Navigating the AI Frontier:

Upholding Trustworthy and Responsible AI without Full Control

Prof. Liming Zhu

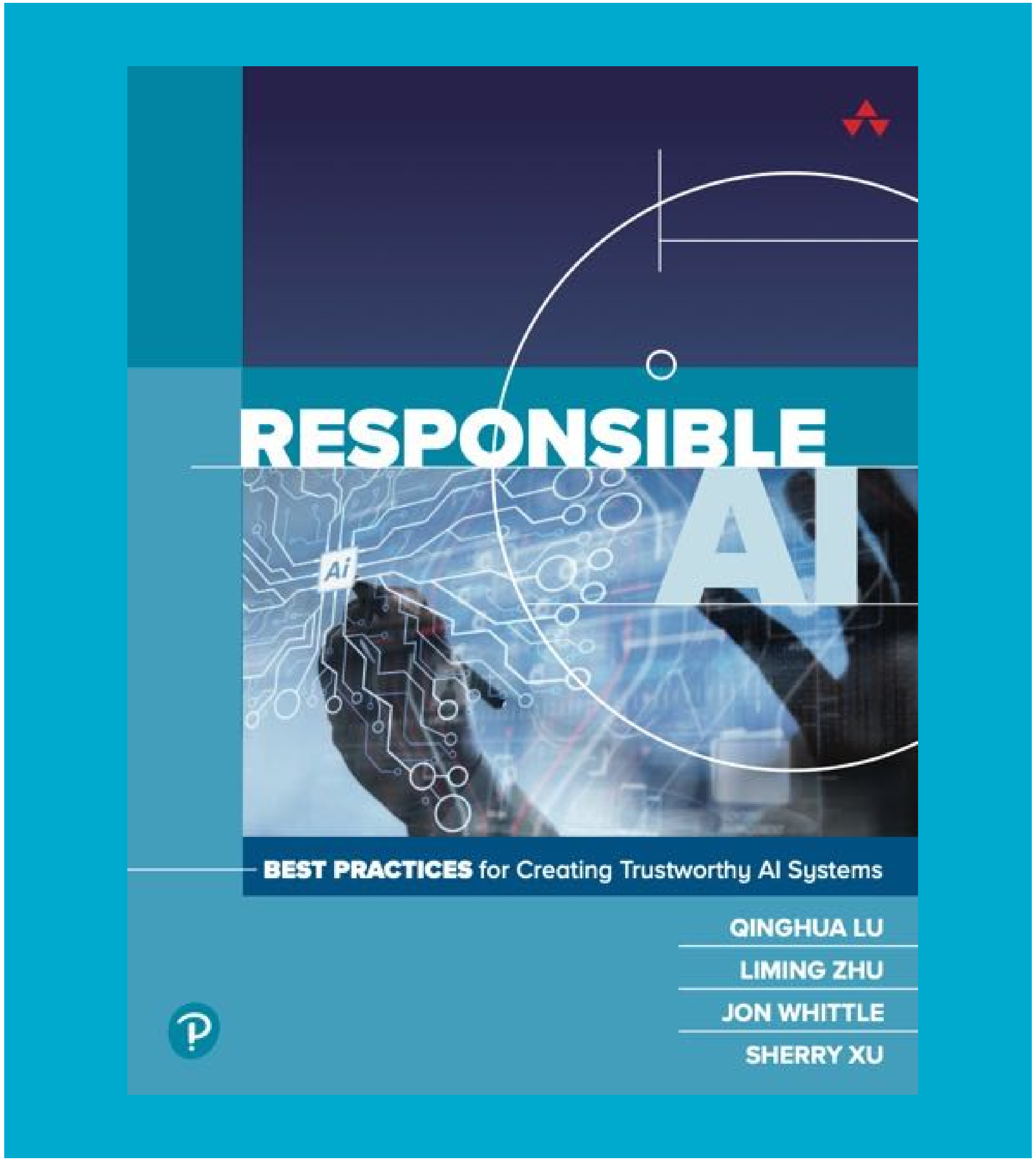
Research Director, CSIRO's Data61

Conjoint Professor, UNSW

Expert in Working Groups

- Australia's AI Safety Standard
- OECD.AI AI Risk and Accountability
- ISO/IEC JTC 1/SC 42/WG 3 – AI Trustworthiness

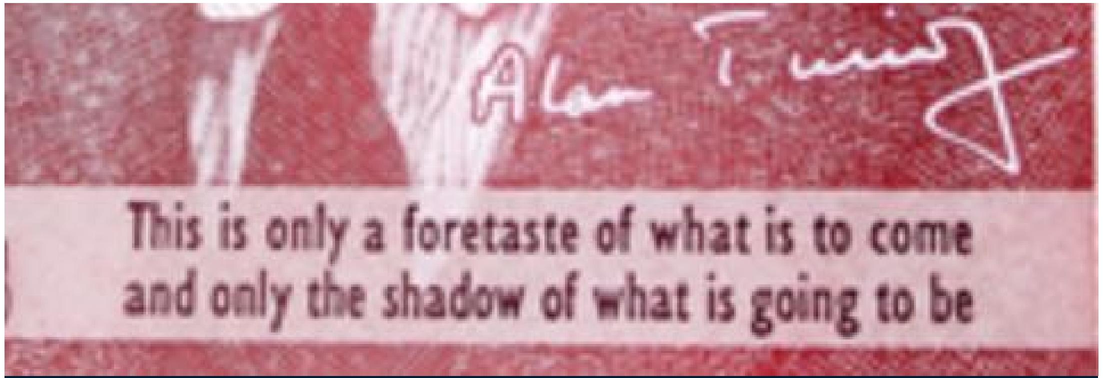
Australia's National Science Agency



Current State of AI

A Forecast…. and a Shadow

Bank of England
£50
Fifty Pounds
AA44 699671
This is only a foretaste of what is to come
and only the shadow of what is going to be


This is only a foretaste of what is to come
and only the shadow of what is going to be

We have to have some experience with the machine before we really know its capabilities.

