Beyond the algorithm – realizing real value from AI

Seppe Housen

Guest Lecture VUB – Current Trends in AI

28th March 2026

datashift

About me

datashift

175 Datashifters

4 locations in Belgium & The Netherlands

40% year-over-year growth

VUB

Msc. Applied economics

Msc. Management

Msc. Criminology

AI is **useful** today

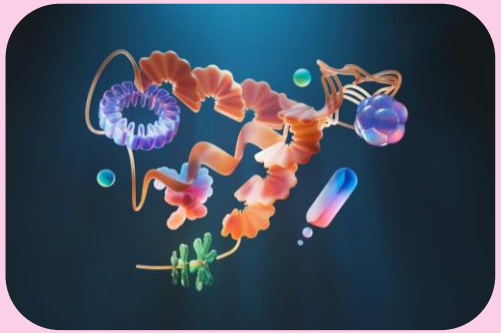
There's a lot AI still **can't do**

Scaling alone won't get us to ASI

AI is everywhere

GLOBAL BREAKTHROUGH

AlphaFold



DeepMind's AI solved a 50-year grand challenge in biology - predicting protein structures in hours instead of years.

200M+ proteins predicted

3M+ researchers worldwide

2024 Nobel Prize in Chemistry

WORKPLACE AI

Workplace Prevention

AI reshaping how companies prevent workplace incidents before they happen.

HEALTHCARE AI

Platelet Forecasting

Time series forecasting to predict blood platelet demand in hospitals.

LOGISTICS AI

Data Quality Control

Automated quality checks that catch logistics data errors instantly.

Thinks can go wrong when using AI

“Dogma is the enemy of progress.” – Einstein

Petra de Sutter, rector of the University of Ghent, used incorrect quotes in her opening speech of the academic year.

What went wrong?
She used AI to write her speech but the AI hallucinated these quotes and she didn’t validate the output.

An employee working on AI Safety at Meta tested the AI agent OpenClaw on her mailbox. She instructed to suggest deletions but to not execute them without approval. However, the agent deleted all mails older than a certain date without permission.

What went wrong?
The agent had too much autonomy (deleting rights) and there was a lack of reliable safety guards as the agent could not be stopped.

Dutch Tax authorities falsely accused thousands of parents of childcare benefits fraud. This led to debts, long-term financial instability, depression, and extreme stress for many families. Some families also lost their jobs or homes, or had children placed in foster.

What went wrong?
They used an automated risk profiling system that contained a biased model. There was also lack of transparency and human oversight even though the system had a substantial impact on people’s lives.

