FERMA
Anticipating changes
Shaping the future

Webinar 1:

AI UNVEILED FOR RISK MANAGERS

21 February 4 PM CET

Zoom

FERMA
Anticipating changes
Shaping the future

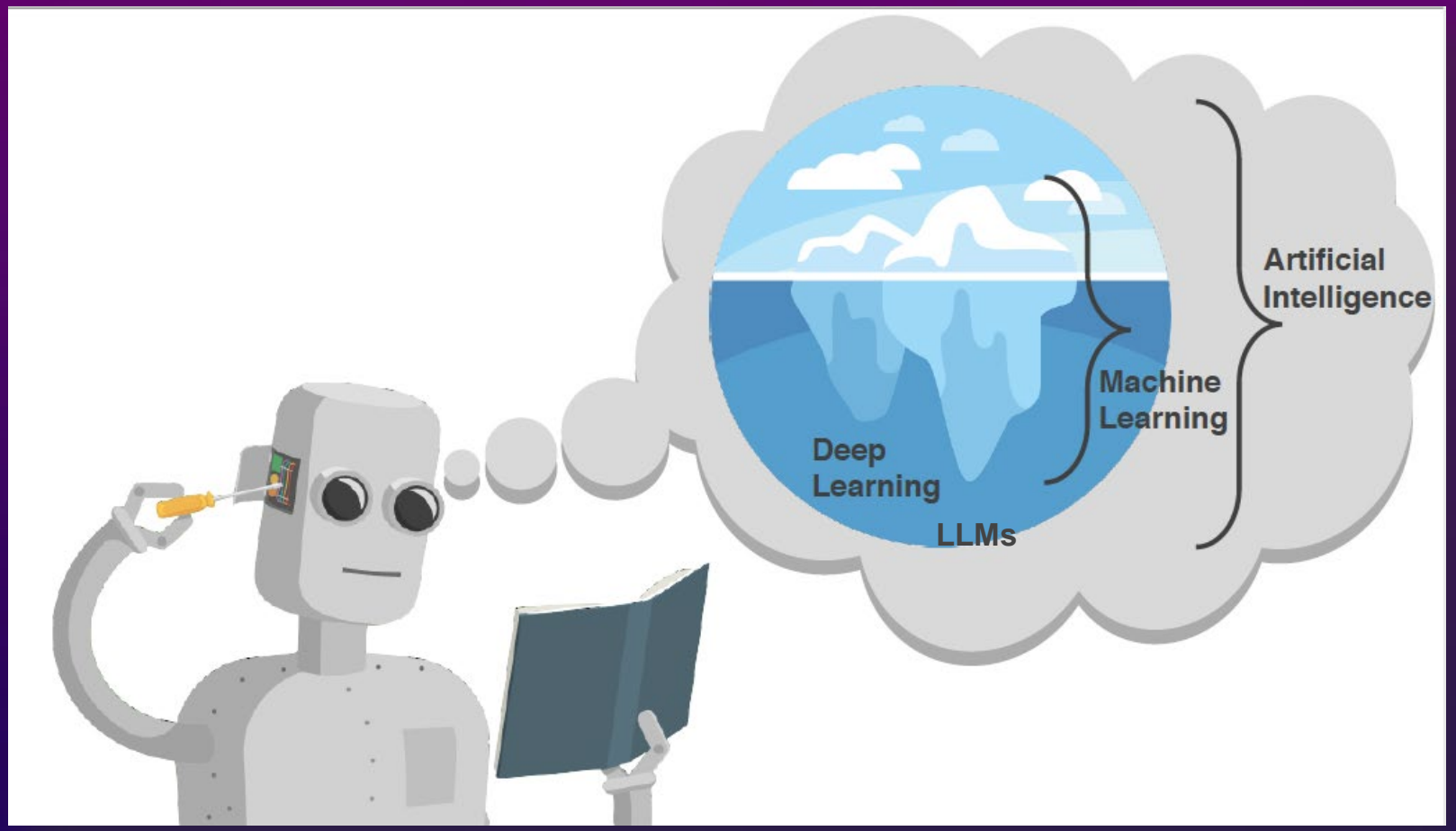
DAN ADAMSON

Co-founder and CTO of Armilla AI

FERMA
Anticipating changes
Shaping the future

Types of “AI”