It may take years before we settle down to the new possibilities, but I do not see why it should not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms.

Alan Turing, 1949

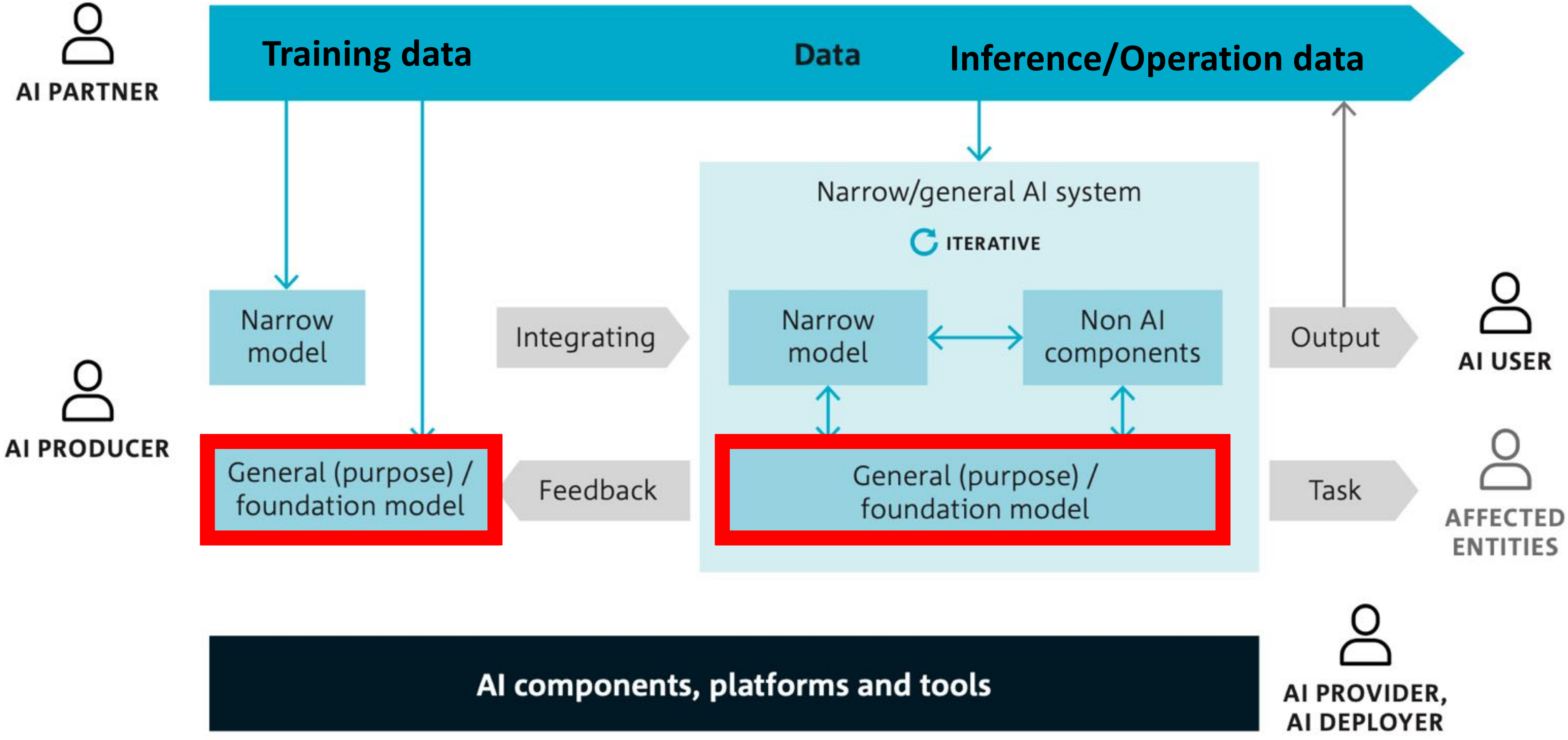
How Is AI’s Nature Shifting? - Updated Definitions

An ***AI system*** *is a machine-based system 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.*
- *Different AI systems vary in their level of autonomy and* **adaptiveness after deployment.**

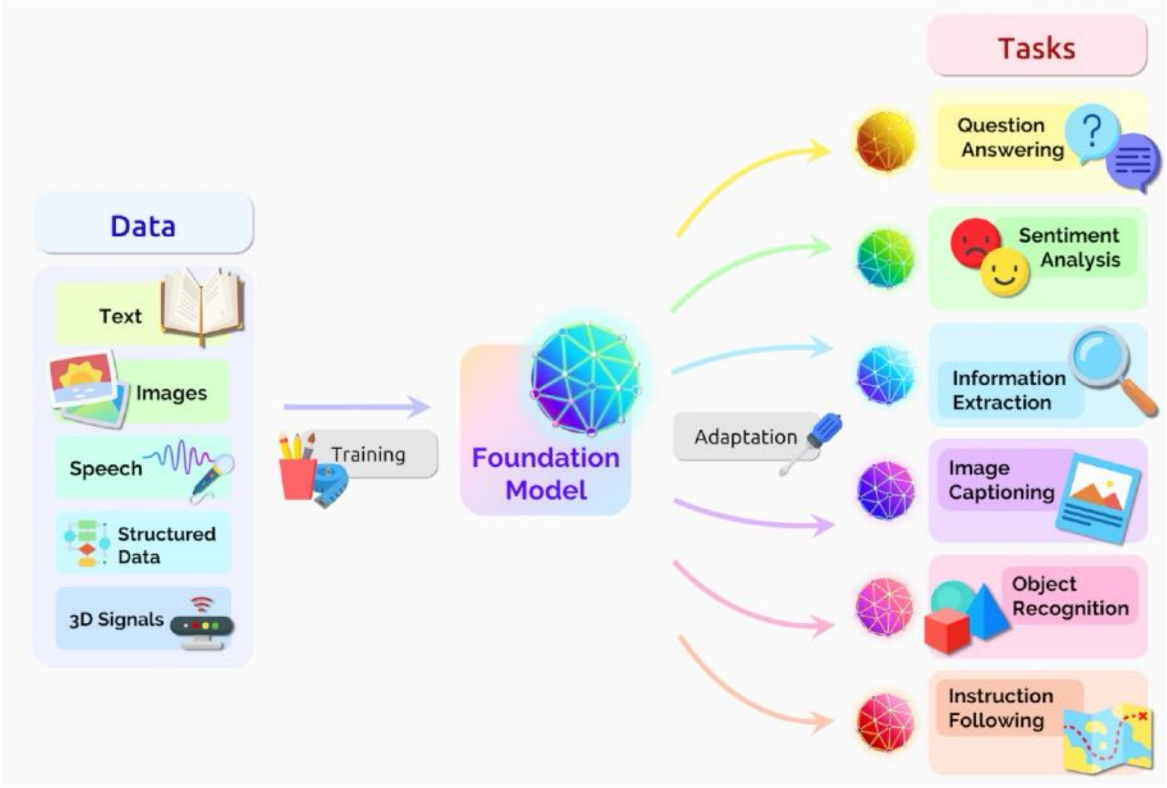
(OECD, 2023)

