Trustworthy AI:
from Model to System to Agent

Prof. Liming Zhu
Research Director, CSIRO's Data61
Conjoint Professor, UNSW
Expert in Working Groups

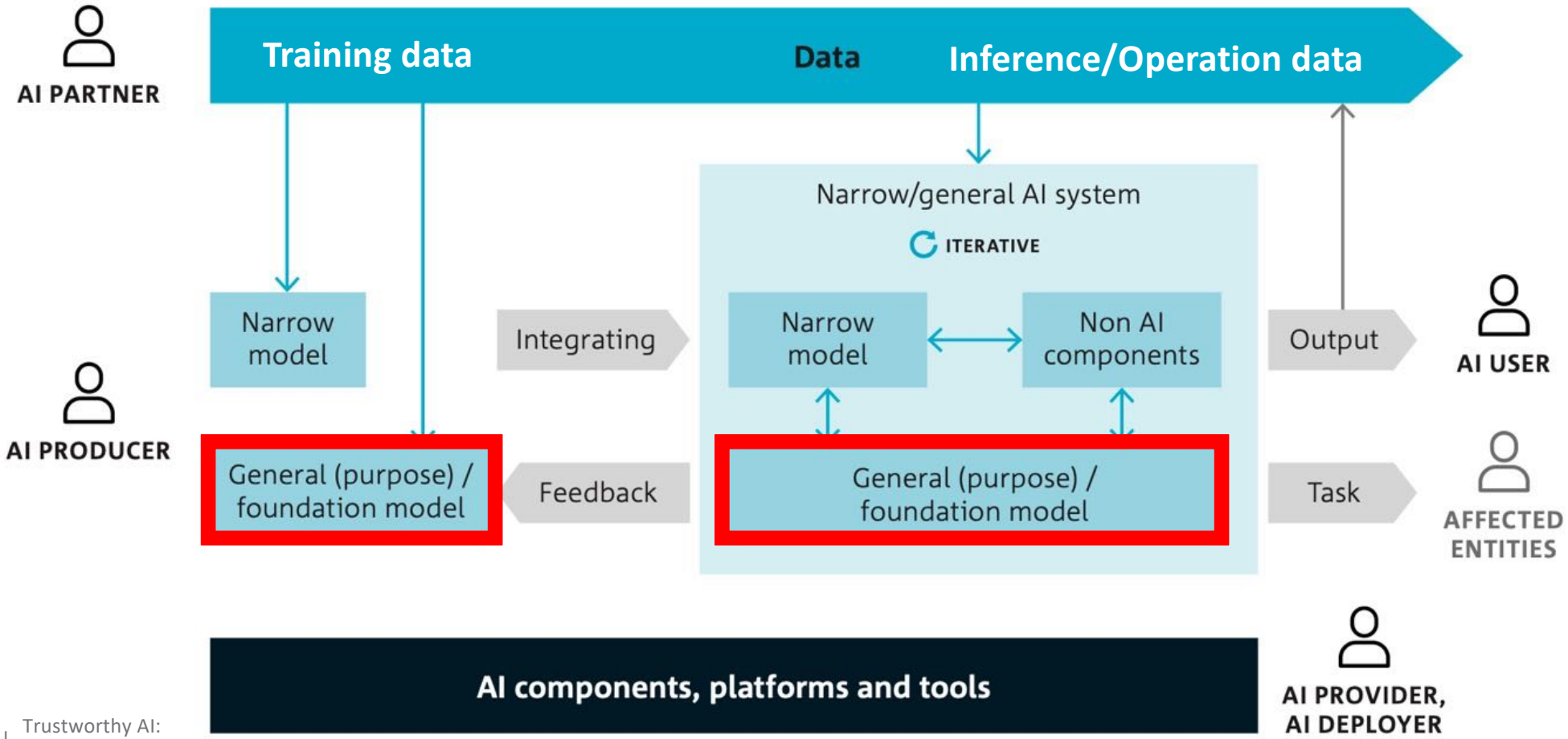
- Australia's AI Safety Standard
- OECD.AI AI Risk and Accountability
- ISO/IEC JTC 1/SC 42/WG 3 – AI Trustworthiness

Trends & Challenges

What's AI (System)?

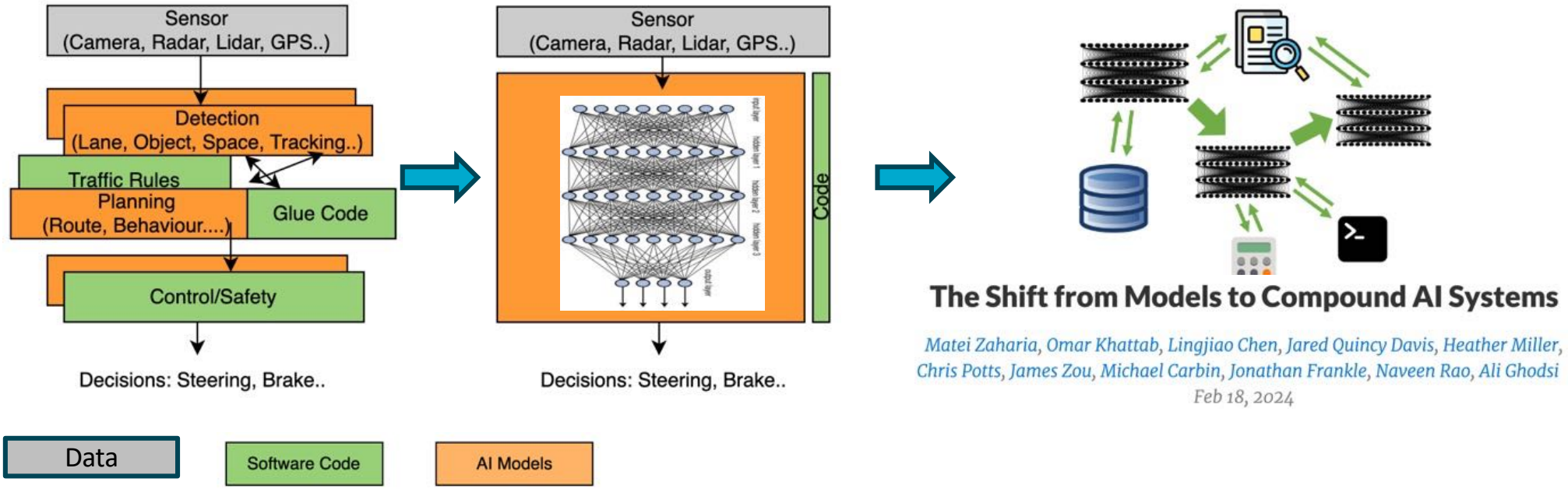
*a ~~designed~~ **machine-based system** that, for ~~a given set of human-defined~~ explicit or **implicit** objectives, infers, from the input it receives, how to generate outputs such as **predictions**, **content**, **recommendations**, or **decisions** that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and **adaptiveness after deployment**. (OECD)*

