Innovation in Automated and AI-Driven Decision Making

Real-World Case: AI Doc Evaluation

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

Why Defining “Decision-Making” Is So Contentious

Execution (apply rules) vs. **Deliberation** (reason) vs. **Discretion** (choose under undetermined rules)

- AI can execute, deliberate, even apply discretion and act
- Humans may have discretion but no real choices

Design-Time vs. Operation-Time

– Humans encode rules; AI learns rules & infers beyond them

Agency Mismatch

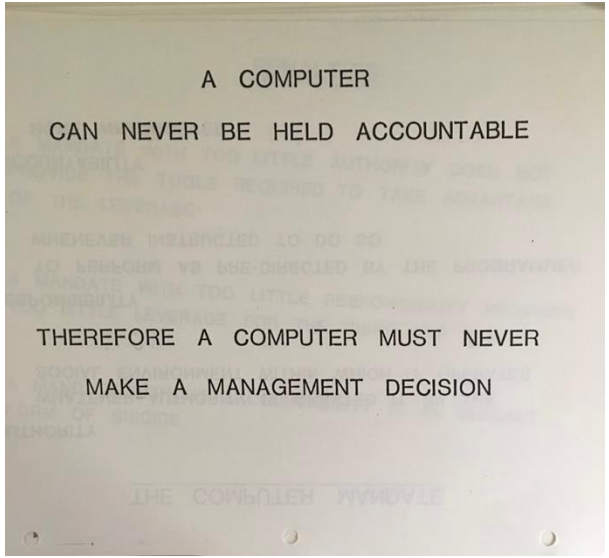
– AI recommends, human approves — real control unclear

Legal vs. Functional Views

– Law sees final acts; AI influences/shapes earlier steps

Blame Game

– Some always want a human liability sponge; others recognise AI is functionally making decisions



Famous (or infamous) IBM slide from 1979

The Magic Bullet of Meaningful Human Oversight

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 != Actual 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.

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.

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

CSIRO

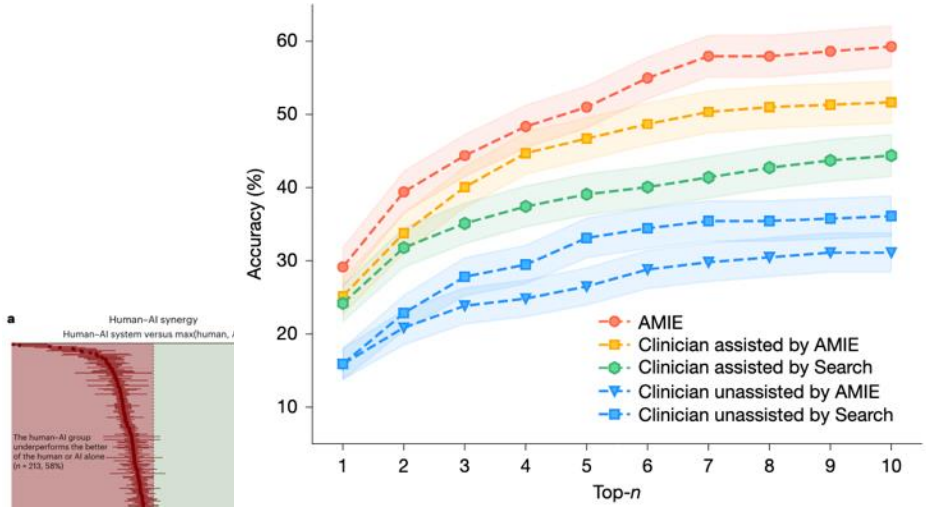
Myth #1 - Human in the Loop Guarantees Better Decisions

Automation bias

A human tendency to over-rely on AI recommendations, leading to degraded performance in human–AI teams even when the human alone would outperform the AI.

Automation aversion

A human tendency to under-use or reject AI recommendations, leading to degraded performance in human–AI teams even when the AI alone may outperform the human.