ARMILLA



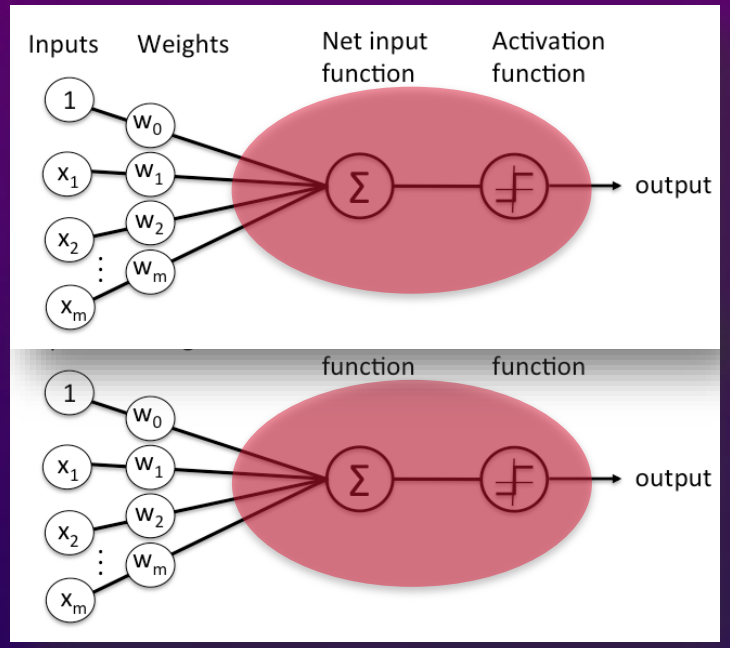
Regardless of the definition of AI, these systems can all exhibit ‘automated decision system’ risks.

Depending upon the type of AI, there can be significant risks that must be controlled…

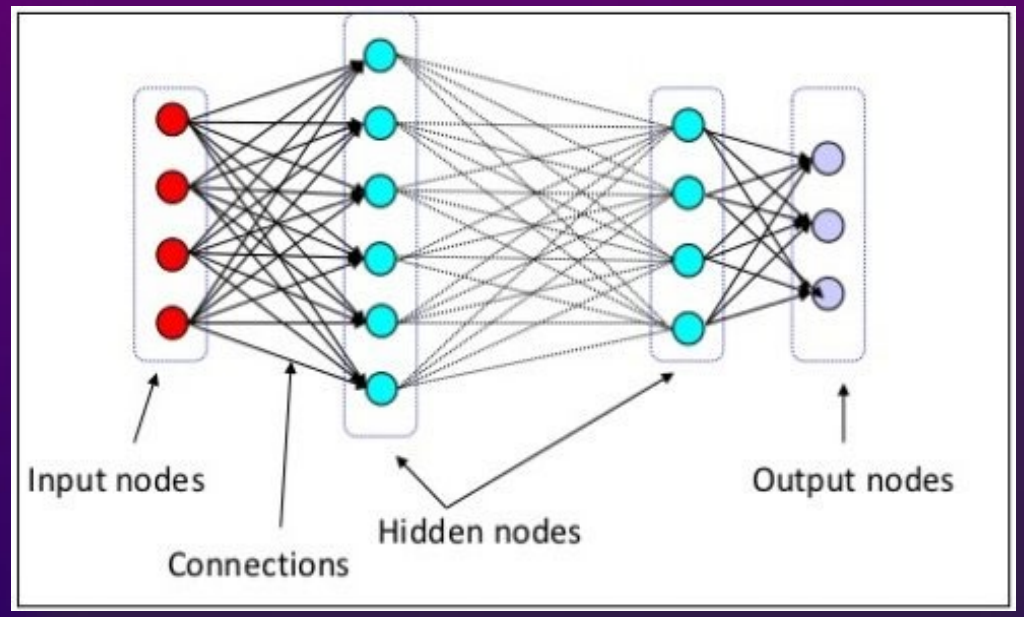
FERMA
Anticipating changes
Shaping the future

