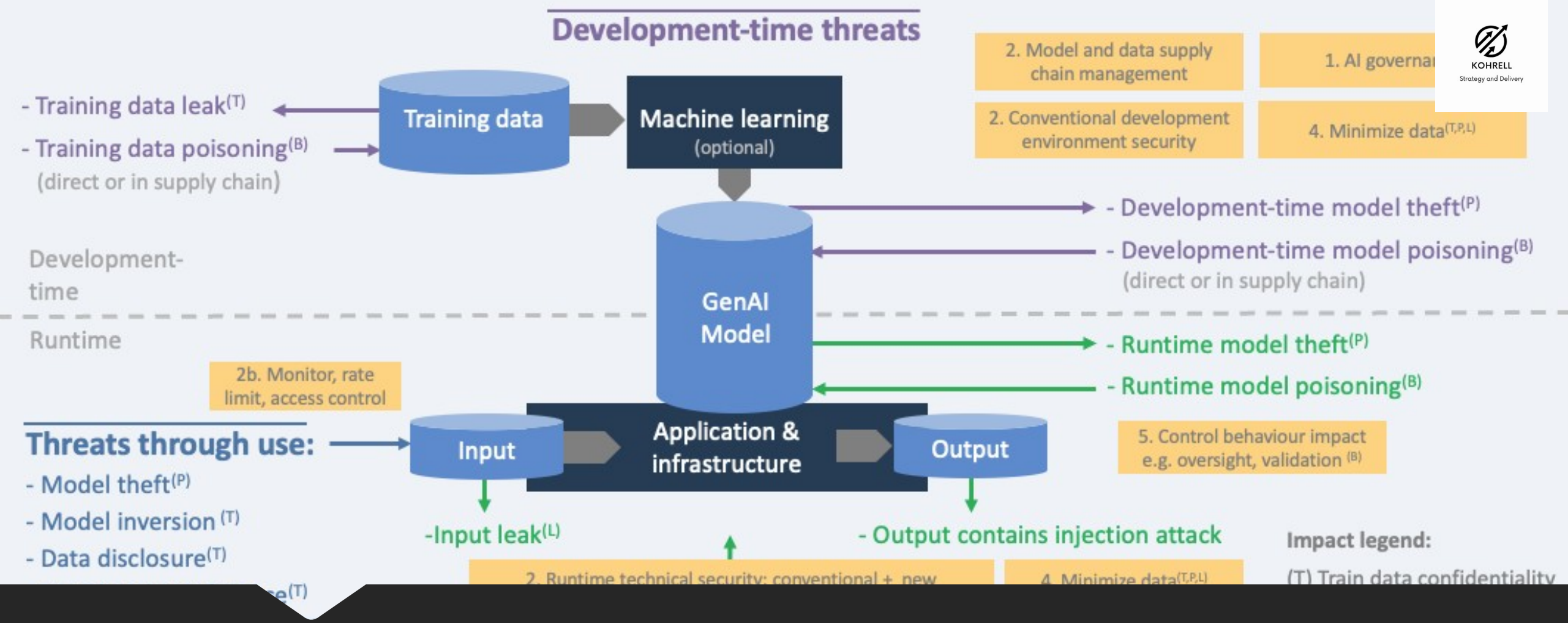
https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

KOHRELL
Strategy and Delivery

Risk Management Life Cycle

https://www.alertmedia.com/blog/risk-management-lifecycle/

AI Risk Program Management



Development Time Threats – AI going bad

https://owaspai.org/docs/ai_security_overview/

KOHRELL
Strategy and Delivery

AI Threat Mitigation and Robustness

https://cybercx.com.au/blog/designing-an-effective-ai-risk-mitigation-strategy/

2025 OWASP Top 10 List for LLM and Gen AI

LLM01:25

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02:25

Sensitive Information Disclosure

Sensitive info in LLMs includes PII, financial, health, business, security, and legal data. Proprietary models face risks with unique training methods and source code, critical in closed or foundation models.

LLM03:25

Supply Chain

LLM supply chains face risks in training data, models, and platforms, causing bias, breaches, or failures. Unlike traditional software, ML risks include third-party pre-trained models and data vulnerabilities.

LLM04:25

Data and Model Poisoning

Data poisoning manipulates pre-training, fine-tuning, or embedding data, causing vulnerabilities, biases, or backdoors. Risks include degraded performance, harmful outputs, toxic content, and compromised downstream systems.