AI Model vs. AI System/Agent

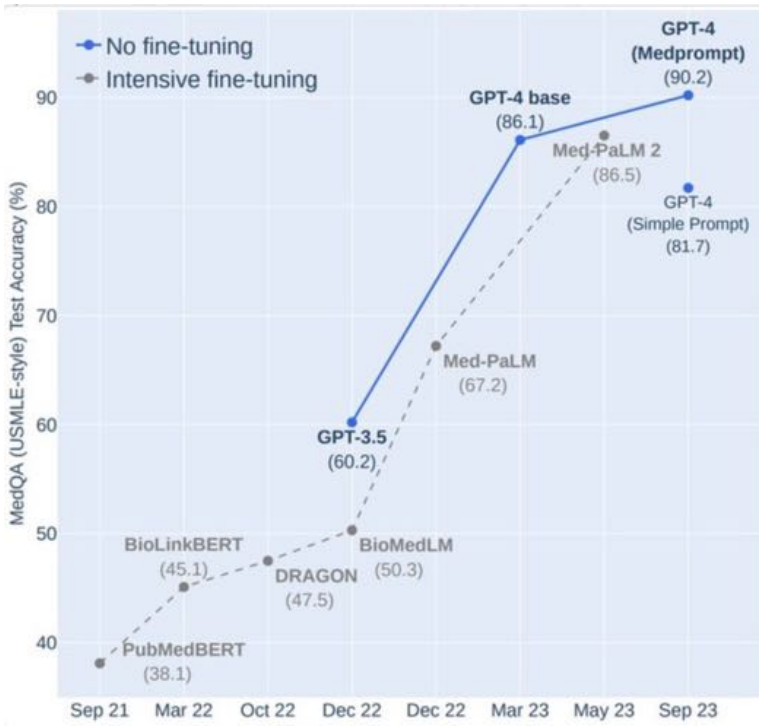


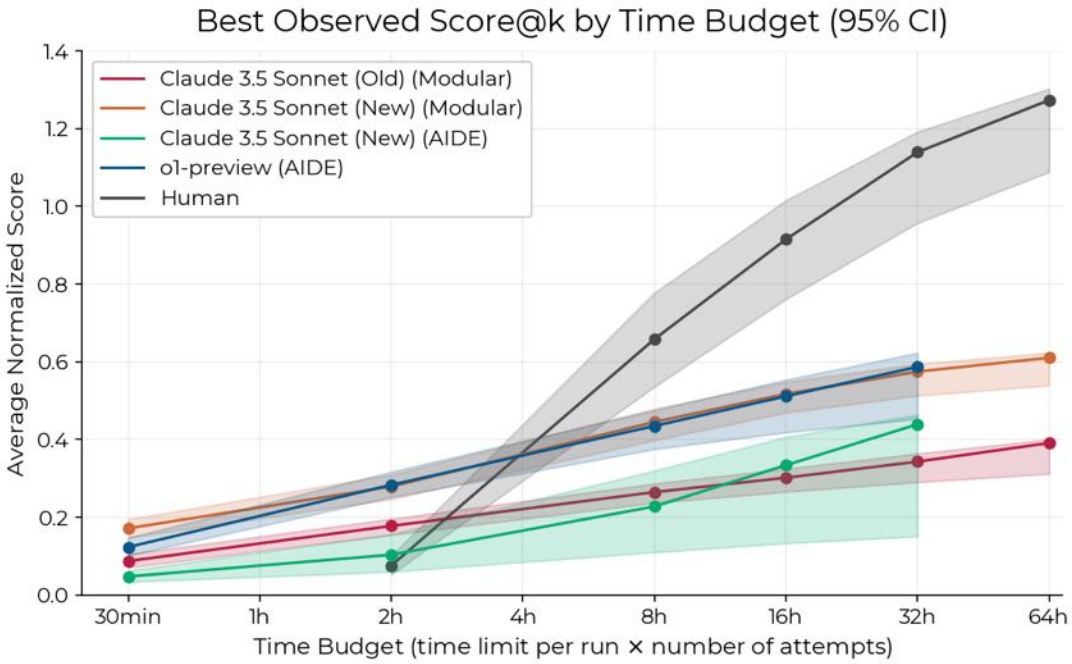
From Small to Large Model to Compound System

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



General vs Specific vs Human – Who Wins





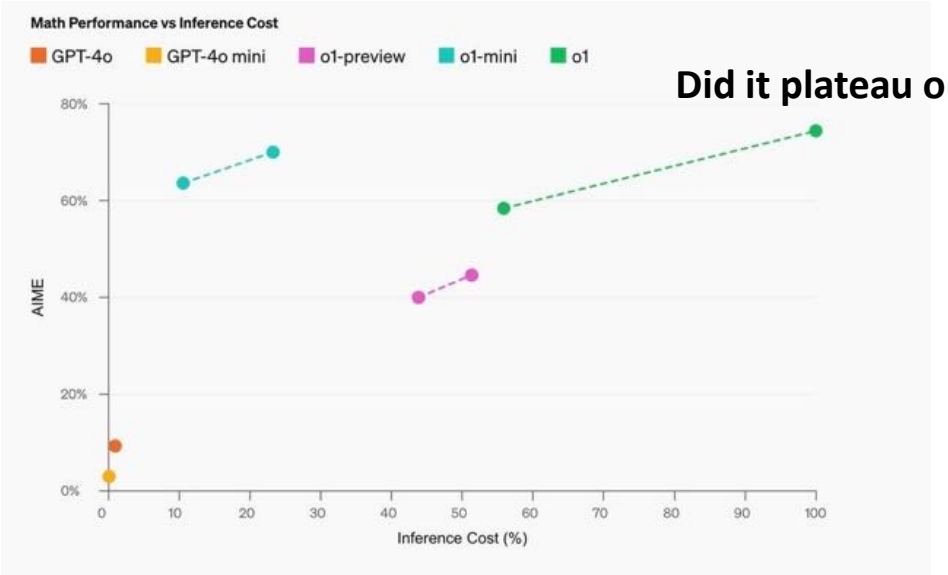
Value of unique data in training (vs inference)?

Is time budget for AI relevant?

Wijk, H. *et al.* (2024) 'RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts'. https://metr.org/blog/2024-11-22-evaluating-r-d-capabilities-of-llms/

Inference Time Scaling Law?

LLM + Python + for loop -> 15% capability increase…
and what about raw compute? Tools?



Did it plateau or ….?

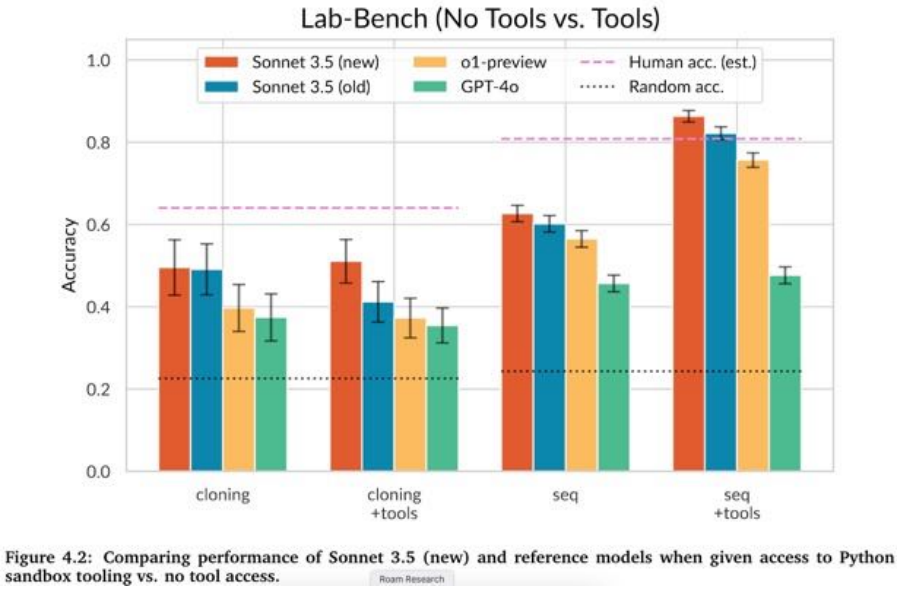
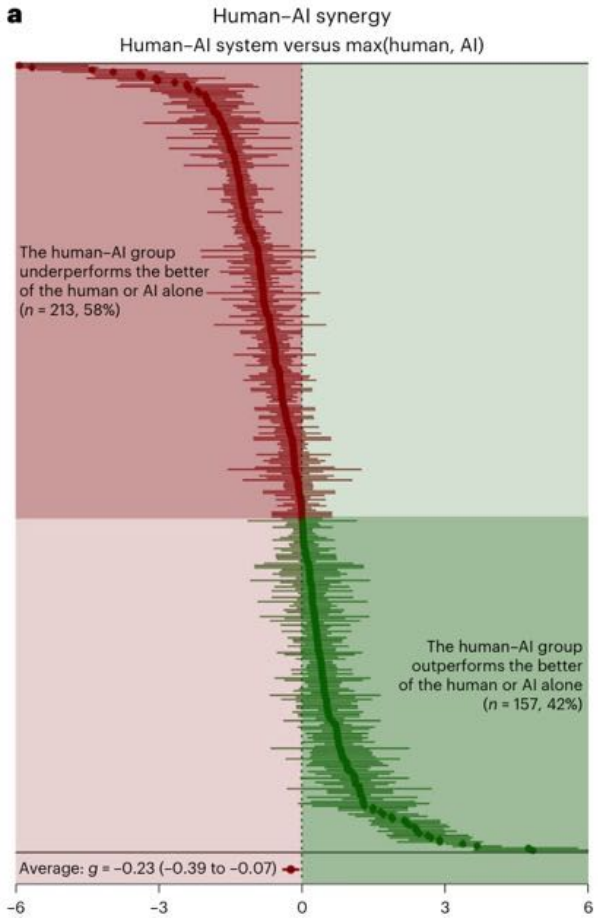


Figure 4.2: Comparing performance of Sonnet 3.5 (new) and reference models when given access to Python sandbox tooling vs. no tool access.

OpenAI (2024) *OpenAI o1 System Card*. OpenAI. https://openai.com/index/openai-o1-system-card/

US AISI and UK AISI Joint Pre-Deployment Test. UK/US AISI. https://www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet

Human+AI Less Trustworthy than AI/Human Alone?



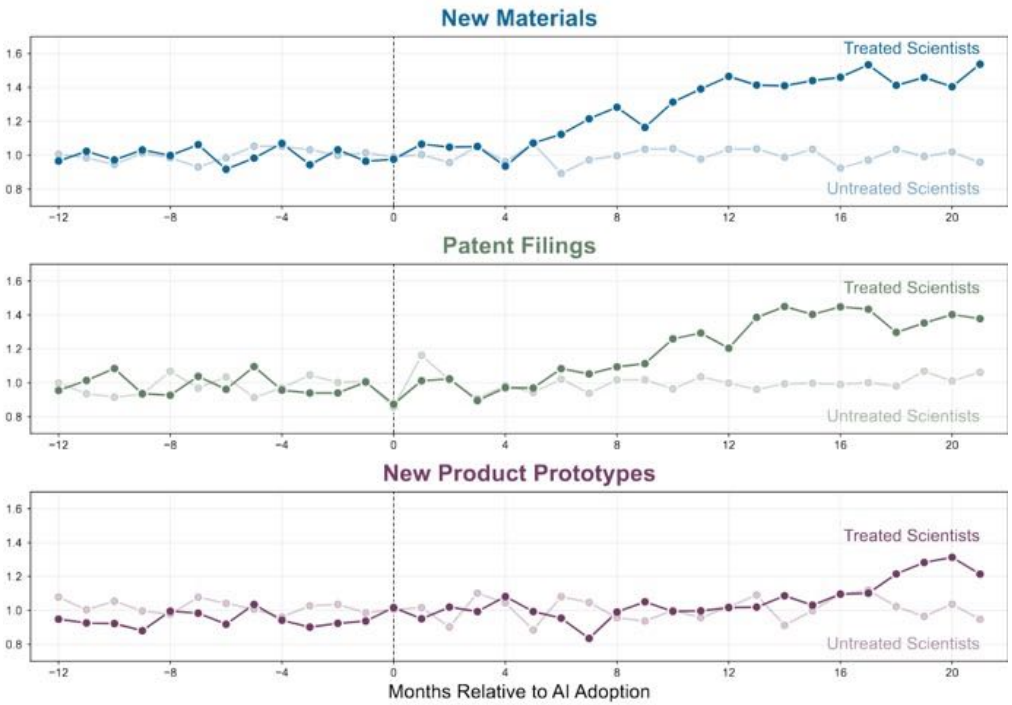
When the human outperformed the AI alone, performance gains occurred in the human–AI systems