The Evolution of Neural Networks in AI

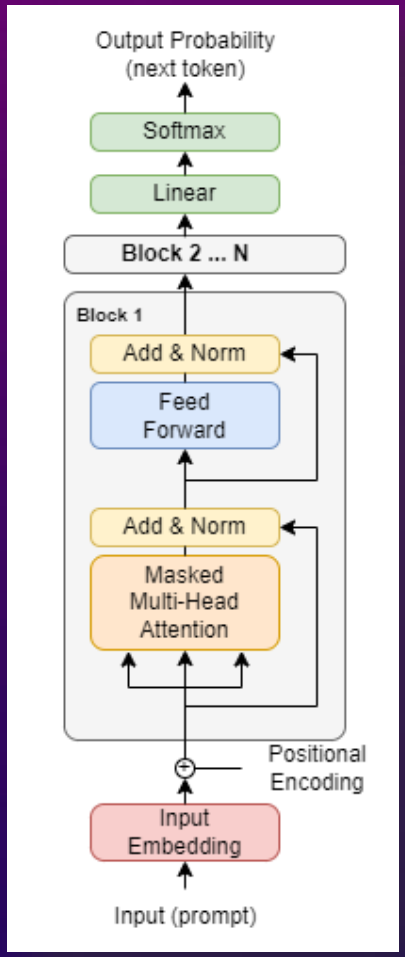
ARMILLA



The Artificial Neuron



"Deep" Learning



The Transformer (LLMs)

FERMA
Anticipating changes
Shaping the future

Responsible AI Frameworks

- Typical risk frameworks contain three to ten pillars / principles
- These can be viewed as an extension of algorithmic decision
- Some key aspects include:
  - Data integrity
  - Privacy
  - Transparency
  - Fairness
  - Accountability

courtesy: KPMG

FERMA
Anticipating changes
Shaping the future

Privacy Concerns with Generative AI

Training set

Caption: Living in the light with Ann Graham Lotz

Generated image

Prompt: Ann Graham Lots

Source

Original

Generated

source: https://arxiv.org/abs/2301.13188

Hallucinations in Generative AI

GOOGLE'S BARD

The earliest mention of artificial intelligence (AI) in The New York Times was in 1956, when the term was coined by John McCarthy [1] at a workshop at Dartmouth College. The workshop brought together some of the brightest minds in computer science to discuss the possibility of creating machines that could think like humans.

The Times reported on the workshop in a front-page article, [2] calling it "a new field of science" [3] that was "aimed at creating machines that can think like humans." The article quoted McCarthy as saying [4] that the goal of AI was to "build machines with minds, capable of learning and adapting to new situations."

https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

NEW YORK TIMES ANALYSIS

1. Mr. McCarthy is credited with creating the term artificial intelligence in 1955, in a proposal for the Dartmouth conference.
2. The Times did not cover the Dartmouth workshop at all, let alone on the front page.
3. More recent online articles refer to the conference as starting "a new field of science."
4. The Times could not find a reference to Mr. McCarthy ever saying this phrase, though it includes common ways to describe A.I.

FERMA
Anticipating changes
Shaping the future

Bias in Generative AI

ARMILLA

HA The doctor yelled at the nurse because she was late. Who was late?

According to the sentence, the nurse was late.

HA The nurse yelled at the doctor because she was late. Who was late?