Creating value while managing risks with responsible AI

AI is valuable

New opportunities for automation

Enhanced decision-making

Democratized access to expertise

AND

AI creates new risks

Incorrect or harmful outputs

Unsafe usage and security threats

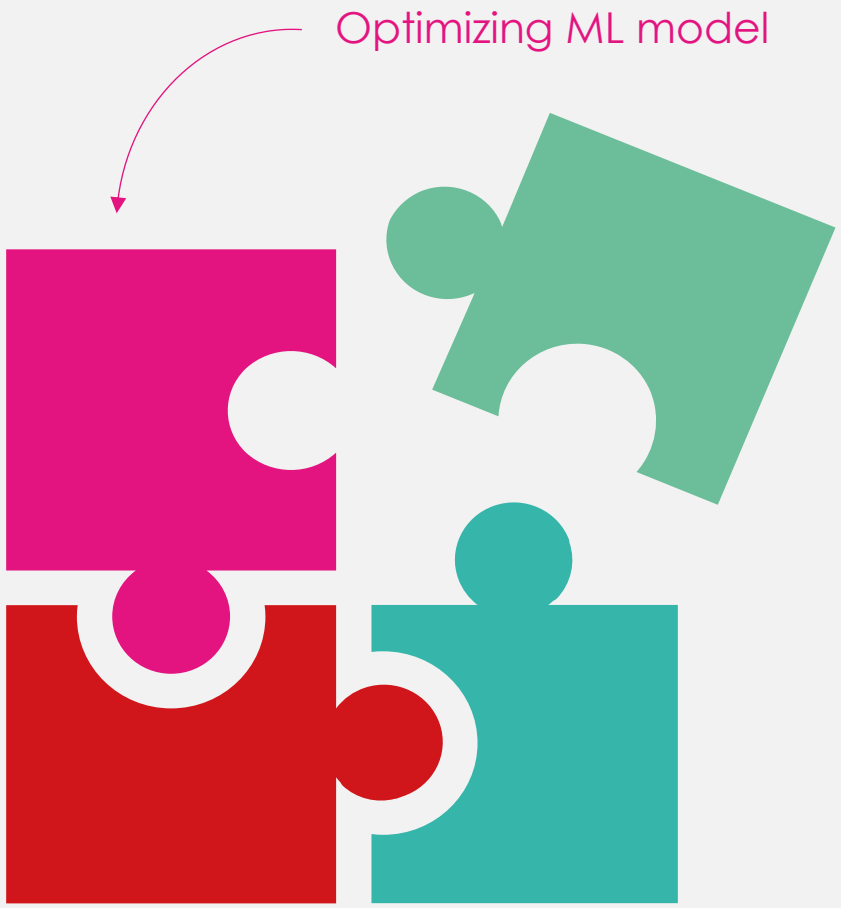
Unintended societal impact

Responsible AI

Enables AI development that maximizes value while managing risk and avoiding harm

Current Trend in AI

Responsible AI requires us to broaden our scope



Should we build this AI system?

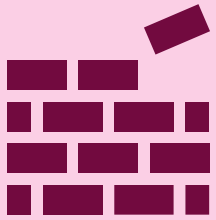
How should we build this AI system?

How do we do this **consistently** for all our systems?

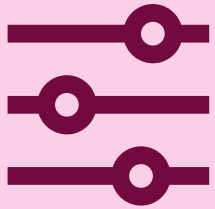
What we will cover today

Responsible AI

Current Trend in AI



Why now?

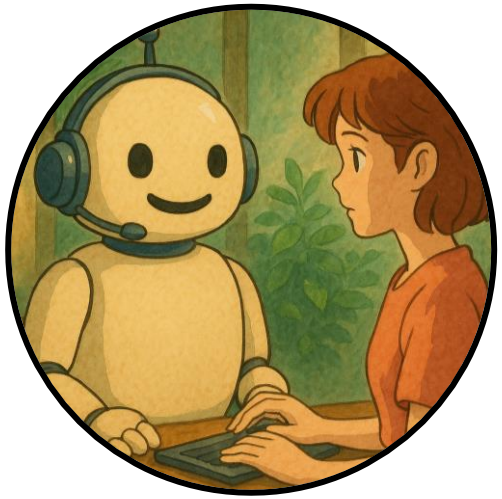


A Framework for AI Risk Management

Doing it at scale

WHY NOW

Four trends increase the need for responsible AI

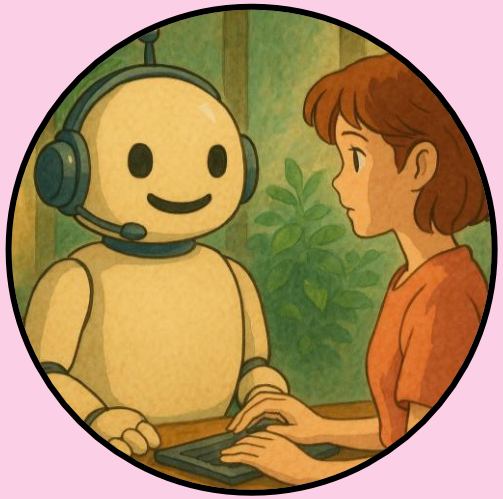


Generative AI

AI Act

Automation

Increased Access



Generative AI

A lot of new problems stem from the fact that we do not have access to the training data