When the AI alone outperformed the human alone, ***substantial performance losses*** *occurred in the human–AI systems.*

humans rely too little on AI (under-reliance), ignoring its suggestions ***because of adverse attitudes towards automation***

Vaccaro, M., Almaatouq, A. and Malone, T. (2024) 'When combinations of humans and AI are useful: A systematic review and meta-analysis', *Nature Human Behaviour*, pp. 1–11. https://doi.org/10.1038/s41562-024-02024-1

Different Effects on High/Low Performers?



Scientist: *While the bottom third of researchers see minimal benefit from the tool,* ***the output of top-decile scientists increases by 81%.***

Customer support agents: *14% increase in productivity,* ***with the most substantial gains observed among novice and low-skilled workers****, while experienced and highly skilled workers experienced minimal impact.*

Programmers: *50% increase in productivity, with statistically* ***significant productivity gains primarily among junior staff****, whereas the impact on more senior employees was less pronounced.*

Toner-Rodgers, A. (2024) 'Artificial Intelligence, Scientific Discovery, and Product Innovation'. https://conference.nber.org/conf_papers/f210475.pdf

The System Trust Gap

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

Data61 work: 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.

The Data Trust Gap – Trusted at Scale?

Synthetic Data
- **Trusted ground truth**: delay, quality, explainability, data drift, scalable eval.
- **Trusted provenance**: watermarking/metadata

System-Level Data
- **Trusted knowledge:** inconsistency
- **Trusted trade-off**: multi-dim tradeoff (privacy, fairness, accuracy…)

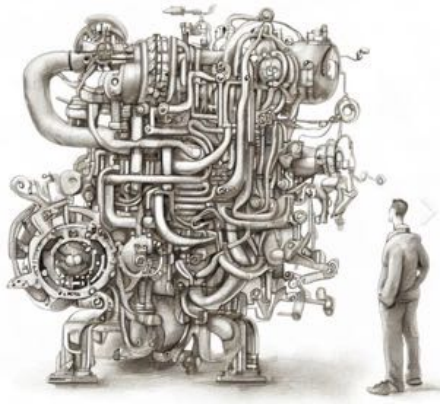
Testing/Validation Data
- **Trust in eval data:** quality, data leak..
- **Trust in human feedback**

Training Data – bias/repr.
- **Distributed trust:** cross-org/on-device, trust without seeing
- **Zero trust:** data poisoning/fingerprinting
- **Trusted "license" to use:** data rights (vs. ownership/copyright)
- **Trusted artificial:** synthetic data

Trustworthy Whole out of Untrustworthy Parts

Do we have to fully understand and trust AI models part?

Can system-level understanding, guardrails and design assure trustworthiness?

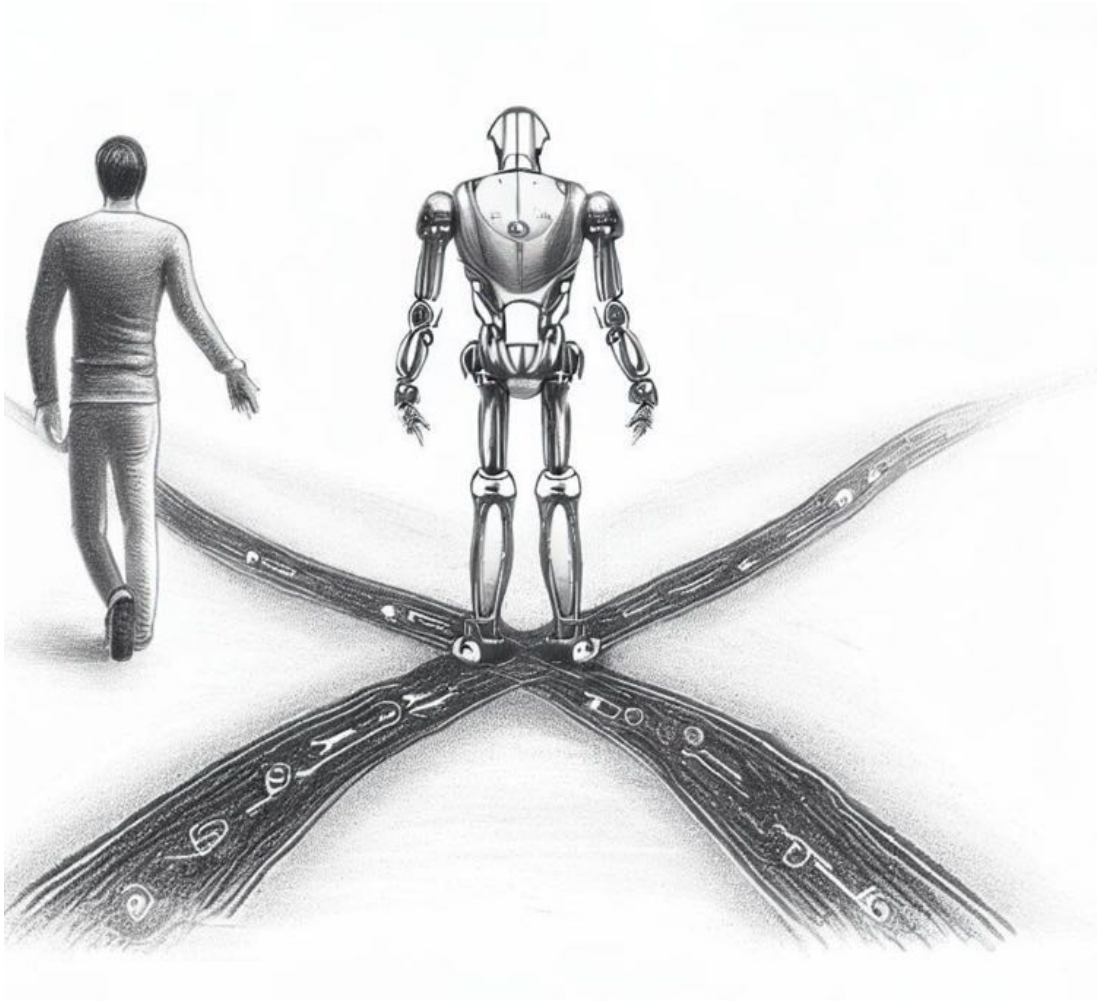


Increasingly, the study of these trained (but un-designed) systems seems destined to become a kind of natural science…

… 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

Science Approaches



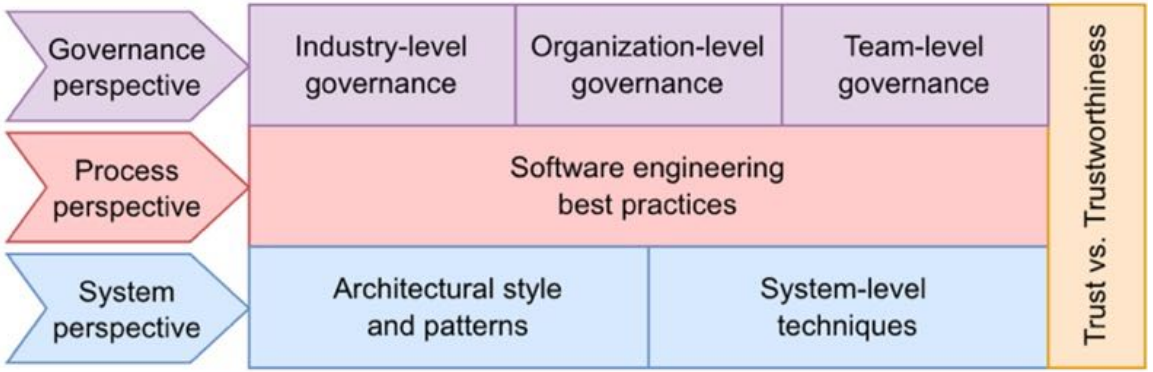
Design-time Trustworthiness: AI Engineering

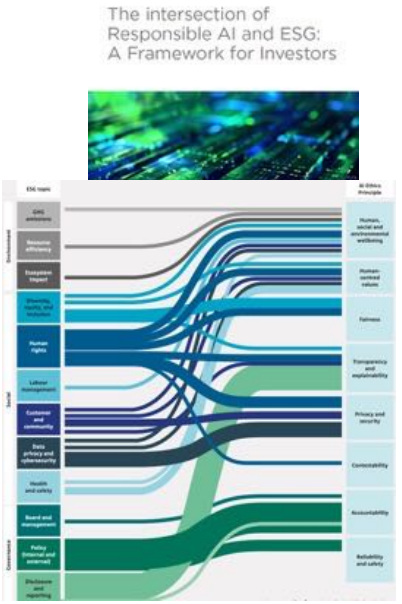
Standards
Frameworks

Australia's AI ethics framework | OECD AI principles | EU AI Act | ...