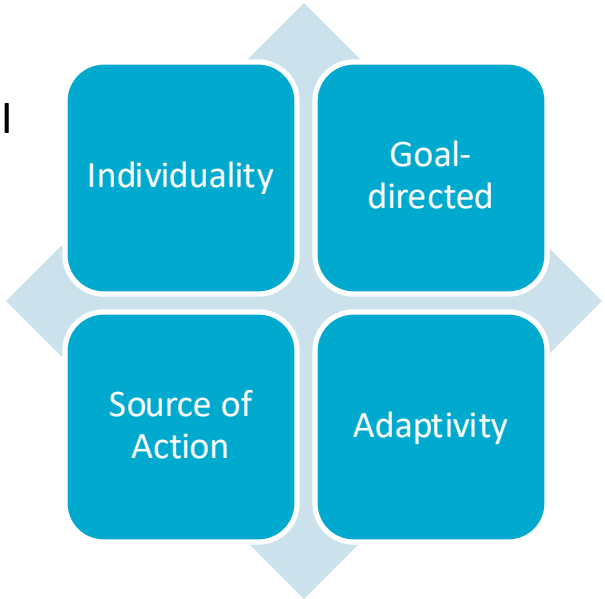
McDuff, D. *et al.* (2025) 'Towards accurate differential diagnosis with large language models', *Nature*, pp. 1–7. Available at: https://doi.org/10.1038/s41586-025-08869-4.

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

Myth #2 – Human Can Exert Actual Agency

Example: Weak human agency in LLM-enabled decision or content generation

System-boundary dependent: human within AI framing or just input source



AI sets most sub-goals instrumental goals: *"make it more interesting!"*

Illusion of source of control: prompting for AI-generated images

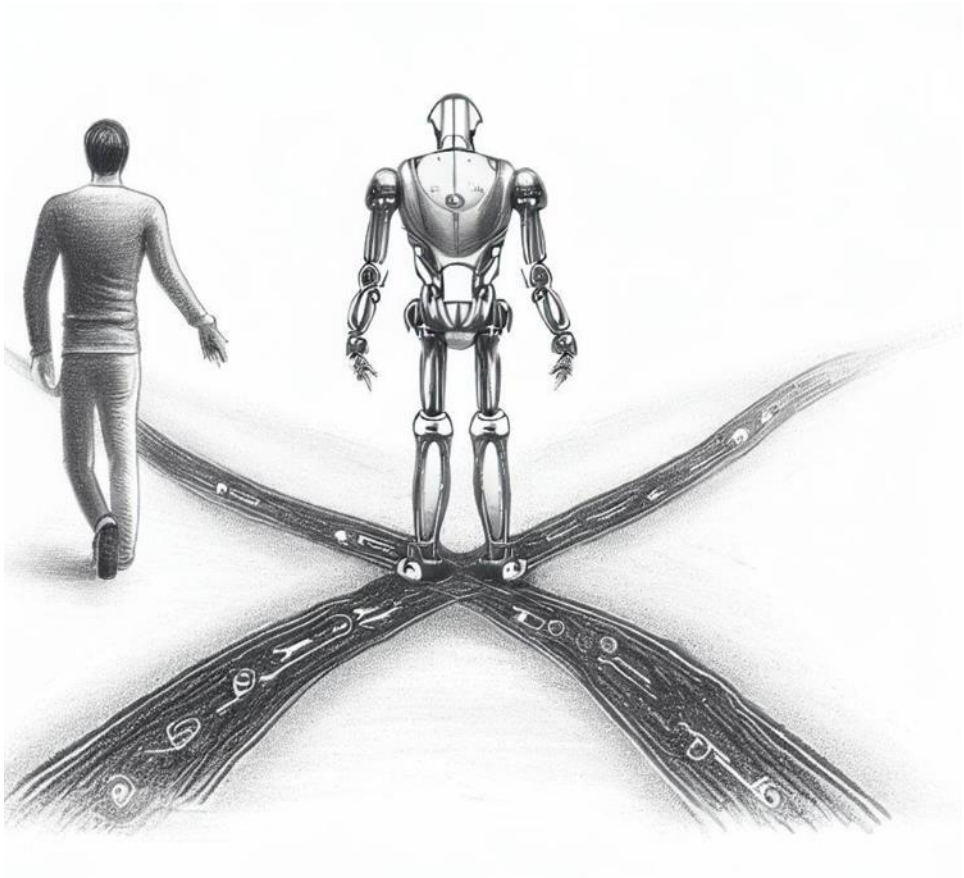
AI learns and adapts rules; human reacts without shaping them