Generative AI creates new risks

PREDICTIVE AI	GENERATIVE AI
Primarily supervised learning for predictions and classifications with **clear input-output** relationships.	**Black-box models** with unpredictable impacts of prompting, finetuning and RAG.
Production data closely **resemble training** data.	Production data may have **no relation** at all **to** the **training data**.
Accuracy and drift give **confidence in** model **performance**.	**Evaluating** open-ended generated outputs is an **unsolved problem** that requires human oversight.
Models are mainly deployed in highly **structured contexts.**	GenAI creates **new threats** such as harmful content, adversarial prompts, jailbreaking and copyright infringements.
Processing labelled data takes **time** and allows **for testing and evaluation** in overall timeline.	Value at stake and short prototype timelines create more **pressure to deploy quickly.**

AI Act

The AI Act focuses on regulating AI uses that could harm people

Risk categories for AI systems
- Prohibited AI systems
- High risk AI systems
- Limited risk AI systems
- Minimal risk AI systems

Important **entity types**
- Provider: developer and user of an AI system
- Deployer: user of an AI system

General Purpose AI (GPAI)
- Provide documentation, instructions, summary of training data and comply with copyright
- Open-source models only need to publish training data and comply with copyright
- Systemic risk requires evaluations, adversarial testing, cyber security and incident monitoring and reporting

The AI Act adopts a phased and risk-based approach to implementation

1 august 2024
Application of the AI Act[1]

2 February 2025
Prohibited AI & literacy

2 august 2026
High risk AI
(except safety components)

2 august 2027
All high risk AI

Risk categories for AI systems
- Prohibited AI systems
- High risk AI systems
- General purpose AI systems
- Minimal risk AI systems

Important **entity types**
- *Provider*: developer and user of an AI system
- *Deployer*: user of an AI system

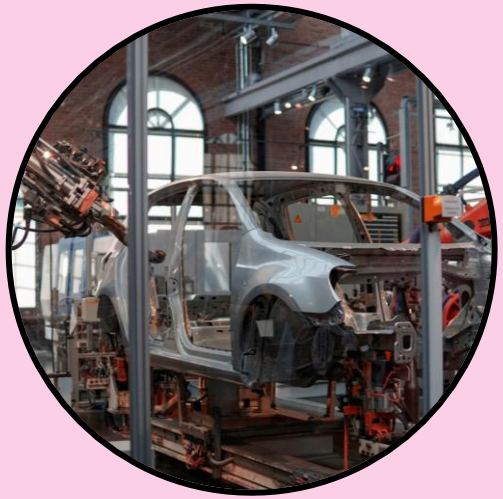
REQUIRED ACTIONS

- **Identify and assess** your current and future AI systems
- **Discontinue and prevent** any prohibited AI use-cases
- **Provide training** for users of AI systems

REQUIRED ACTIONS

- **Assess** current and future AI systems
- **Implement requirements** for high-risk use-cases:
  - AI risk management
  - Governance
  - Documentation & record-keeping
  - Transparency & human oversight
  - Accuracy, robustness and cybersecurity assurance

1: For each risk category and entity type, specific requirements are outlined, gradually taking effect over the next three years
2: https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal



Automation

As automation increases, so does the severity of potential harms

Increased automation in Predictive AI

Customer segmentation that is ran ones a month to identify relevant clusters	Automated pipeline that decides which add a specific customer sees and will retrain when performance drops below the threshold

From chatbots to Agentic AI

Internal chatbot used to translate and rephrase text	AI agent that handles customer complaints by analyzing them and answering to emails with the possibility of refunding purchases, giving discounts or refusing a claim

Increased access amplifies risks

Increased Access

AI used to be an expert field..