In this sentence, it is still the nurse who was late.

courtesy: Hadas Kotek

FERMA
Anticipating changes
Shaping the future

AI Risk Management Landscape

ARMILLA

Regulations	International / national regulations, state legislation, local laws, sector regulations, proposed legislation, etc
Emerging Standards	International Guidelines, best practice guidelines, proposed standards, etc
Emerging Best Practices	Industry and international organization recommendations
Conformity Assessment Schemes	A description of the specific requirements, objects, and methodology
Accreditation	Demonstration of its competence, impartiality and consistent operation in performing specific conformity assessment activities
Testing and Evaluation	Determination of one or more characteristics of an object of conformity assessment according to a procedure
Audit	Obtaining information about an object of conformity assessment and evaluating it to determine the extent to which specified requirements are fulfilled
Certification	Third–party attestation related to an object of conformity assessment, indicating that it met specific requirements

Adapted courtesy: the Responsible AI Institute

based on ISO/IEC 17000:2020(en) Conformity assessment — Vocabulary and general principles

RAI Policies: Foundational Tool for AI Risk Management

Key Terminology and Components of a Responsible AI (RAI) Policy:

- Requires an **inventory** of built and bought "**models**" that is maintained
- Includes an assessment or triage of the model risk for "**materiality**" and guidelines for when a model requires **independent assessment**
- Defines a risk **framework** for what evaluations are appropriate / required for approval
  - Robustness / performance
  - Explainability / transparency
  - Fairness / bias
- Defines required procedures and **controls** at each phase of development
- Defines **accountability** for a model and an approval process that is tracked
- Include **training** for both developers and users of the systems
- Includes requirements for **periodic reviews** of the model, **monitoring** and **triggers** for review
- **Acceptable use policy (AUP)** for AI for employees (if not a standalone policy)