Clarifying AI: Model or System?



Trends & Challenges

Flipping the Script: General First, Specific Later

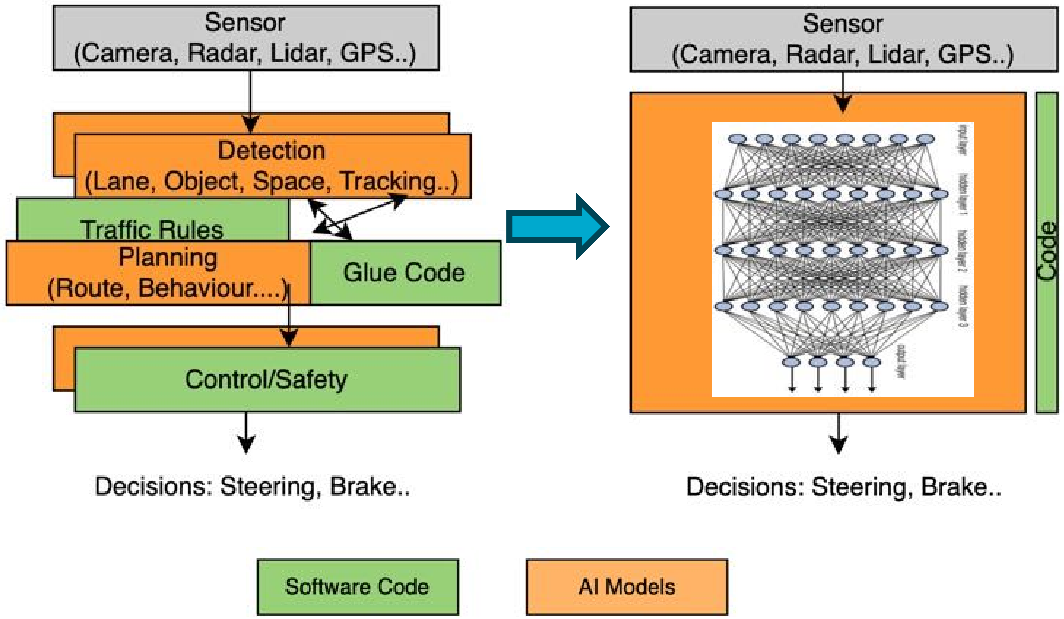


Generality is free?

Bommasani, R. et.al , 2022. On the Opportunities and Risks of Foundation Models.

Turning the Page: From Specific to End-to-End AI

End-to-End AI: Data In, Decision out, No Code


Sensor
(Camera, Radar, Lidar, GPS..)
Detection
(Lane, Object, Space, Tracking..)
Traffic Rules
Planning
(Route, Behaviour....)
Glue Code
Control/Safety
Decisions: Steering, Brake..
Sensor
(Camera, Radar, Lidar, GPS..)
Code
Decisions: Steering, Brake..
Software Code
AI Models

Where is human expertise here?

- ***No code/expert-derived smarts***, ”dumb” learning algorithm + big/synthetic data
- ***Non-domain experts*** improve learning efficiency

Rethinking Data and Expertise Value

Proprietary task data is more valuable.	General learning often outperforms domain-specific training.
AI can’t surpass the intelligence of its human training data.	Hard tasks can be mastered from easy task data and synthetic data.
AI only memorises, retrieves, and generates variants—it can’t reason	The line between retrieval and reasoning is blurred; humans often retrieve and apply familiar reasoning templates rather than true reasoning.

Reclaiming the Reins: Human Control Over End-to-End AI

Principles
Standards
Frameworks

Australia's AI ethics framework

OECD AI principles

EU AI Act

…

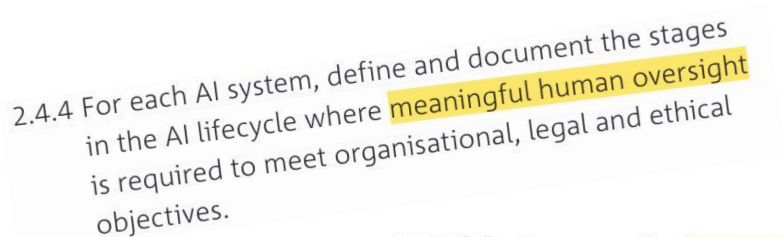
AU Safety Standard

ISO Standards

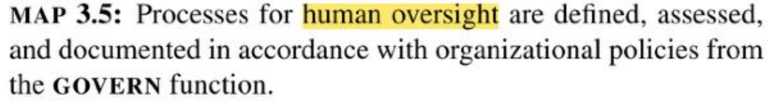
NIST AI RMF

Principles/Regulations/Standards != Eng. Practices

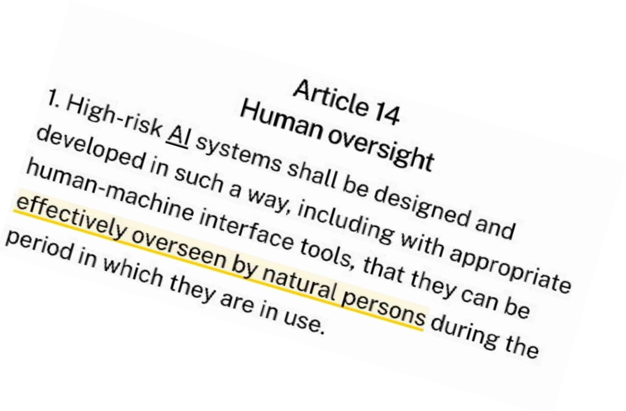
?