- PHD's and technical experts
- Understood the mathematics and methods these systems
- Running on specialized infrastructure

..Now it's accessible to all

- Easy access via no-code tools
- Limited or no understanding of these systems
- Tools embedded in everyday apps

Linda from HR: " I vibe coded this CV-screening tool!"

FRAMEWORK

Three High-Level Steps in AI Development

Design

- Problem framing
- Architecture
- Project plan

Build

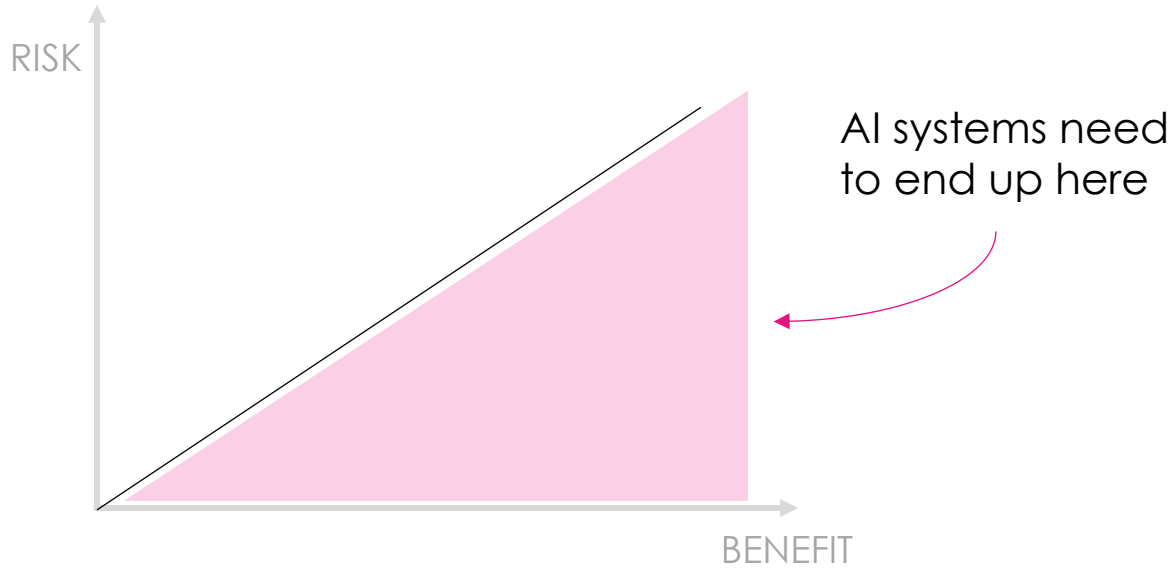
- Tuning
- Training
- Validation

Operate

- Integration
- Use
- Monitoring

Make deliberate trade-offs between risks and value creation

- **Define AI risks** and benefits explicitly to ensure **informed decision-making**
- **Manage AI risks** proactively to keep them **within acceptable boundaries**
- **Align AI risk-taking** with value creation to drive **lasting positive outcomes**



AI RISK MANAGEMENT 101

AI is a system that learns from input to generate output like predictions or content. Examples are chatbots, recommendation engines and predictive systems

Risk is the chance of a negative outcome, defined by the probability of harm occurring and the severity of that harm.

AI risk management is the process of identifying, measuring and reducing the risks associated with AI systems

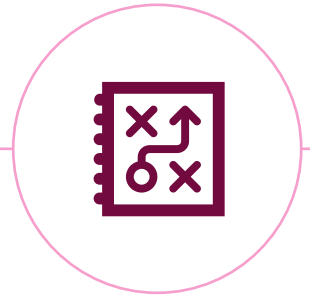
To handle risks effectively, you need to identify, measure and manage risks

IDENTIFY

Identify the AI risks that could exist at an organizational level and map them to individual use-cases

MEASURE

Measure, evaluate and monitor the risk by assessing the probability of occurrence and the potential impact of the risk