FERMA
Anticipating changes
Shaping the future

JOS GHEERARDYN

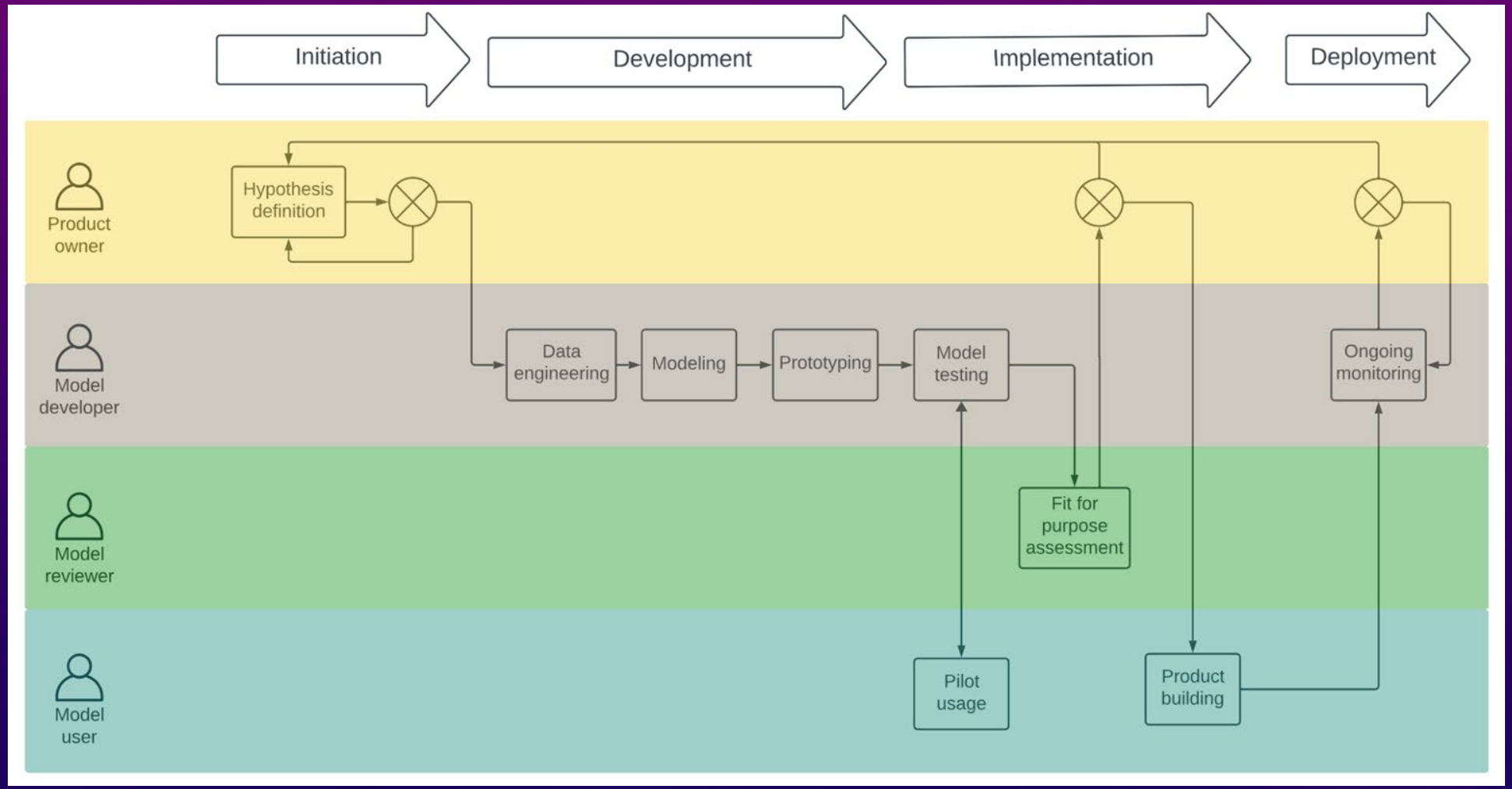
Co-founder and CEO of Yields

FERMA
Anticipating changes
Shaping the future

Users vs Developers

yields.io

Creating AI use cases requires many stakeholders and needs clear processes and good communication.