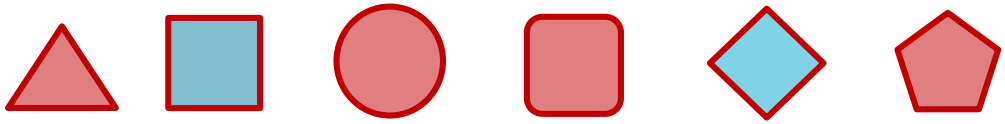
AU Safety Standard | ISO Standards | NIST AI RMF

Responsible AI (RAI) Engineering





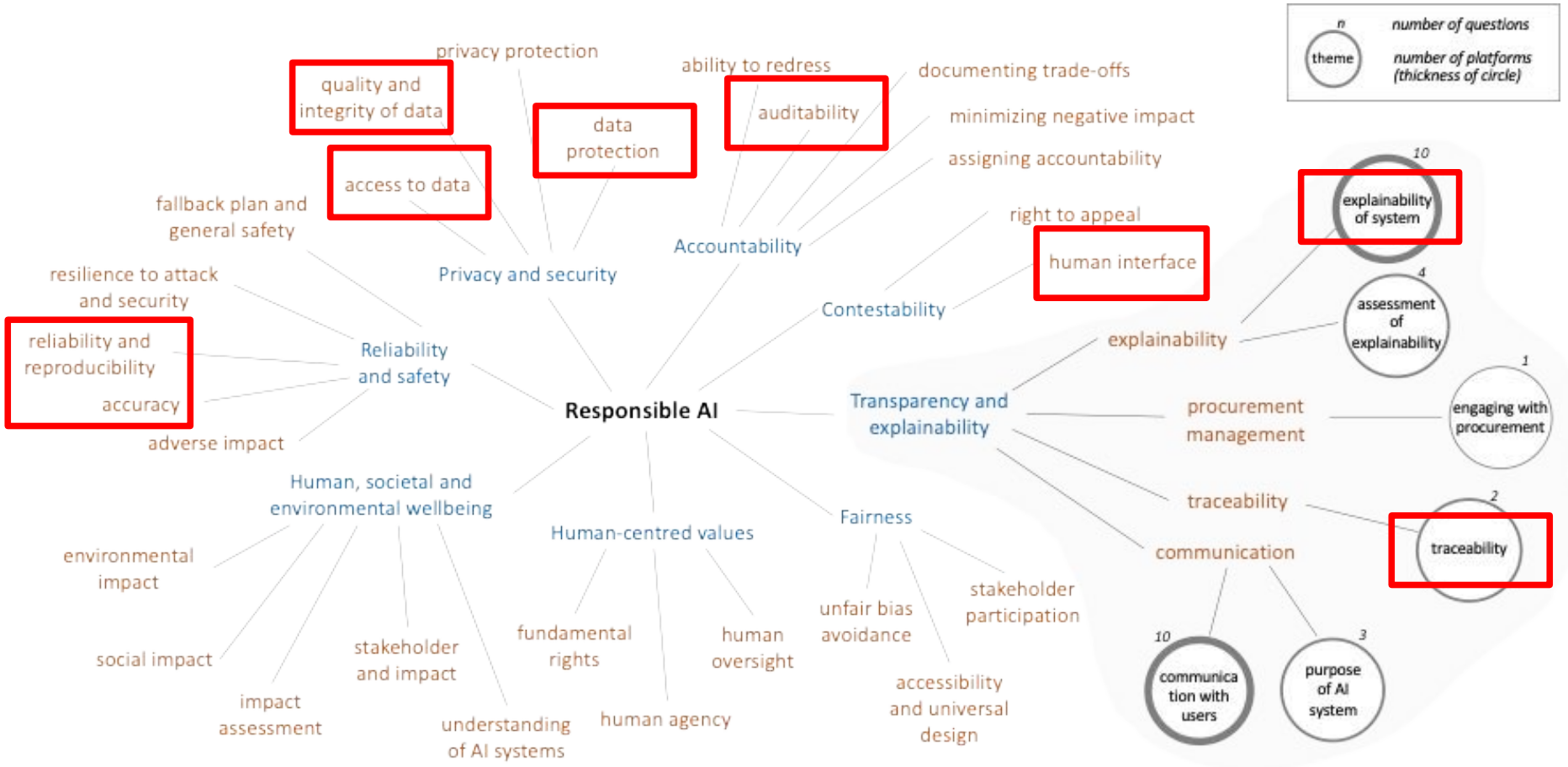
Models



Trustworthy AI:
from Model to System to Agent

Data61 work: 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)

Question Bank for Stakeholders



Data61 work: Lee, S.U., Perera, H., Xia, B., Liu, Y., Lu, Q., Zhu, L., Salvado, O. and Whittle, J., 2023. QB4AIRA: A Question Bank for AI Risk Assessment. https://arxiv.org/pdf/2305.09300.pdf

AI Accountability Metrics Catalogue

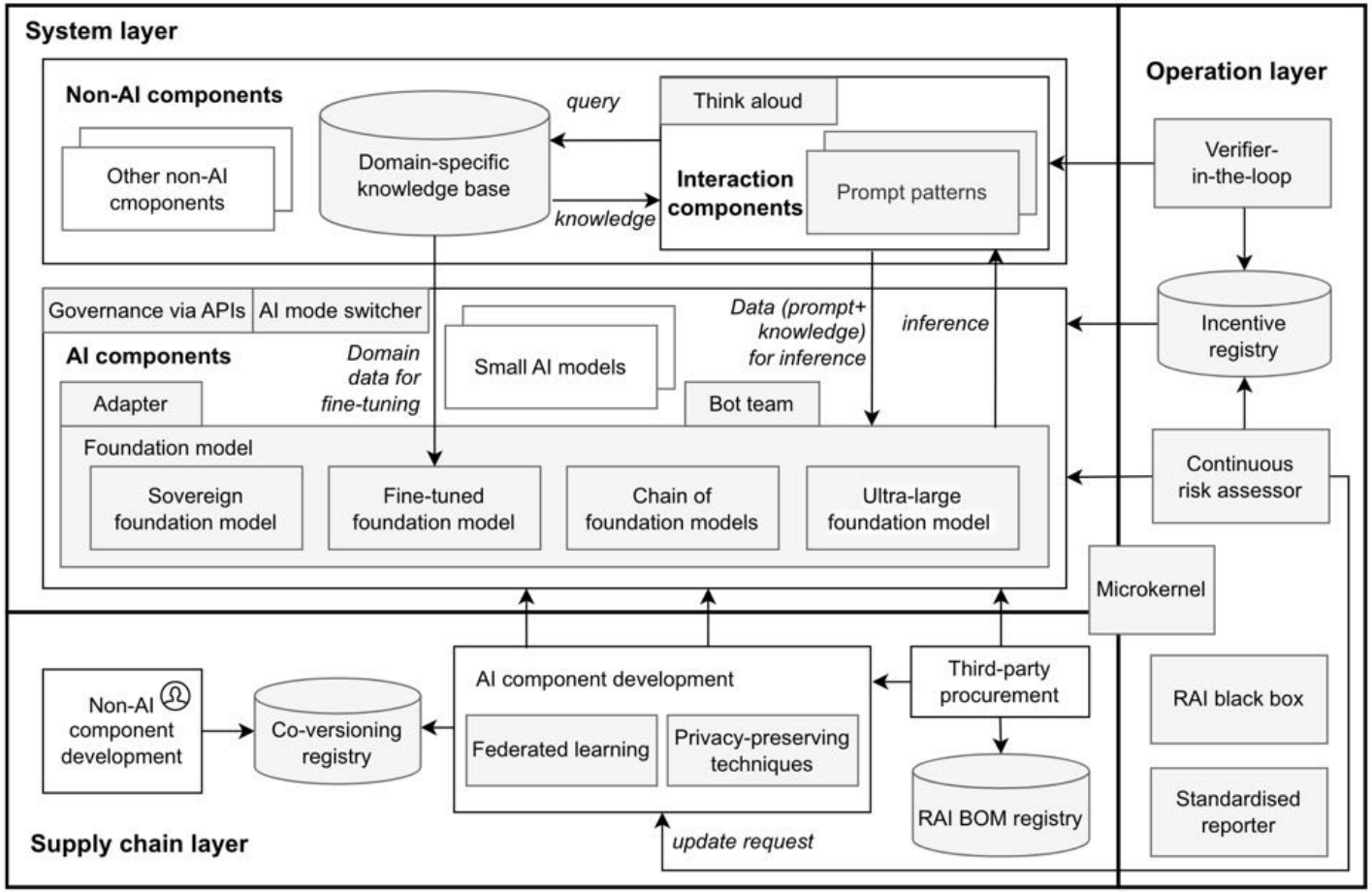
Table 2: System-Level Metrics Catalogue for AI Accountability