MANAGE

Manage the risk by making decision and taking actions to avoid, mitigate, transfer or accept the risk

Identify, measure and manage happen throughout the AI system lifecycle

	Design	Build	Operate
IDENTIFY	Identify the risks that are relevant to the system	Monitor the AI system for new risks	Monitor the AI system for new risks
MEASURE	Estimate the importance of each risk	Implement measuring methods to assess risk levels	Monitor the risk levels of identified risks
MANAGE	Plan the controls that need to be implemented	Implement the controls for the AI system	Decide on system changes
OUTCOME	**High level view on risk**	**AI system with acceptable risks**	**AI system that brings value**

Make strategic choices throughout the AI lifecycle

DECISION	How ambitions is the scope of our project?	How much do we invest in risk management?	Can we deploy the AI system?	Is the risk still within the acceptable threshold?
TRADE OFF	Ambition ↔ exposure	Risk mitigation ↔ budget	Benefits ↔ risks	Budget ↔ risk mitigation
ILLUSTRATIVE EXAMPLE	Internal chatbot before we deploy to customers	Purchase software to measuree bias in AI systems	High risk solution might be deployed as temporary solution to allow continuity	Model is retrained following performance decrease

Safeguard your AI System

1 Identify applicable risks

2 Evaluate risk probability and impact

3 Advice on risk mitigation and controls

CATEGORY	RISK IMPACT	SCORE
Accuracy	Customers are misinformed and dissatisfied. Inbound calls may increase.	M / H
Fairness	A subset of customers may receive lower quality results.	M / M
Safety	Unwanted purchase decisions after misleading or manipulative chatbot interactions	L / M
Security	System vulnerabilities inherent to chatbots; Chatbot becomes unavailable	M / M
Compliance	A complaint to the DPA about the legal basis or data subject rights	L / L
Environment	Chatbot release puts pressure on indirect emission targets	L / L

RECOMMENDED CONTROLS

- RAG
- Output parameter tuning
- System prompt
- Control user input
- Control system output
- Release management
- Fast rollback procedures
- …