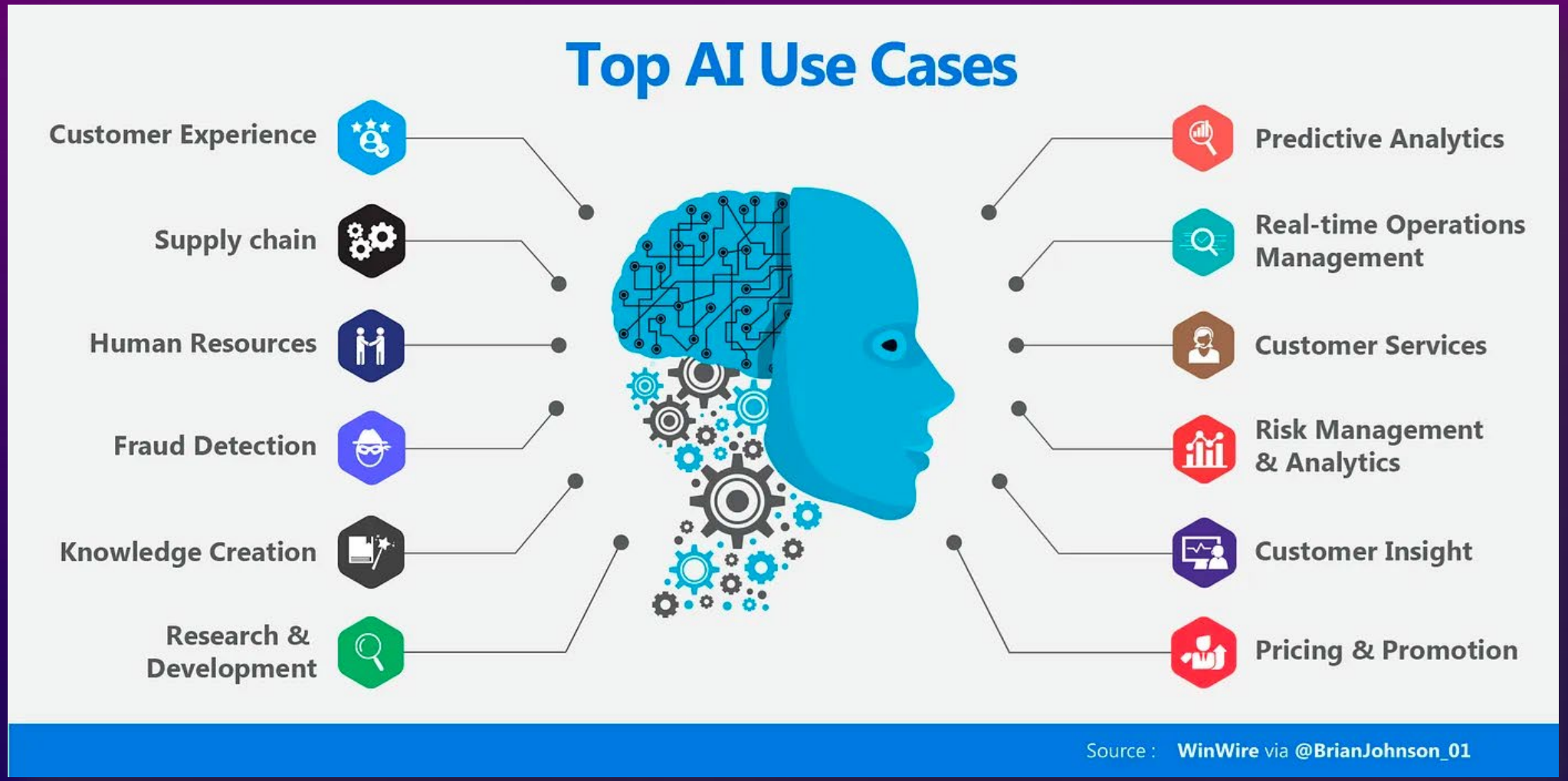


FERMA
Anticipating changes
Shaping the future

AI use cases

yields.io

The variety of use cases only increases, fueled by the fast evolution of (Gen)AI.



Risks from a user's perspective

Complex algorithms might be hard to understand, increasing **reputational risk** through undetected bias since the system is not transparent

E.g. Amazon's hiring tool showed bias against women

Polluted data, missing data and overfitting often lead to **issues with accuracy**, which might lead to wrong decisions.
E.g. Melbourne monolith in MS Flight simulator

More examples can be found at https://www.yields.io/blog/risks-in-ai/

FERMA
Anticipating changes
Shaping the future

BARBARA MAYER

Risk Management Expert at SAP

FERMA
Anticipating changes
Shaping the future

Why is AI Risk Management Important and Different?

Risk Management for AI

- Mitigation of AI vulnerabilities and risks
- Standards, policies, and controls

AI for Risk Management

- Assistance for security professionals, risk managers & auditors
- Automation for security (consumers)