2.4.4 For each AI system, define and document the stages in the AI lifecycle where meaningful human oversight is required to meet organisational, legal and ethical objectives.



MAP 3.5: Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the **GOVERN** function.

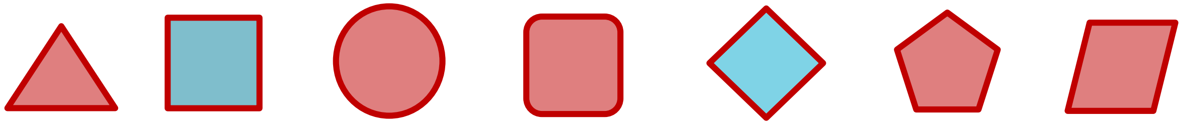


Article 14
Human oversight

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.

Model Alignment != System Alignment

Algorithms
Models

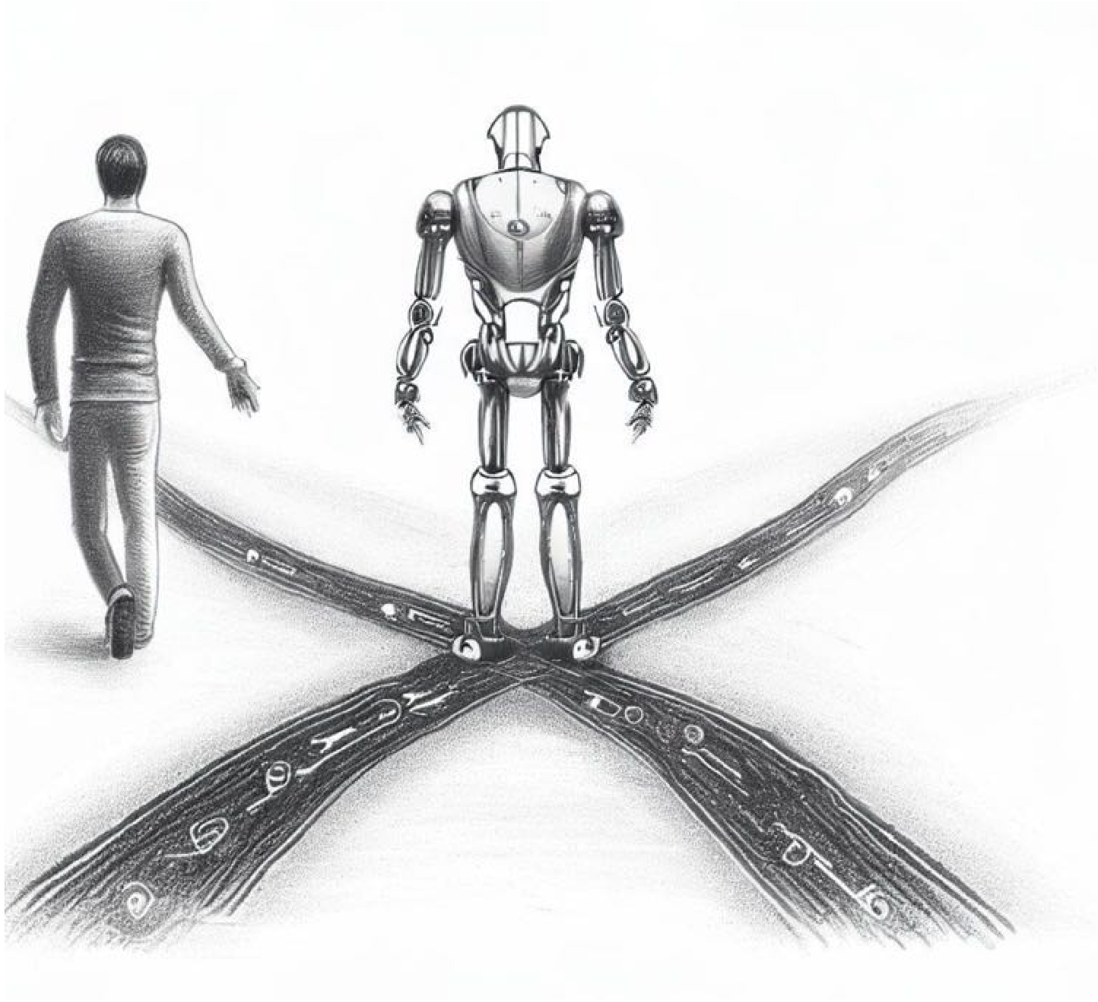


Lu, Q., Luo, Y., Zhu, L., Tang, M., Xu, X., Whittle, J., 2023. Operationalising Responsible AI Using a Pattern-Oriented Approach: A Case Study on Chatbots in Financial Services. IEEE Intelligent Systems.