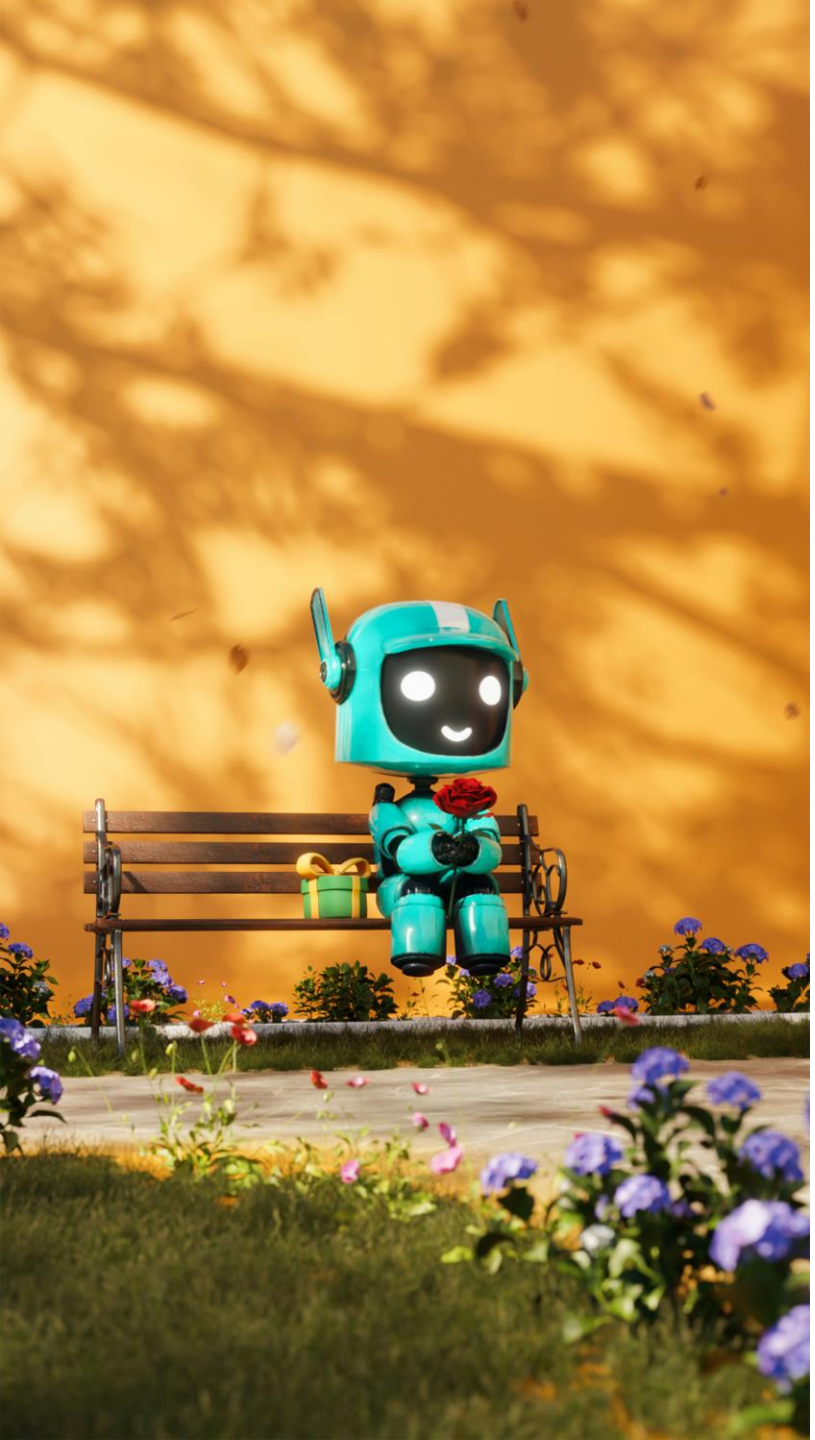
RECOMMENDED MEASUREMENTS

- Performance testing
- User testing
- Adversarial testing
- …

Summary of risk events by category

Risk score at an individual and organizational level

Recommendation risk mitigation actions



Example – AI Chatbot

Performance measure
How do you measure how often the chatbot is correct?

Prompt injection detection
How do we make sure adversaries do not take control?

GDPR compliance
Within which perimeters can we process personal data?

Example – AI Tools Like Claude Code

RISKS

- Production issues due to incorrect code
- Compliance breach
- ESG target failure