LLM05:25

Improper Output Handling

Improper Output Handling involves inadequate validation of LLM outputs before downstream use. Exploits include XSS, CSRF, SSRF, privilege escalation, or remote code execution, which differs from Overreliance.

LLM06:25

Excessive Agency

LLM systems gain agency via extensions, tools, or plugins to act on prompts. Agents dynamically choose extensions and make repeated LLM calls, using prior outputs to guide subsequent actions for dynamic task execution.

LLM07:25

System Prompt Leakage

System prompt leakage occurs when sensitive info in LLM prompts is unintentionally exposed, enabling attackers to exploit secrets. These prompts guide model behavior but can unintentionally reveal critical data.

LLM08:25

Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities in RAG with LLMs allow exploits via weak generation, storage, or retrieval. These can inject harmful content, manipulate outputs, or expose sensitive data, posing significant security risks.

LLM09:25

Misinformation

LLM misinformation occurs when false but credible outputs mislead users, risking security breaches, reputational harm, and legal liability, making it a critical vulnerability for reliant applications.

LLM10:25

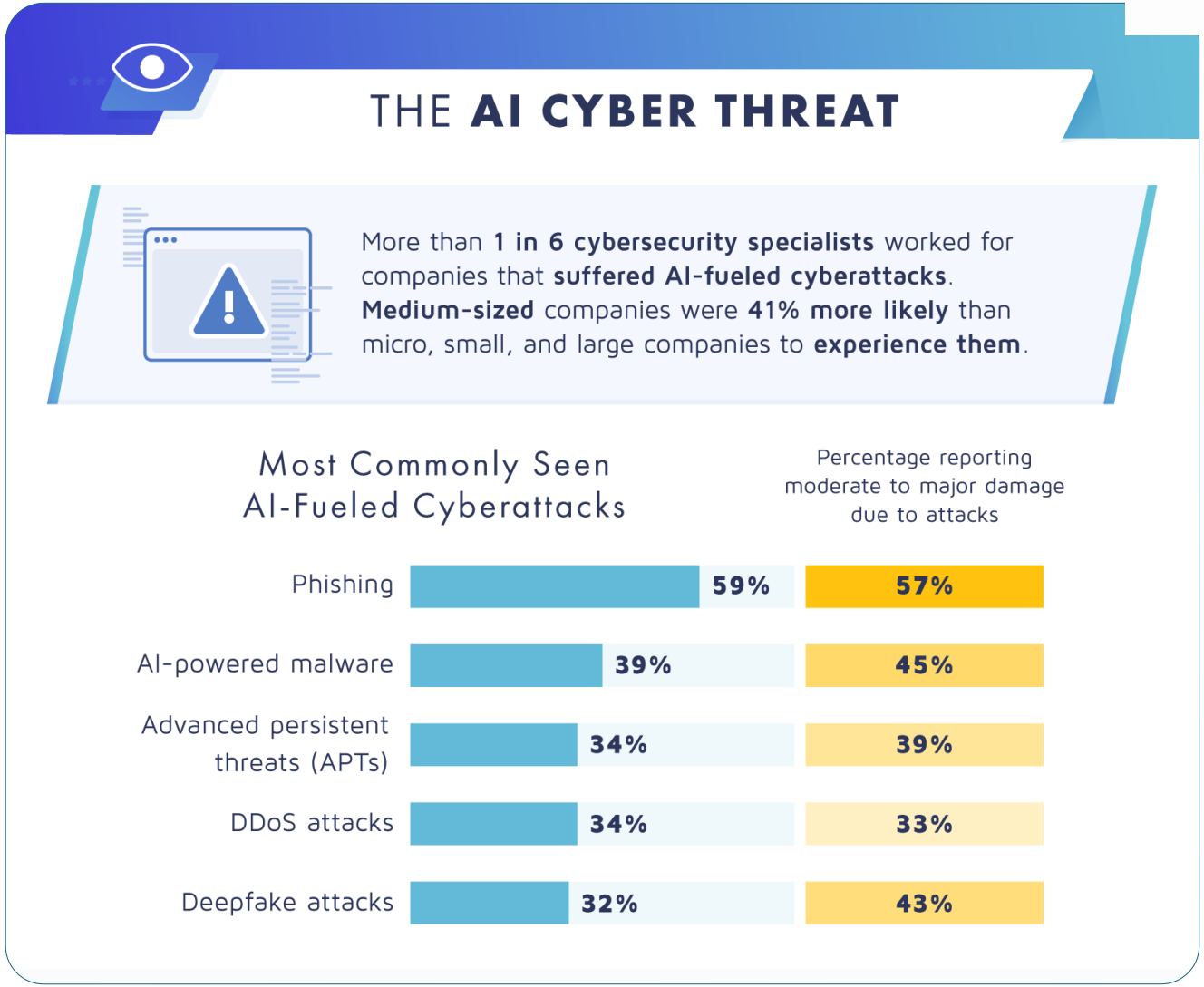
Unbounded Consumption

Unbounded Consumption occurs when LLMs generate outputs from inputs, relying on inference to apply learned patterns and knowledge for relevant responses or predictions, making it a key function of LLMs.