Abel, D. __et al.__ (2025) 'Agency Is Frame-Dependent'. arXiv. https://doi.org/10.48550/arXiv.2502.04403

Myth #5 – Existence of Procedural Oversight is Sufficient

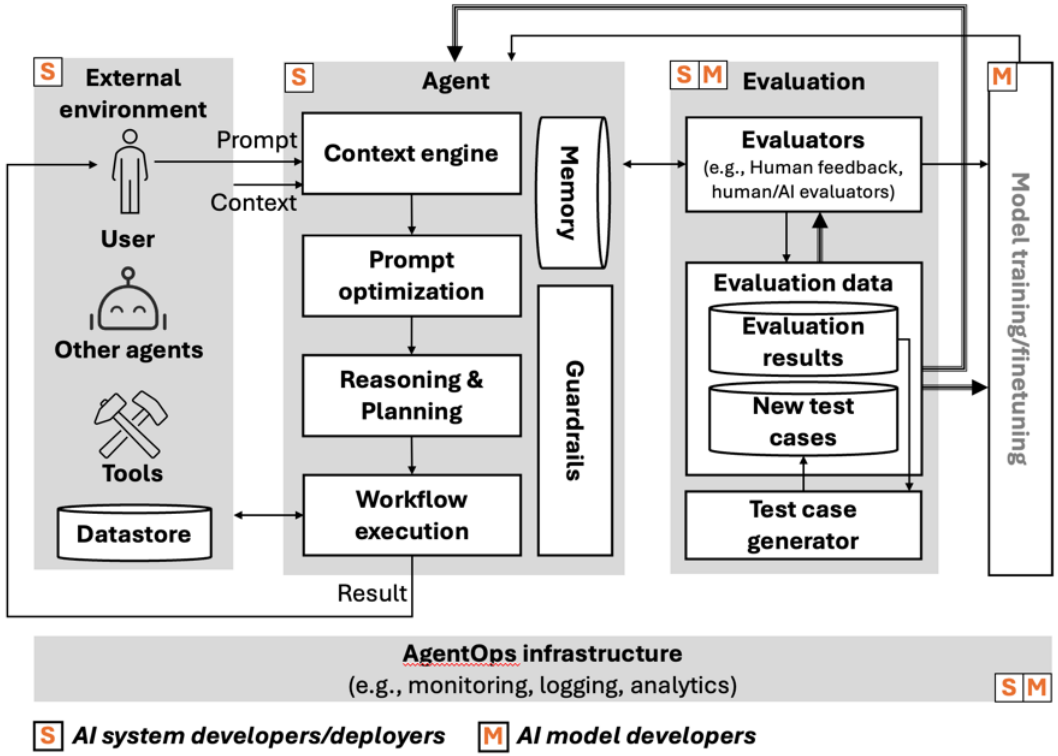
- **Human review** – present, but not evaluated for effectives
- **Transparency** – disclosed, but not comprehensible
- **Explainability** – required, but not faithful to model inner working or expert standards
- **Contestability** – appeal exists, but overwhelmed
- **Accountability** – named person, but no real control
- **Audit and reporting** – logged and reported, but not linked to failures or redress

Innovations
&
Case Study



Evaluation-Driven Performance & Meaningful Oversight

- Evaluation-driven **system-level** learning
- **Stackable** decision improvements
- **Effectiveness** evaluation of each oversight and safeguards
- **Continuous** Compliance & Conformance