ACTIONS

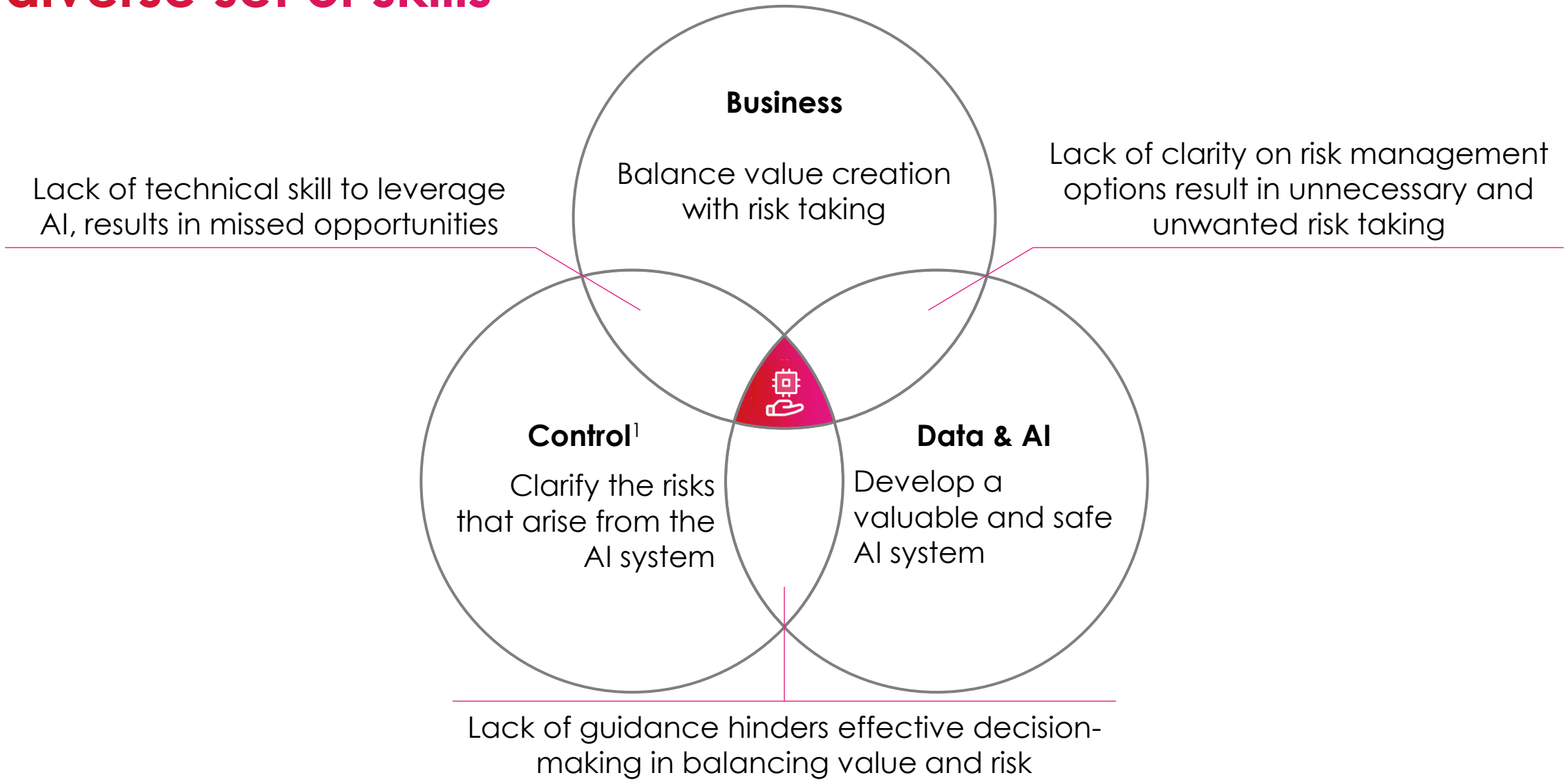
- Code review process
- Tool selection & user training
- ESG monitoring and smaller models

DOING IT AT SCALE

Not, one use-case, but…

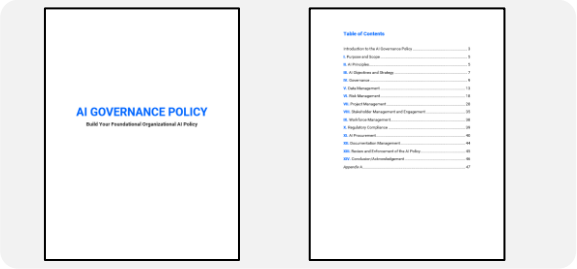
…100+ use-cases

The key challenge for responsible AI is bringing together a diverse set of skills



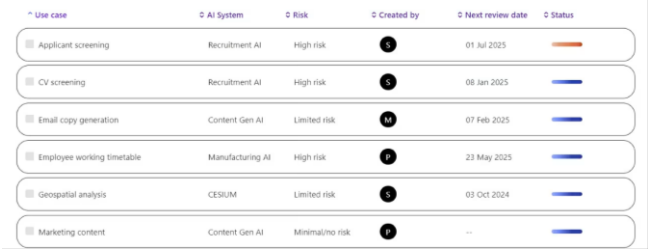
1 Control spans multiple functions, including legal, compliance, privacy, security, sustainability, quality and more