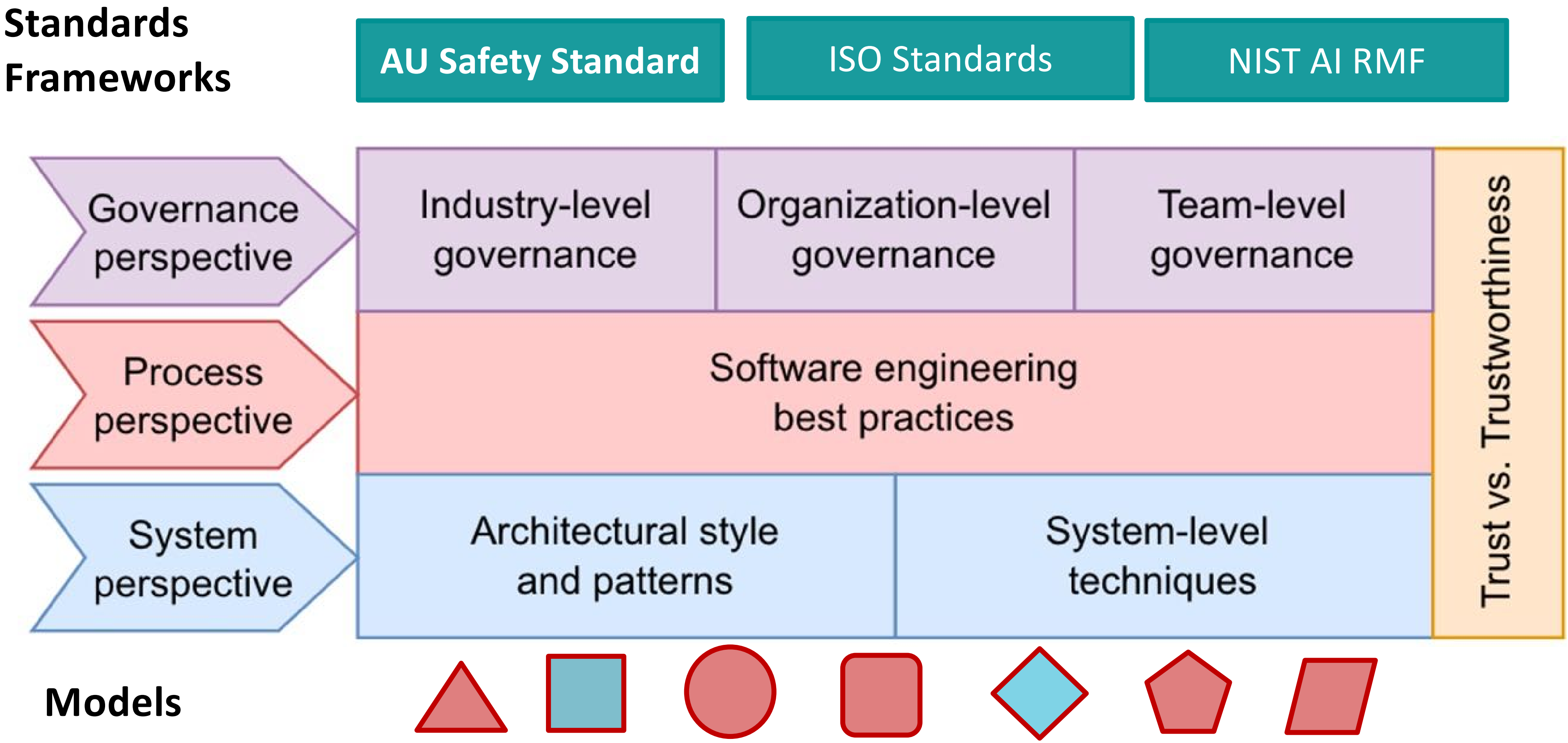
Rethinking Human Control

Human in the loop solves everything.	Humans can be liability sponges, especially without the right tools.
AIs are just tools; human oversight and decision is needed at every step.	AI can embed consistent agency, freeing humans to focus on critical and special cases.
AI always augments human capabilities, leaving interesting tasks to humans.	AI can lead to deskilling, automating the interesting and complex while leaving humans with the boring.