Criteria	Sub-Criteria	Process Metrics	Key Considerations	Resource Metrics	Product Metrics
Responsibility	RAI Oversight	Roles and Responsibilities	• Comprehensive role clarity: - Design and development - Deployment and operations - Procurement and integration - Governance and compliance - AI as a service	• Soft laws (e.g., best practices, guidelines standards etc) • Hard laws (e.g., EU AI Act)	• Procedure Manuals • Contracts or agreements • Position descriptions • Recruitment practices • Workforce dev strategy
		RAI Governance Committee	• Multidisciplinary composition • Strategic leadership involvement		• Policy doc on Committee
		Organizational AI Risk Tolerance	• Tiered risk-based categorization • Balancing competing interests		• Policy doc on org's risk tolerance and mitigations
	RAI Competence	RAI Training	• Holistic training content • Targeted training for diverse roles • Adaptive and ongoing education		• Training certificates
		RAI Capability Assessment	• Multifaceted assessment • Standard alignment • Organizational RAI maturity • Continuous enhancement		• Assessment reports
Auditability	Systematic Oversight	Data Provenance	• Detailed data record-keeping • Data version control • Data integrity and risk mitigation • Legal and ethical compliance	• Soft laws (e.g., auditing guidelines and frameworks etc) • Hard laws (e.g., EU AI Act) • AI documentation tools (e.g., datasheets, model/system cards) • Technical tools (e.g., blockchain, knowledge graph)	• Provenance records • System features (e.g., auto-logging, version control
		Model Provenance	• Detailed model record-keeping • Model selection and validation • Model version control		
		System Provenance and Logging	• Detailed system record-keeping - System version control - Decision/Trade-off • Comprehensive operational logging - User interaction and system response - Incident and response - System configuration changes • Composition Management		• Provenance records (and logs) • System features (e.g., auto-logging, version control
	Compliance Checking	Auditing	• Diversified auditing strategy • Multi-dimensional audit techniques • Ethical and legal compliance • Regular audits • Verifiable audits • Audit-driven improvements		• Audit reports • Compliance certificates and licenses
Redressability	Redress-by-Design	Incident Reporting and Response	• Accessibility and Visibility • Structured Incident Management • Feedback Loop Integration	• Redundancy design case studies • Incident management tools	• Incident and response doc • System features (user feedback and report)
		Built-in Redundancy	• Multi-Modal Redundancy		• System features (redundant components/functionalities)

Data61 work: Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, Zhenchang Xing, Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability, submitted to CAIN'24.

Trustworthy AI Design Patterns

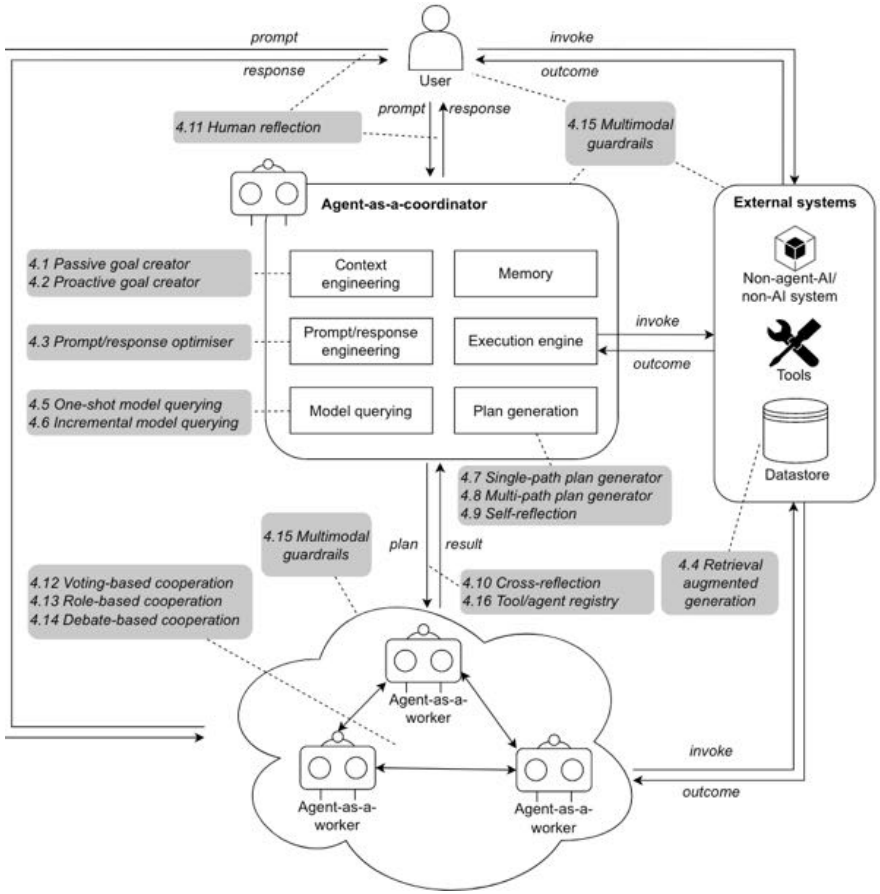
Trustworthy Systems out of Untrustworthy Components Parts



Data61 work: 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

Trustworthy Agent Design Patterns

Trustworthy Outcome out of Untrustworthy Sub-Goals



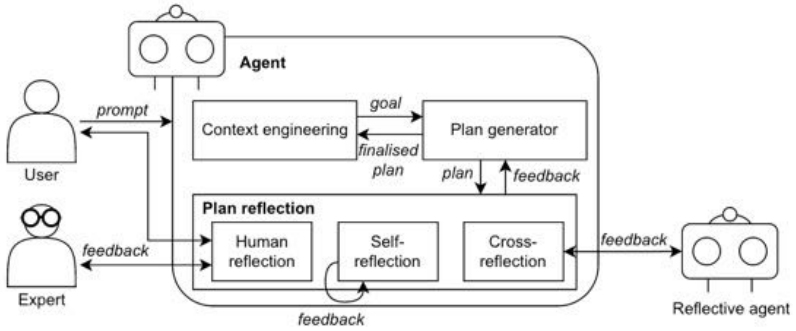


Figure 11: Plan reflection pattern.

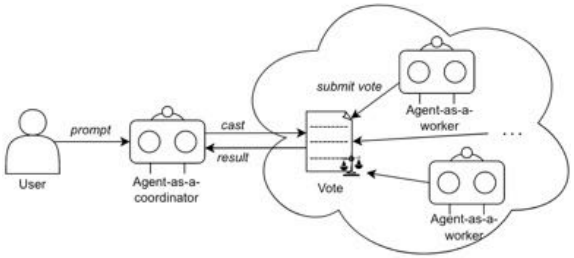


Figure 12: Voting-based cooperation.

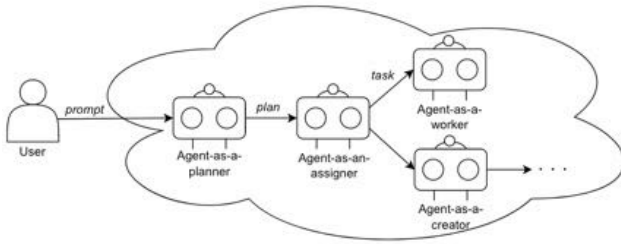


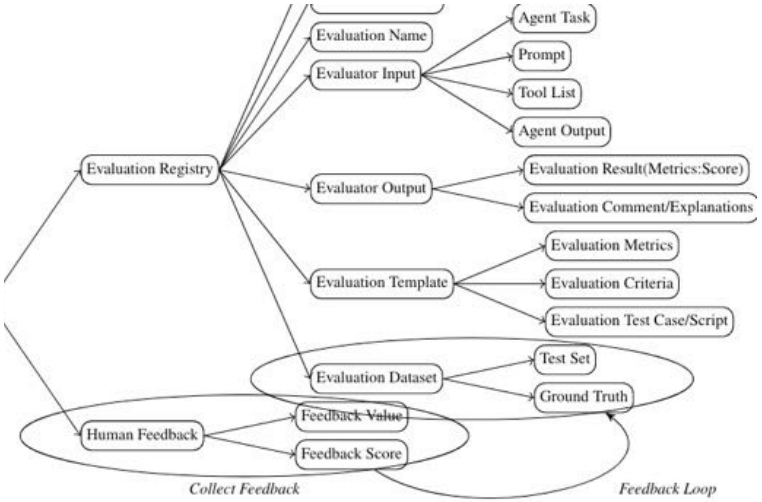
Figure 13: Role-based cooperation.