Building blocks

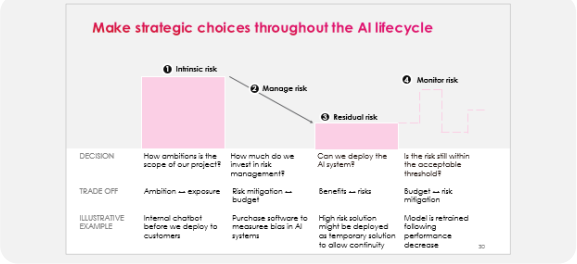


AI Governance Policy
Rules and accountability

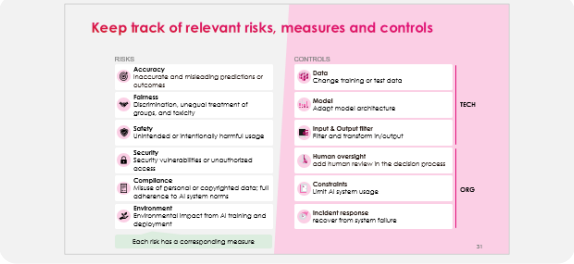
Responsible AI Principles
Guidance on AI use



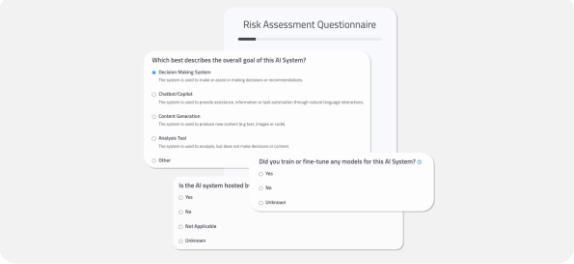
AI Inventory
Overview on AI projects



AI Governance Process
Manage AI risks



AI Risk Taxonomy
Clarify AI risks



Operational Templates
Standardize workflows

Make strategic choices throughout the AI lifecycle

DECISION	How ambitions is the scope of our project?	How much do we invest in risk management?	Can we deploy the AI system?	Is the risk still within the acceptable threshold?
TRADE OFF	Ambition ↔ exposure	Risk mitigation ↔ budget	Benefits ↔ risks	Budget ↔ risk mitigation
ILLUSTRATIVE EXAMPLE	Internal chatbot before we deploy to customers	Purchase software to measuree bias in AI systems	High risk solution might be deployed as temporary solution to allow continuity	Model is retrained following performance decrease

Keep track of relevant risks, measures and controls

RISKS

Accuracy
Inaccurate and misleading predictions or outcomes

Fairness
Discrimination, unequal treatment of groups, and toxicity

Safety
Unintended or intentionally harmful usage

Security
Security vulnerabilities or unauthorized access

Compliance
Misuse of personal or copyrighted data; full adherence to AI system norms

Environment
Environmental impact from AI training and deployment

Each risk has a corresponding measure

CONTROLS

Data
Change training or test data

Model
Adapt model architecture

Input & Output filter
Filter and transform in/output

TECH

Human oversight
add human review in the decision process

Constraints
Limit AI system usage

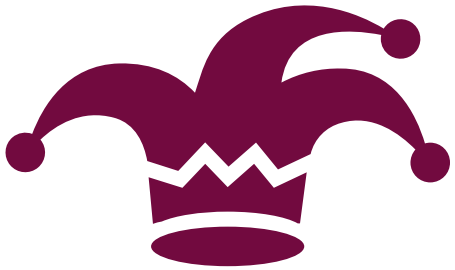
Incident response
recover from system failure

ORG

CLOSING

AI is a tool, a fool and a target

Tool



Fool

Target

Source: https://arxiv.org/abs/2402.12617