Directions & Questions



Design-time Human Control



Standards Frameworks

AU Safety Standard

ISO Standards

NIST AI RMF

Governance perspective

Industry-level governance

Organization-level governance

Team-level governance

Process perspective

Software engineering best practices

System perspective

Architectural style and patterns

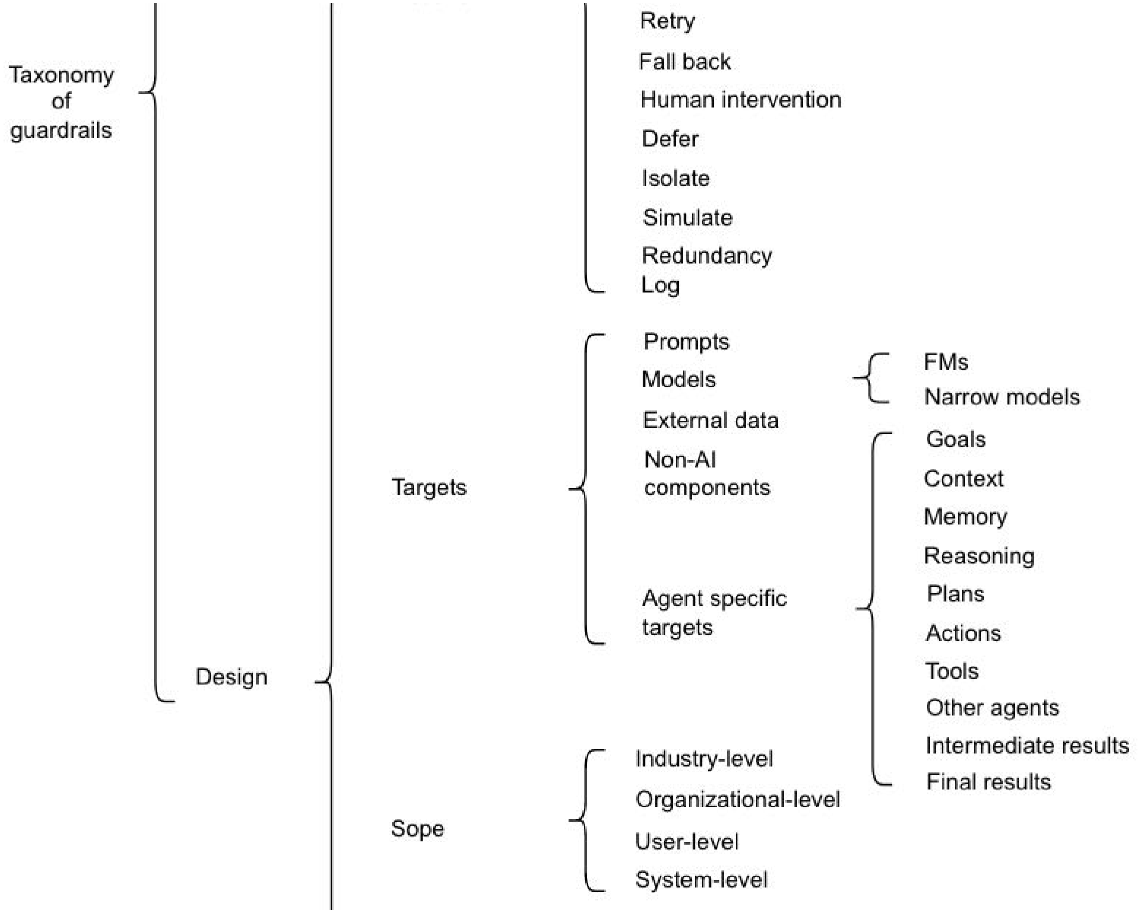
System-level techniques

Trust vs. Trustworthiness

Models

Lu, Q., Zhu, L., Xu, X., Whittle, J., Xing, Z., 2022. Towards a Roadmap on Software Engineering for Responsible AI, in: 1st International Conference on AI Engineering (CAIN)

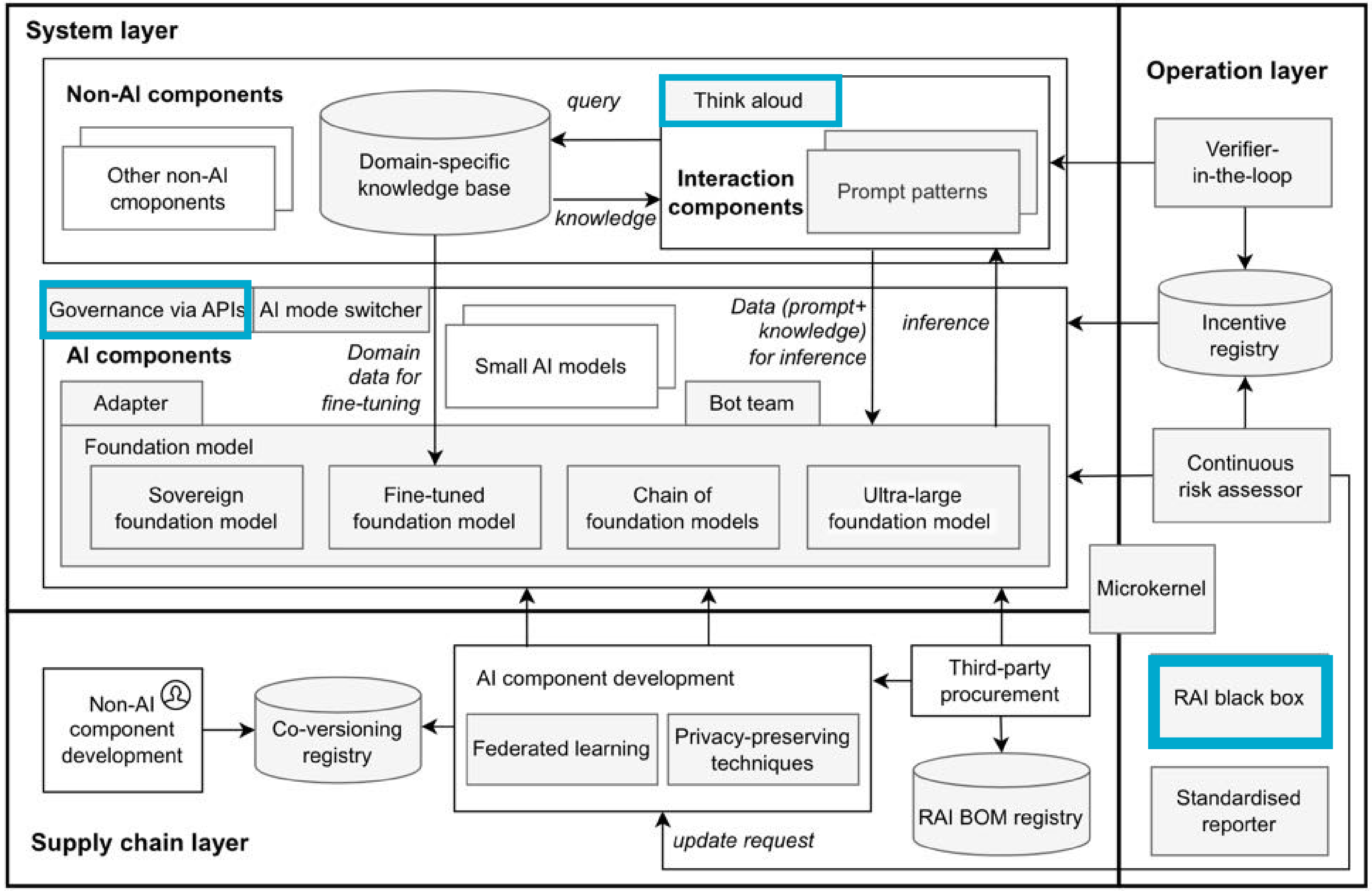
Runtime Control - Guardrails



- Taxonomy of guardrails
  - Design
    - Retry
    - Fall back
    - Human intervention
    - Defer
    - Isolate
    - Simulate
    - Redundancy
    - Log
  - Targets
    - Prompts
    - Models
      - FMs
      - Narrow models
    - External data
    - Non-AI components
    - Agent specific targets
      - Goals
      - Context
      - Memory
      - Reasoning
      - Plans
      - Actions
      - Tools
      - Other agents
      - Intermediate results
      - Final results
  - Sope
    - Industry-level
    - Organizational-level
    - User-level
    - System-level

Shamsujjoha, M. *et al.* (2024) 'Towards AI-Safety-by-Design: A Taxonomy of Runtime Guardrails in Foundation Model based Systems'. arXiv. Available at: https://doi.org/10.48550/arXiv.2408.02205.