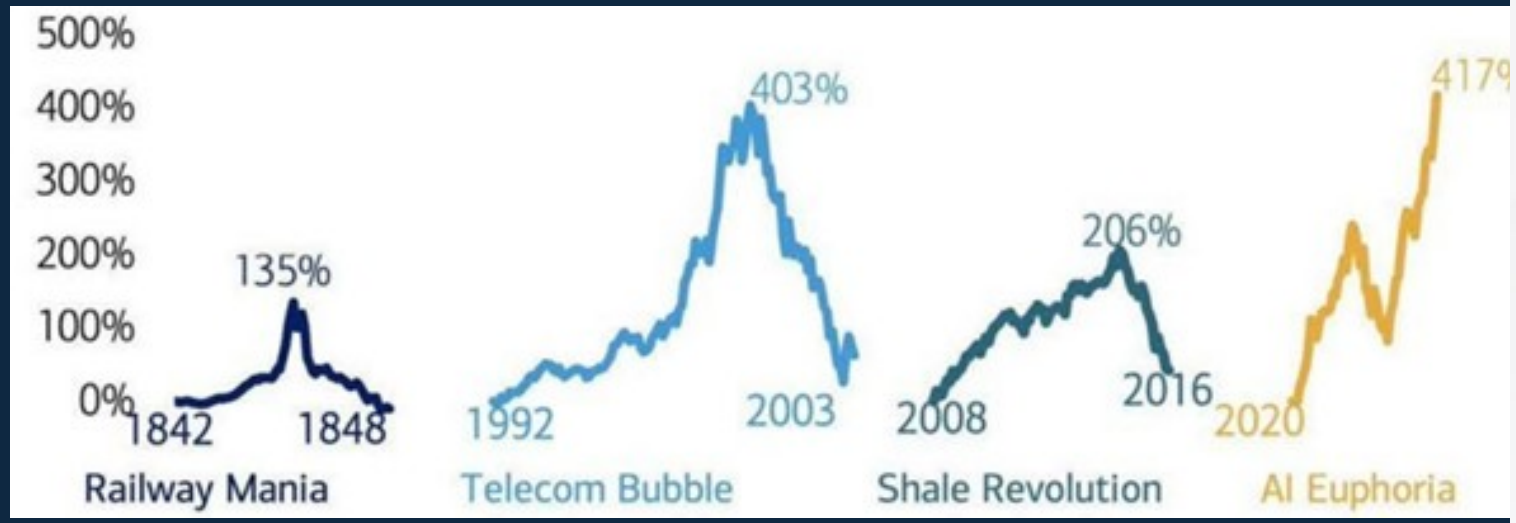
AI doing bad things

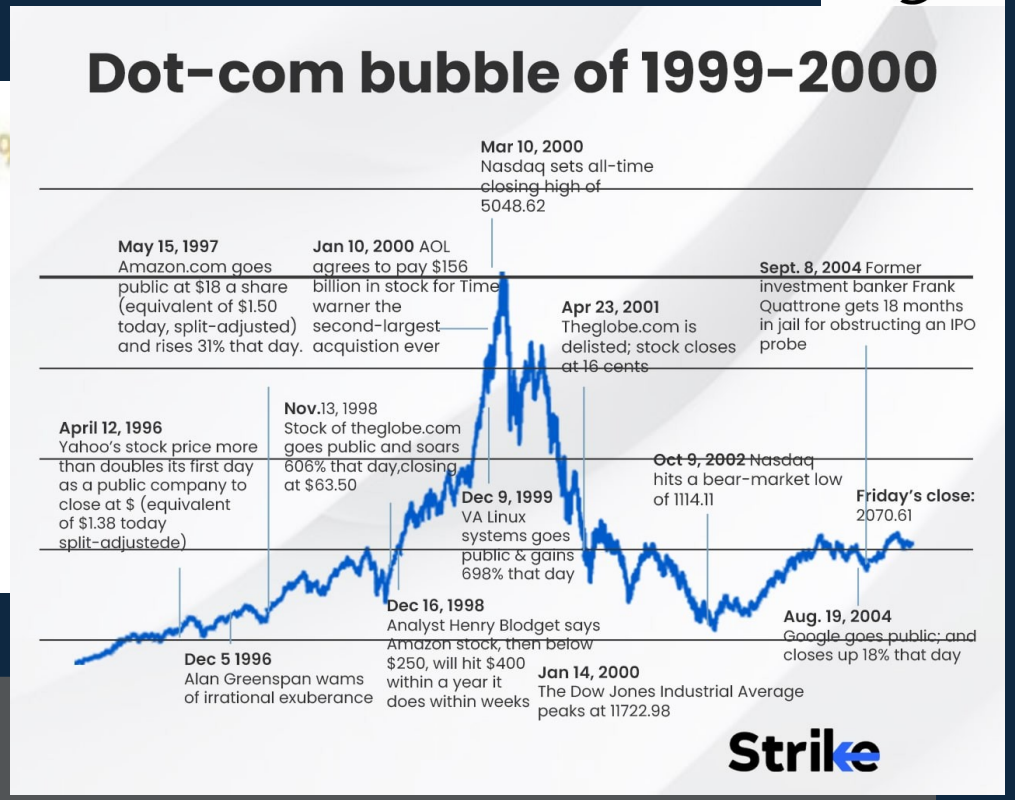
- Most Common Threats aided by AI
- https://www.digitalinformationworld.com/2023/08/a-growing-threat-how-ai-poses-risks-to.html

KOHRELL
Strategy and Delivery



Two Final Thoughts





Economic Risk – Too big to fail What if?