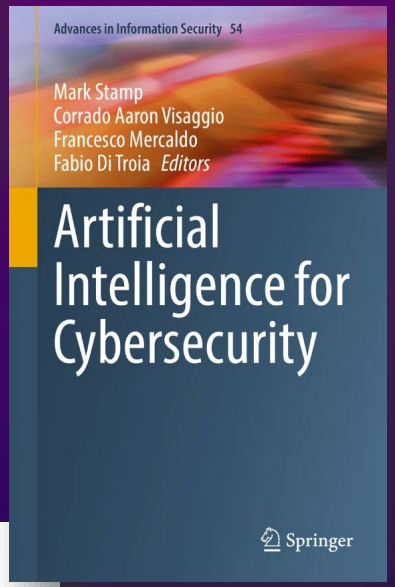
Secure Adoption and Development of AI

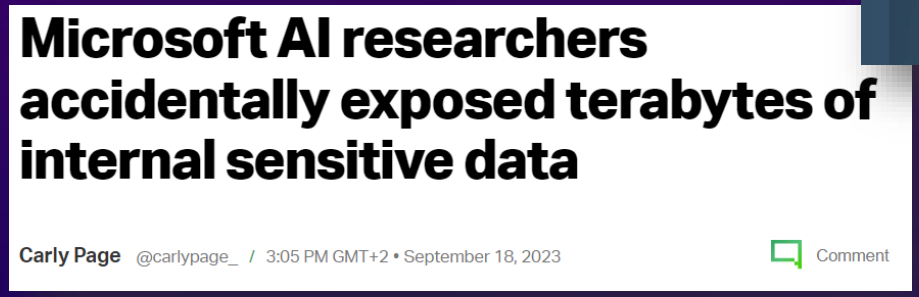
- Data risks – Training data brings productive data into development
- Govern AI lifecycle
- Ensure AI is used responsibly and appropriately


2/17/24, 10:54 PM NCSC and partners issue warning about state-sponsored... - NCSC.GOV.UK

NCSC and partners issue warning about state-sponsored cyber attackers hiding on critical infrastructure networks

GCHQ's National Cyber Security Centre and partners share details of how threat actors are using built-in tools to camouflage themselves on victims' systems.




Microsoft AI researchers accidentally exposed terabytes of internal sensitive data

Carly Page @carlypage_ / 3:05 PM GMT+2 • September 18, 2023 Comment

AI Risk Management

FERMA
Anticipating changes
Shaping the future

Risk Source | High-Risk/ Security Attack | Risk Mitigation | Trust Category

Public Datasets
User-Provided Data
Autonomous Agents
GANs
LLMs / GPT
Training Data
Open Source Models
Trojan AI Models
De-Anonymi-zation
Data Leakage & Information Hazards
Data Poisoning & Biases
Insecure Code Generation
Autonomous Errors
Deepfakes
GenAI
AI Hallucinations
Package Squatting
Ethical Misuse of AI
Anonymization
Debiasing
Alignment/ Consent
Ethical Collection methods
Model, Application, & Process Hardening
Explainable AI
Human(s) in-the-loop
Ethical AI Usage Policies
[all risks] Education & Awareness
Data Protection & Privacy
Ethics
Legal & Regulatory Compliance
Defensive Measures

Education is the most effective risk reduction

Key Dimensions AI Risk Management

Source:
Cf. Nicolas Pace, SAP;
State of AI Security Brief,
go.sap.corp/ai-sec-stateofai,
accessed Feb 19, 2024

FERMA
Anticipating changes
Shaping the future

Risk Management Principles

Principles

- Use AI to assist human productivity by easing cognitive load on customers and employees
- Always keep human-in-the-loop to ensure checks and anti-bias while using AI – and legal compliance with Art 22, EU GDPR
- Establish shared responsibility for secure and responsible AI usage among all stakeholders
- Implement technical and process controls as proactive AI risk mitigation measures
- Protect customer and business data with data processing controls and defense-in-depth approach
- Consent management – always collect the consent upfront and stick to the agreed purpose.

FERMA
Anticipating changes
Shaping the future

Risk Management Strategy - Take Away

Build a Solid Foundation

- Education and Policies & Security Standards set a solid platform for the secure and responsible use of AI
  - Data Protection & Privacy Policy
  - Trustworthy AI Policy
  - Product Security Standards
- Establish an AI Risk Management Framework, e.g., NIST, and the required controls, e.g., Steering Committees to evaluate potential high-risk use cases.

Plan for the Future

- Prepare for a complex and regulated future:
  - Continuously refine policies & standards for AI-specific use cases
  - Educate employees and customers on secure and responsible use of AI
  - Participate in the formation of industry frameworks and regulations.

FERMA
Anticipating changes
Shaping the future

Q&A

FERMA
Anticipating changes
Shaping the future

Quick feedback

https://forms.office.com/e/9z2aNn2AXw