Control of LLM-based AI Systems

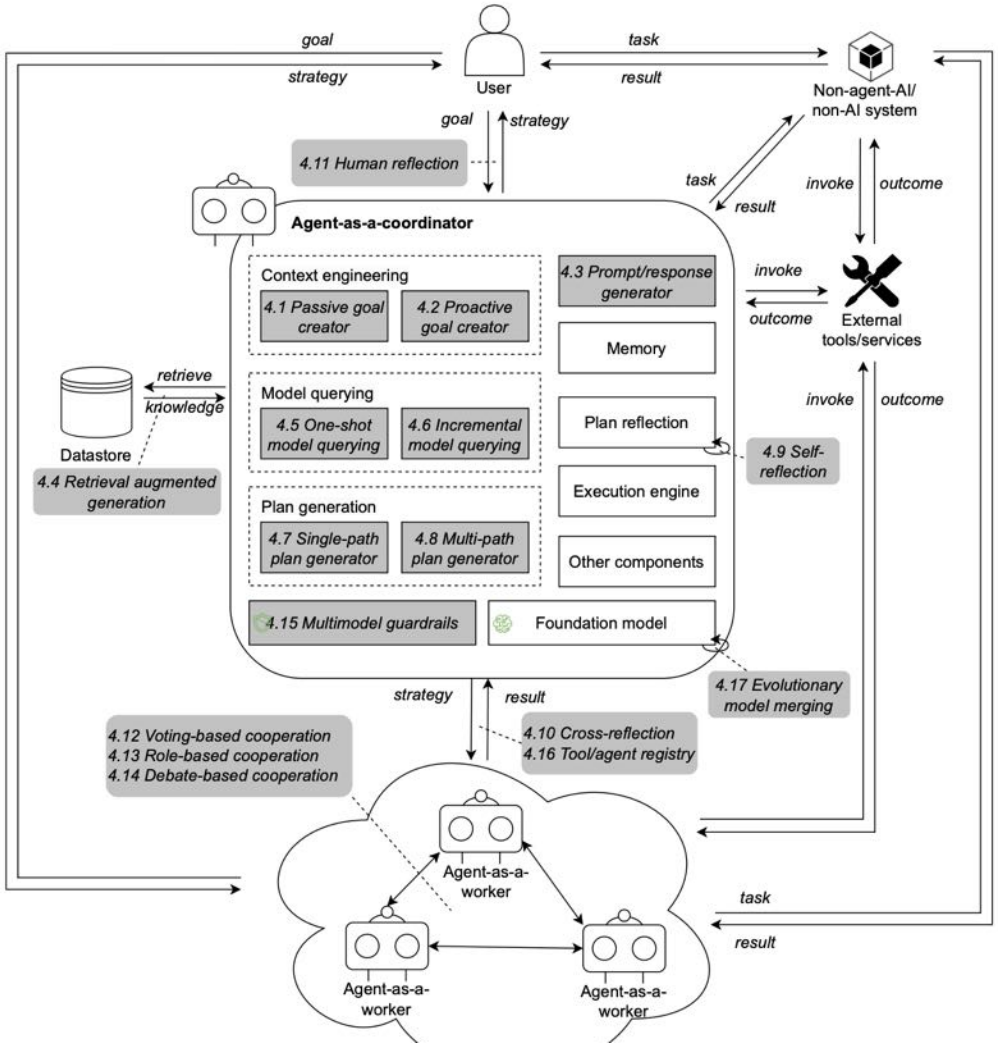


Guardrails and more

Think aloud
Access governance
”Flight recorder”
..

Lu, Q., Zhu, L., Xu, X., Xing, Z., Whittle, J., 2023. Towards Responsible AI in the Era of ChatGPT: A Reference Architecture for Designing Foundation Model-based AI Systems. http://arxiv.org/abs/2304.11090

Control of LLM-based Agents



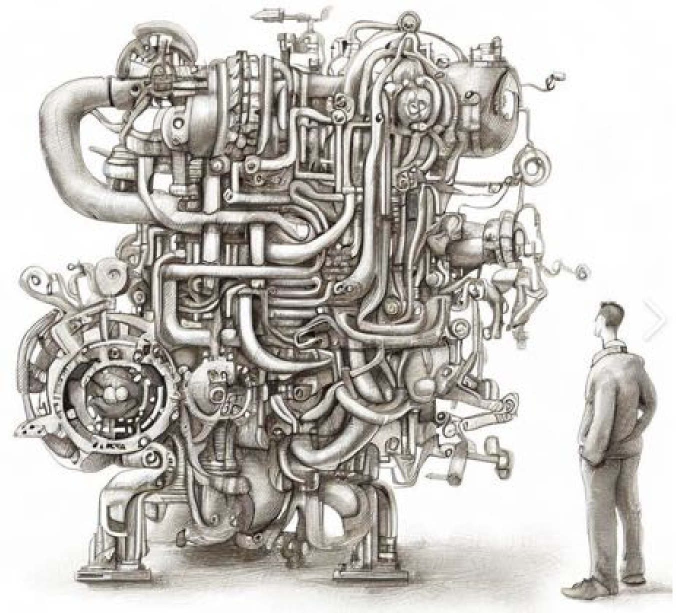
- Goals
- Context
- Memory
- Reasoning
- Plans
- Actions
- Tools
- Other agents
- Intermediate results
- Final results

Liu, Y. *et al.* (2024) ‘Agent Design Pattern Catalogue: A Collection of Architectural Patterns for Foundation Model based Agents’.: http://arxiv.org/abs/2405.10467

Control via System-Level Understanding

Do we have to fully understand AI models?

Can system-level understanding & guardrails help?