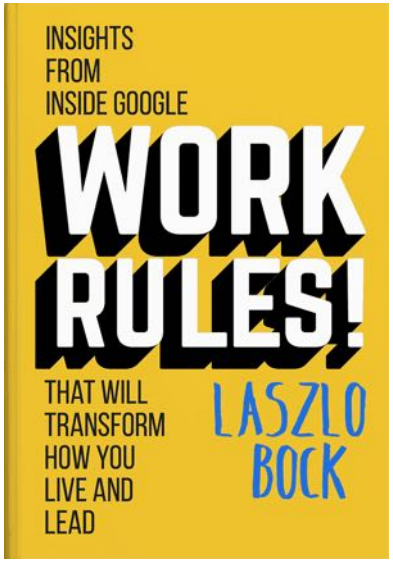
Data61 work: Liu, Y. , Zhu, L. *et al.* (2024) 'Agent Design Pattern Catalogue: A Collection of Architectural Patterns for Foundation Model based Agents'. arXiv. http://arxiv.org/abs/2405.10467

AgentOps – DevOps for Agent-Based Systems

Trustworthy Processes out of Untrustworthy Tasks

Key Aspects	Key Features	Description
Agent Creation	Provision, Custom, Spawn & Deploy Autonomous AI Agents	Create production-ready & scalable autonomous agents.
	Extend Agent Capabilities with Toolkits	Add Toolkits from marketplace to agent workflows.
	Extend Agent Capabilities with Multiple Vector DBs	Connect to multiple Vector DBs to enhance agent's performance.
	Extend Agent Capabilities with (fine-tuned) Models	Custom fine-tuned models for business specific use cases.
Prompt Management	Prompt Versioning and Management	Keep track of different versions of prompts used in agents. Useful for A/B testing and optimizing agent performance.
	Prompt Playground with Model Comparisons	Test and compare different prompts and models for agents before deployment.
	Prompt Injection Detection	Identify potential code injection and secret leaks.
Evaluation and Test	Test Agents Against Benchmarks and Leaderboards.	Create a dataset; Define metrics; Run Evaluations; Comparing results; Track results over time etc.
	Evaluate Agent in Diverse Steps	Evaluate final response- Evaluate the agent's final response.
		Evaluate single step-Evaluate any agent step in isolation (e.g., whether it selects the appropriate tool).
		Evaluate trajectory- Evaluate whether the agent took the expected path (e.g., of tool calls) to arrive at the final answer.
Human Feedback	Collect Explicit Feedback	Directly prompt the user to give feedback, this can be a thumb up or a thumb down.
	Collect Implicit Feedback	Measure the user's behavior, this can be time spent on a page, click-through rate.
Monitoring	Agent Analytics Dashboard	Monitor diverse level and dimension statistics metrics about agents.
	LLM Cost Management and Tracking	Track spend (token cost) with foundation model providers.
Tracing	Trace Agent Execution Process	Trace each agent run, e.g., the whole chain, retrieval, LLM call, Tool Call etc.
		Trace evaluation run
		Trace user feedback





"People Operation"

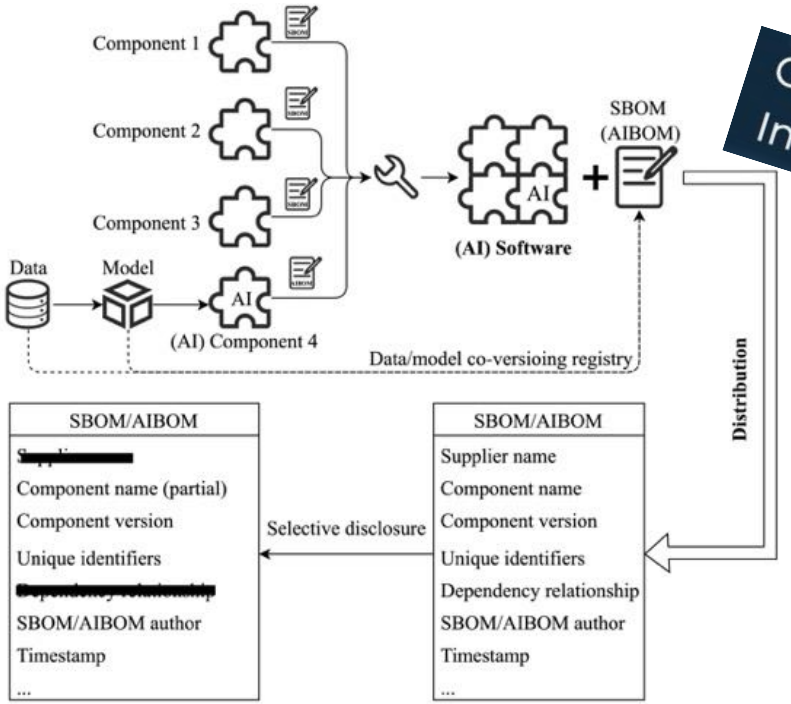
Agent == People?

Data61 work: Dong, L., Lu, Q. and Zhu, L. (2024) 'A Taxonomy of AgentOps for Enabling Observability of Foundation Model based Agents'. arXiv. https://doi.org/10.48550/arXiv.2411.05285

AI/Agent Bills of Materials (AIBOM)

Trustworthy Supply Chain out of Untrustworthy Suppliers

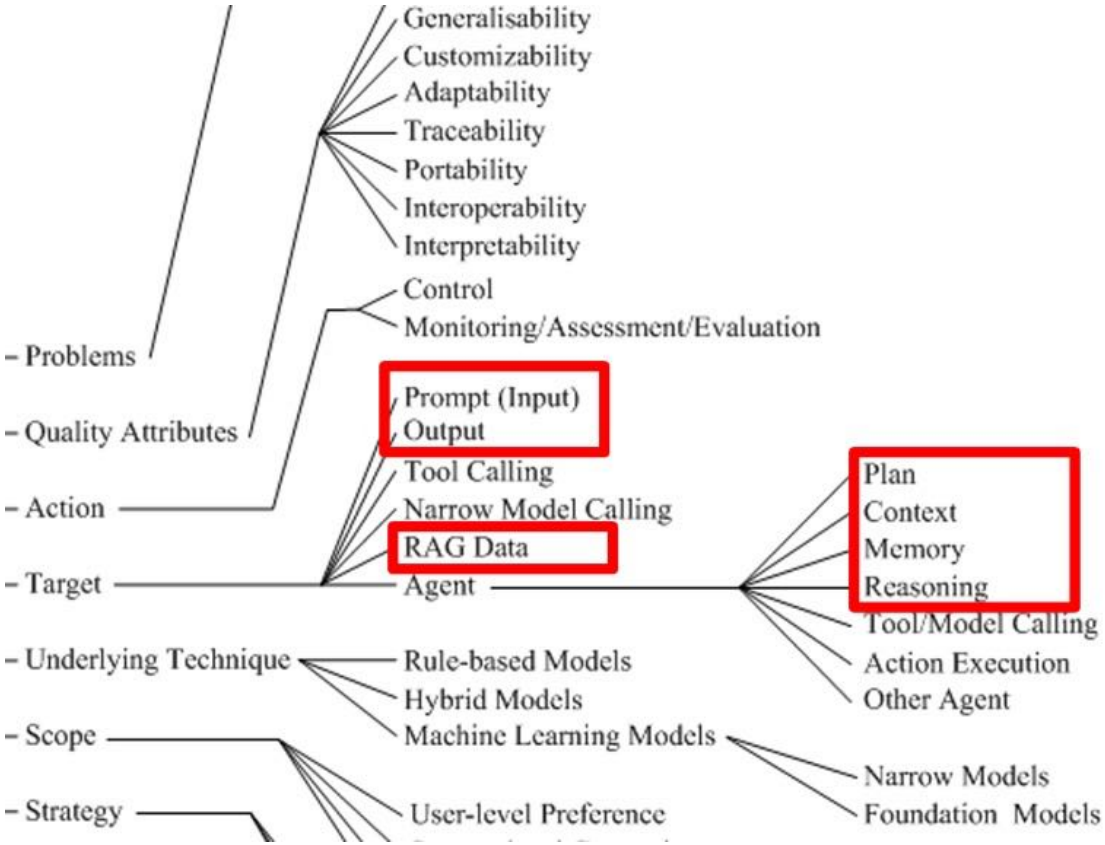
Software Bills of Materials (SBOM)/AIBOM



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

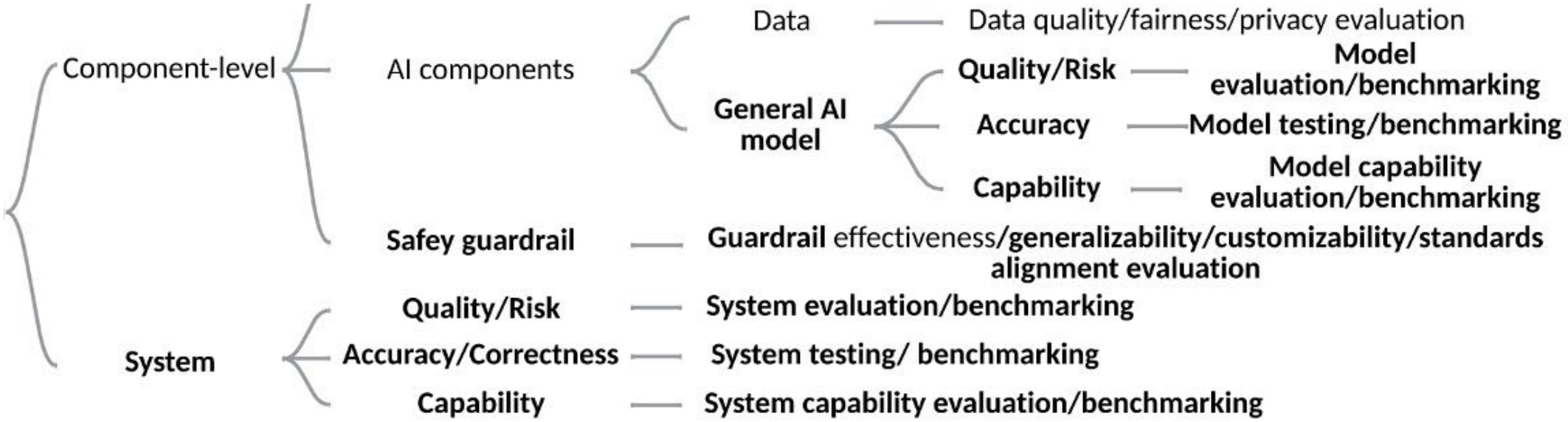
Data61 work: Xu, X., Wang, C., Wang, Jeff, Lu, Q., Zhu, L., 2022. Dependency tracking for risk mitigation in machine learning systems, in: 44th ICSE