Impact on you

Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task△

Nataliya Kosmyna [1]
MIT Media Lab
Cambridge, MA

Eugene Hauptmann
MIT
Cambridge, MA

Ye Tong Yuan
Wellesley College
Wellesley, MA

Jessica Situ
MIT
Cambridge, MA

Xian-Hao Liao
Mass. College of Art and Design (MassArt)
Boston, MA

Ashly Vivian Beresnitzky
MIT
Cambridge, MA

Iris Braunstein
MIT
Cambridge, MA

Pattie Maes
MIT Media Lab
Cambridge, MA

United States

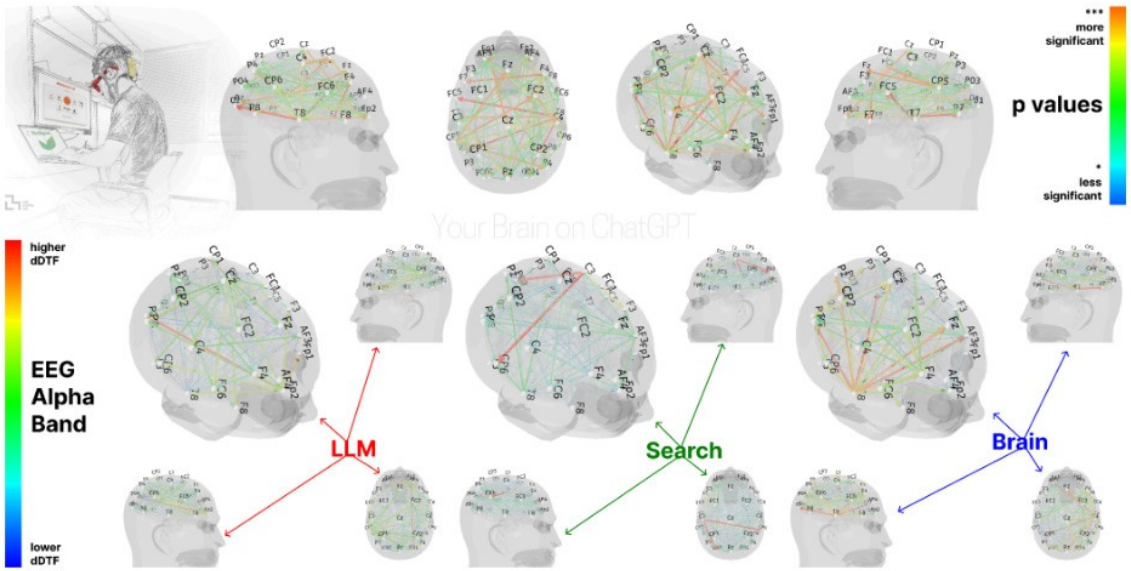
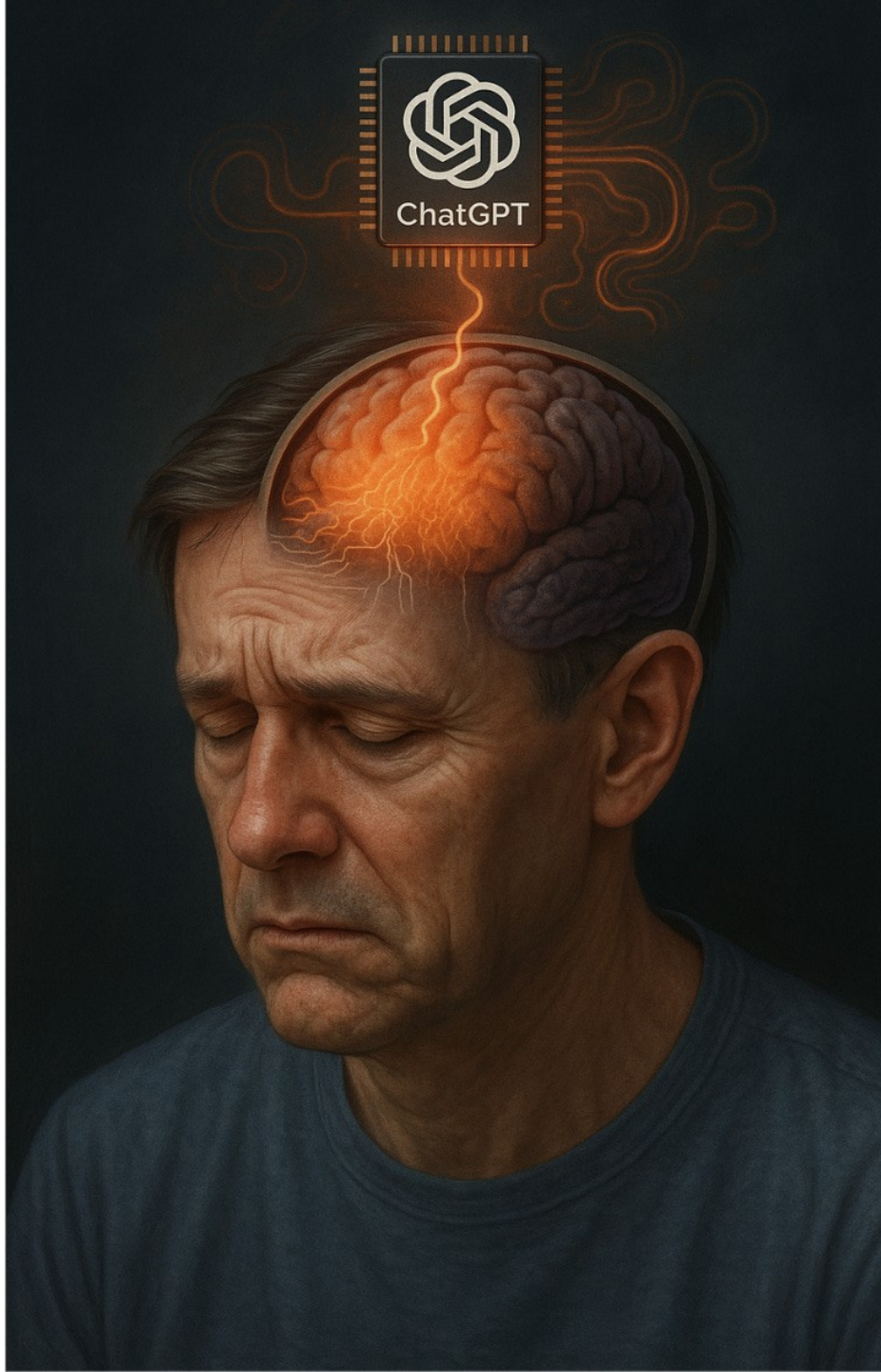


Figure 1. The dynamic Direct Transfer Function (dDTF) EEG analysis of Alpha Band for groups: LLM, Search Engine, Brain-only, including p-values to show significance from moderately significant (*) to highly significant (***).




NEbraskaCERT's
Wisdom

Contact

LINKEDIN – DKOHRELL

DAVID KOHRELL – DAVID@KOHRELL.ORG

HTTPS://WWW.KOHRELL.ORG

KOHRELL
Strategy and Delivery

AI is Risky Business