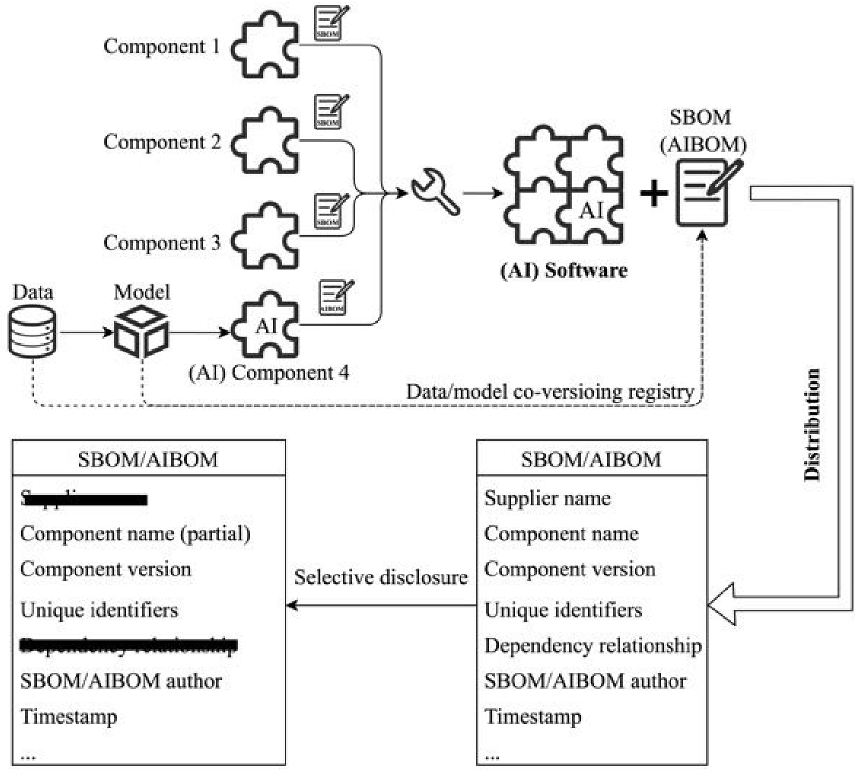
Increasingly, the study of these trained (but un-designed) systems seems destined to become a kind of natural science…

… they are similar to the grand goals of biology, which is to "figure out" while being content to get by without proofs or guarantees …

"AI as (an Ersatz) Natural Science?"
by Subbarao Kambhampati

Control via Supply Chain Accountability

Software Bills of Materials (SBOM)/AIBOM



SBOM/AIBOM

[redacted]

Component name (partial)

Component version

Unique identifiers

[redacted]

SBOM/AIBOM author

Timestamp

...

SBOM/AIBOM

Supplier name

Component name

Component version

Unique identifiers

Dependency relationship

SBOM/AIBOM author

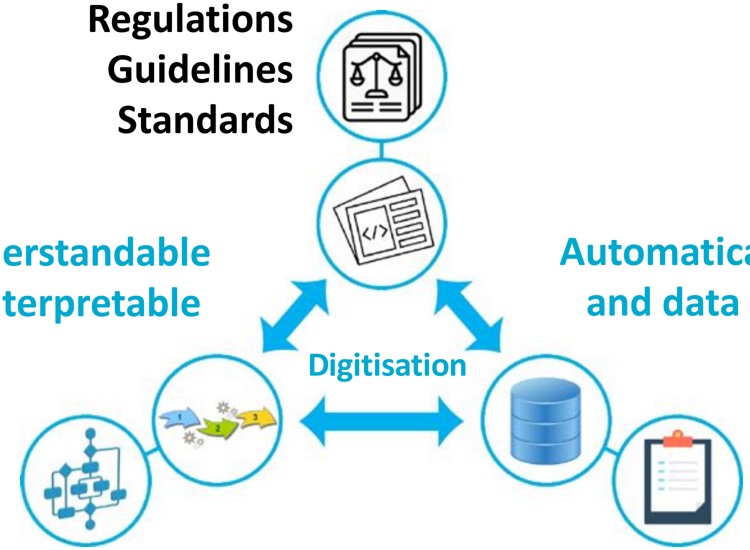
Timestamp

...

Data61 work: Xia, B., Bi, T., Xing, Z., Lu, Q., Zhu, L., 2023. An Empirical Study on SBOM: Where We Stand and the Road Ahead, in: 45th ICSE

Data61 work: Xu, X., Wang, C., Wang, Jeff, Lu, Q., Zhu, L., 2022. Dependency tracking for risk mitigation in machine learning systems, in: 44th ICSE

Can AI Help? Automated Design/Runtime Compliance



Regulations
Guidelines
Standards

Human understandable
Machine interpretable

Automatically verifying processes
and data against requirements

Digitisation

Process

Data

Data61 Technology

DAMOCLES^{TM}: Digital Process Compliance Suite

- Compliance by design
- Runtime monitoring
- Projected disruptions
- Preventative adaptation

Australian AI Safety Standard

1. Globally leading and accessible to small and medium enterprises (SME)
2. Globally leading in Diversity and Inclusion
3. Coherence with select international regulations, standards, principles & governance
4. Agile, modular and evolving
5. Practical & comprehensive – beyond just governance standards
   - Initial Focus: **Testing**, **Transparency** and **Accountability**

- **Deployer module** released; **Developer module** underway.

Australian Government
Department of Industry, Science and Resources

National Artificial Intelligence Centre

CSIRO

Voluntary AI Safety Standard

August 2024

Reclaiming the Reins:
Human Control in an End-to-End AI World

Debunk the myths: data/expertise value, human control
Focus on AI Systems, not just AI Models

- Design time control – encoding human oversight/agency responsibly
- Run time control - system-level guardrails, tools for humans
- Control of LLM-based AI systems and agents

- Close the regulation/standard-model gaps via **Responsible AI Engineering**

Australia's AI Safety Standard – v1 released
Mandatory Guardrail for High-Risk AI – consultation underway

#3 On Amazon