Runtime Trustworthiness: Guardrails



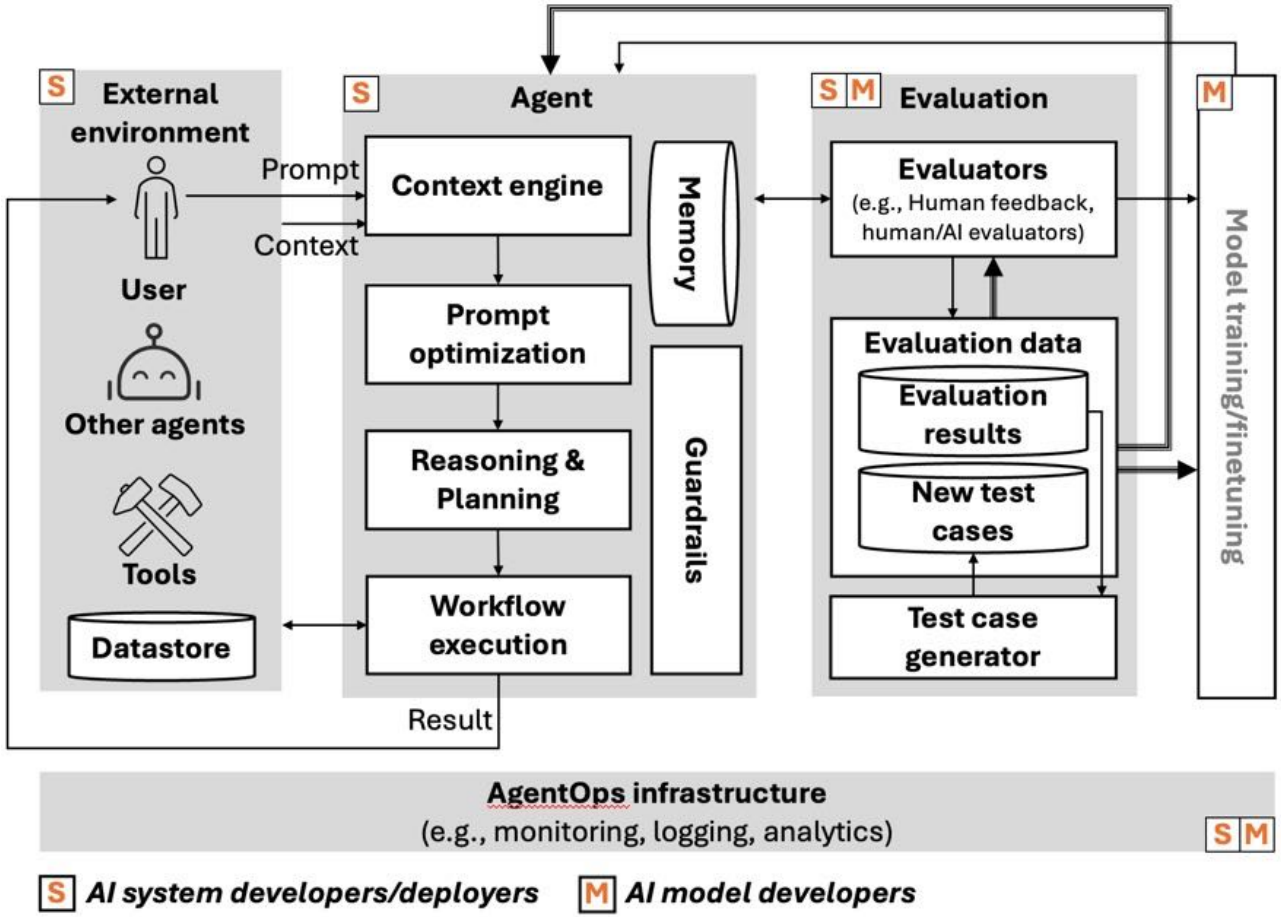
Data61 work: Shamsujjoha, M. *et al.* (2024) 'Towards AI-Safety-by-Design: A Taxonomy of Runtime Guardrails in Foundation Model based Systems'. arXiv. Available at: https://doi.org/10.48550/arXiv.2408.02205.

Evaluation at the System Level (beyond Model)



Data61 work: Xia, B., Lu, Q., Zhu, L., Xing, Z., 2024. Towards AI Safety: A Taxonomy for AI System Evaluation. https://doi.org/10.48550/arXiv.2404.05388 AIWare 2024

Evaluation-Driven (Out-of-Model) Learning

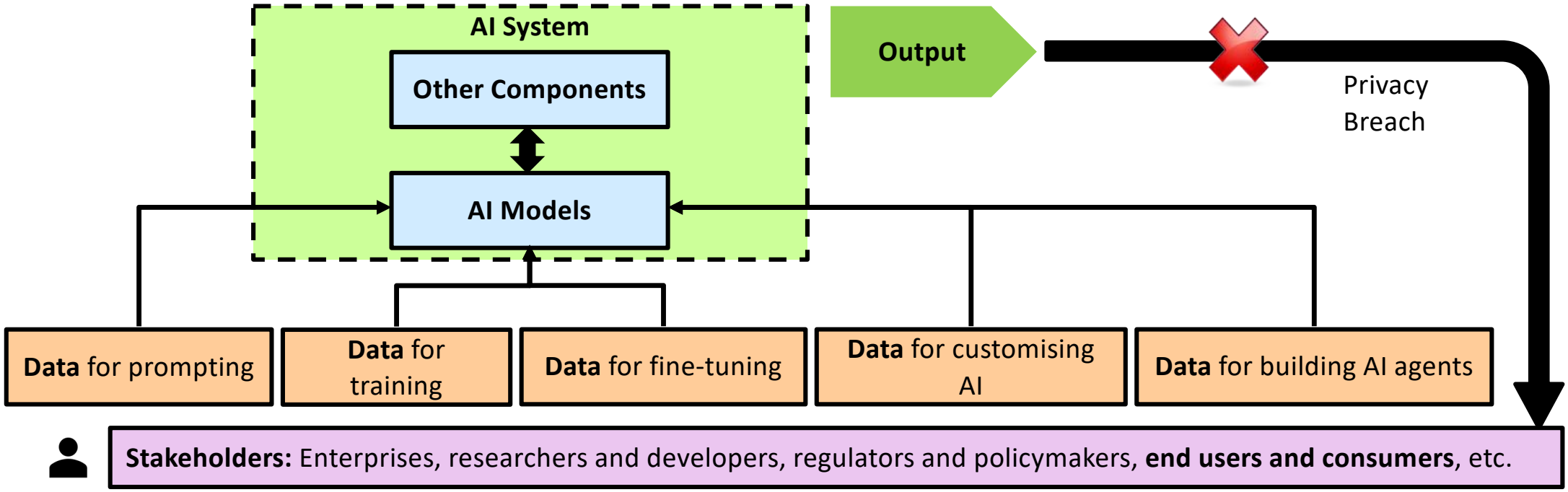


Test Driven Development

Evaluation-Driven Learning

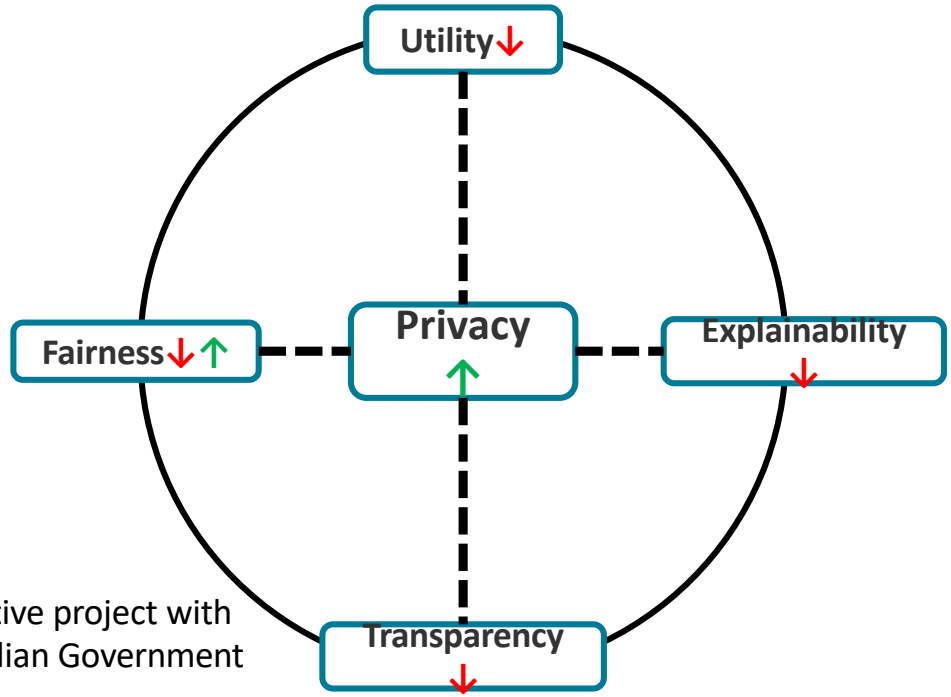
Trade-offs in Trustworthiness - Privacy

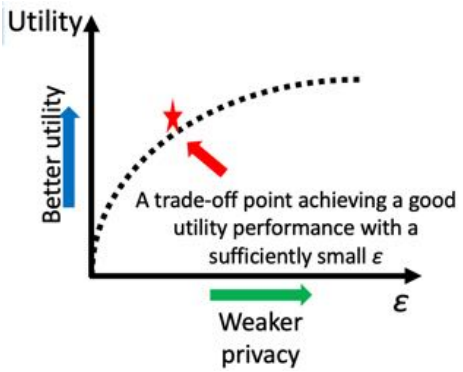
- AI output shall not reveal any private information of users (e.g., address)



Trade-offs Between Privacy and Other Principles

- Privacy-preservation methods may cause AI **utility** degradation, **fairness** loss, decreased **transparency**, and reduced **explainability**, etc.





Data61 work: DP-Copula: A collaborative project with **Department of Social Services**, Australian Government
The dataset (~5M rows, 27 attributes)

Transparency vs Commercial Confidentiality

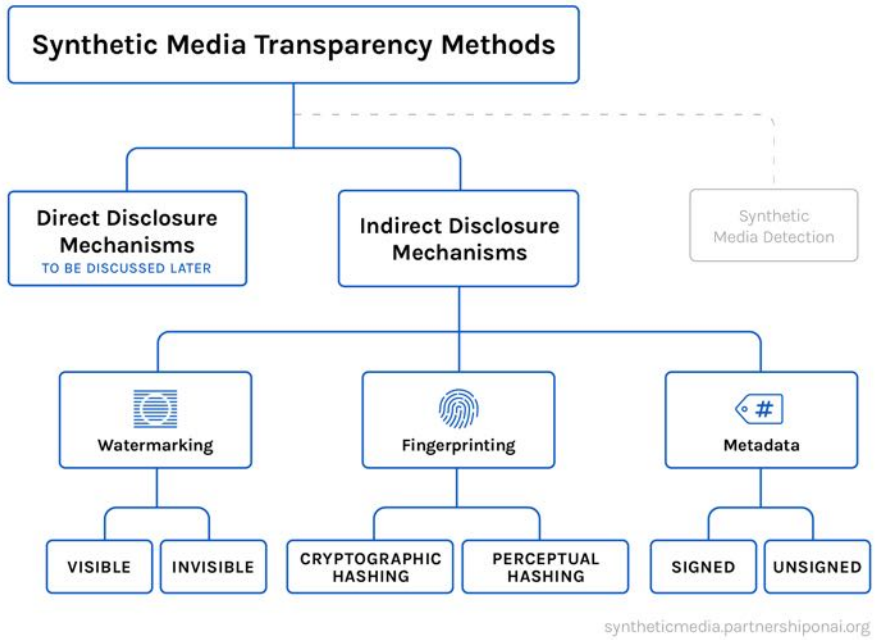
Benefits and limitations of **black/grey-box/out-of-box** evaluation

	Access Level	Black-Box	Grey-Box	De facto White-box	White-Box	Outside-the-box
Test sets (Section 3)	Queries	✔	✔	✔	✔	✗
Manual attacks (Section 3)		✔	✔	✔	✔	✗
Transfer-based attacks (Section 4.1)		✔	✔	✔	✔	✗
Gradient-free attacks (Section 4.1)		✔	✔	✔	✔	✗
Sampling-probability-guided attacks (Section 4.1)	Probabilities	✗	✔	✔	✔	✗
Gradient-based attacks (Section 4.1)	Gradients	✗	✗	✔	✔	✗
Hybrid attacks (Section 4.1)		✗	✗	✔	✔	✗
Latent space attacks (Section 4.1)	Weights/ Activations	✗	✗	✔	✔	✗
Mechanistic interpretability (Section 4.2)		✗	✗	✔	✔	✗
Fine-tuning (Section 4.3)	Fine-tuning	✗	✗	✔	✔	✗
Methodological evaluations (Section 5)	Outside-the-Box	✗	✗	✗	✗	✔
Data evaluations (Section 5)		✗	✗	✗	✗	✔
Complementary evaluations (Section 5)		✗	✗	✗	✗	✔
Using source code (Section 5)		✗	✗	✗	✗	✔
Copying system parameters (Section 6)	Unrestricted	✗	✗	✗	✔	✗

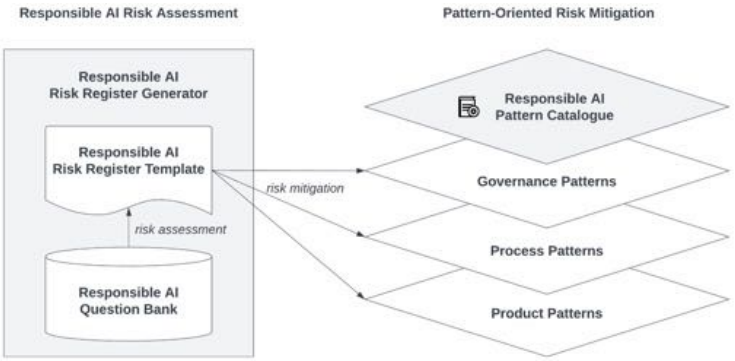
Casper, S.,D., Tegmark, et al., 2024. Black-Box Access is Insufficient for Rigorous AI Audits.

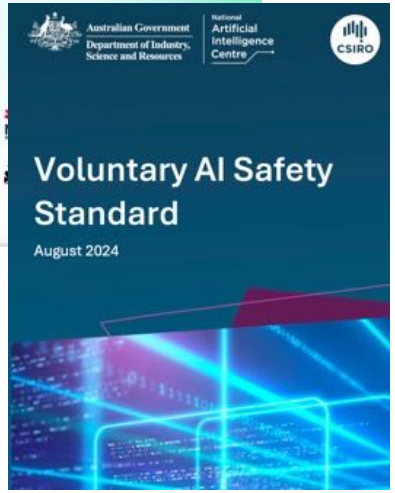
Trustworthy Synthetic Content

- Which Role
  - during/post generation and distribution?
    - vs. **censorship/moderation**
- On What
  - Multimedia, Text, Code…
    - vs **robustness (FP, FN)**
- Easily removable or not
  - vs. **privacy concerns**



Put it Together: Data61's Best Practice Guides





deployer v1
developer v2 coming

TBD

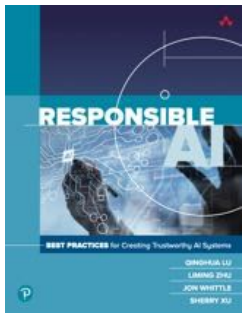
DISR

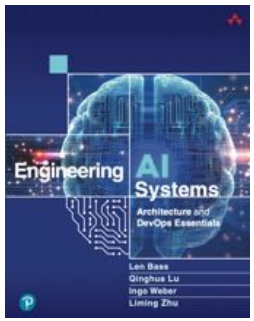
AISI AISI AISI AISI AISI AISI AI Office

Inaugural Convening of International Network of AI Safety Institutes

Best practice catalogue

300k+ impressions in 9 months





Australia's 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
   - Part of the international AI Safety Research Network
4. Agile, modular and evolving
5. Practical & Technical – beyond just governance/management
   - Initial Focus: **Testing**, **Transparency** and **Accountability**

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

Trustworthy AI: Model->System->Agent

System-level challenges

- Humans are no longer the ground truth
- We may never understand the model
- Inference-time scaling + tool -> capability jump

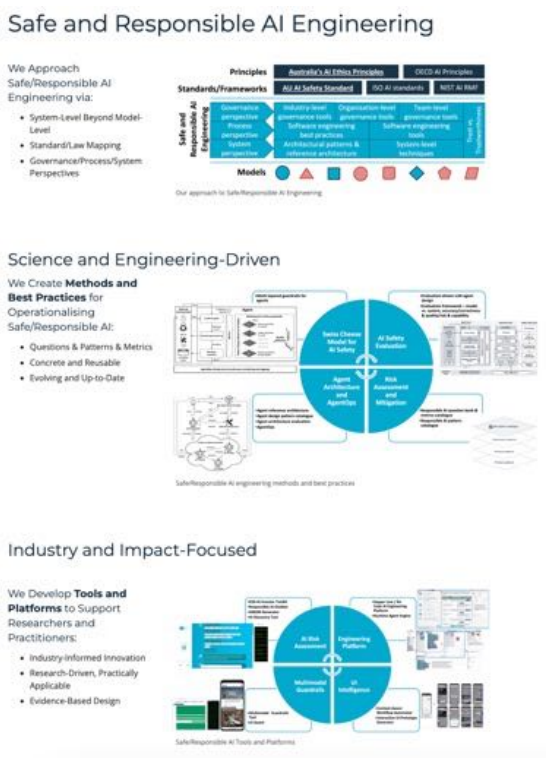
System/Agent-level AI Engineering

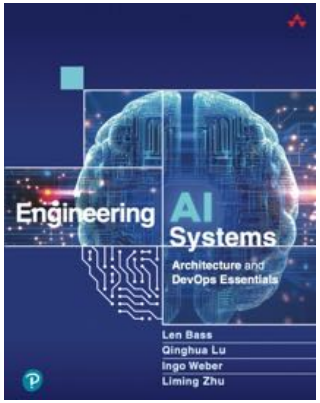
- Patterns and Guardrails
- Out-of-model learning
- Tradeoffs

Australia's AI Safety Standard

- v1 released, v2 underway

International Network of AI Safety Institute





https://research.csiro.au/ss/team/se4ai/responsible-ai-engineering/