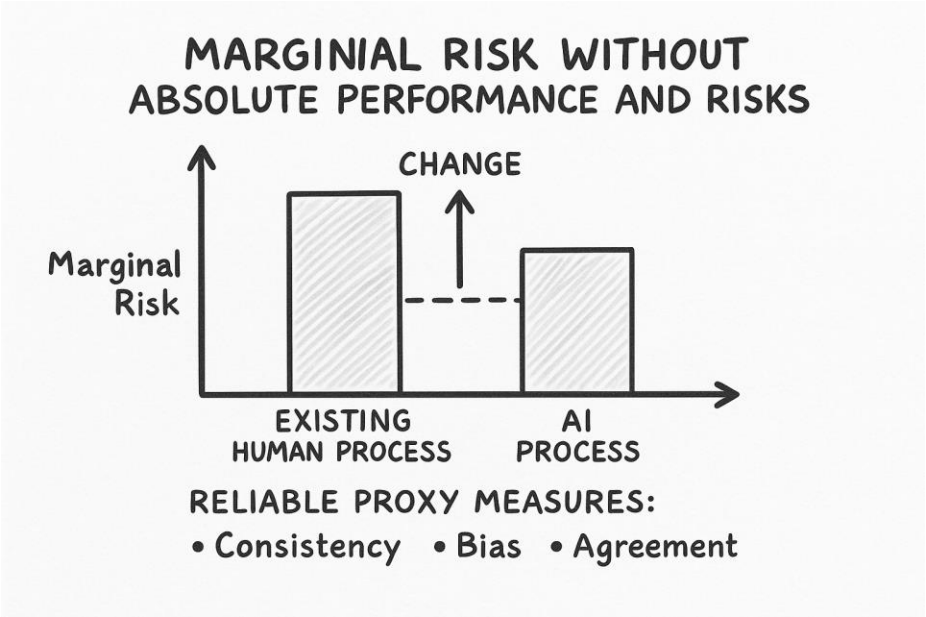
Data61 Work: Xia, B. *et al.* (2025) 'Evaluation-Driven Development of LLM Agents: A Process Model and Reference Architecture' https://doi.org/10.48550/arXiv.2411.13768

Meaningful Eval: Marginal Performance/Risk Assessment

without Ground Truth and Absolute Performance/Risk

- **Challenges**: No ground truth; No eval for existing process, stakeholder resistance; privacy issues

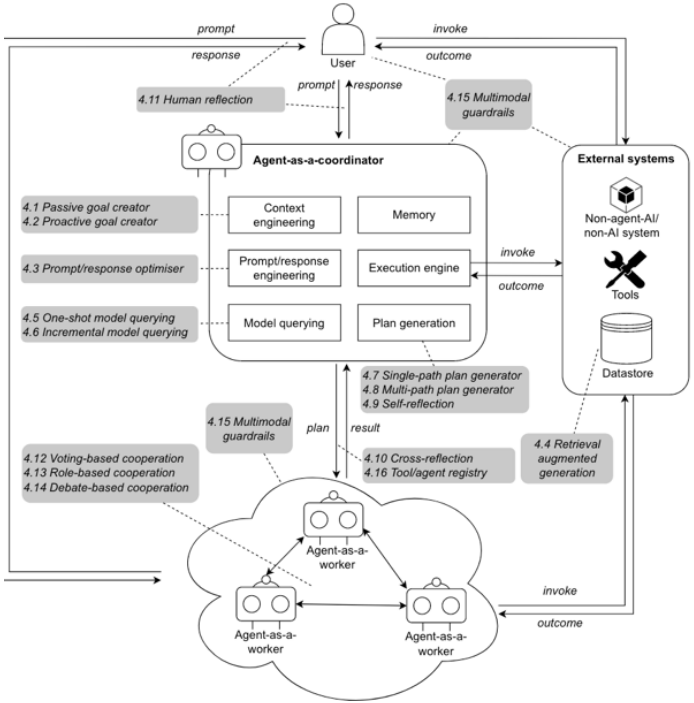
- **Solution**:
  - Marginal risk assessment using consistency, variance, bias…
  - Use existing KPIs
  - Selective downstream human audits



Data61 Work: Paper in progress; applied in major commercial projects

Design-time Oversight: AI/Agent Design Patterns

Trustworthy Whole out of Untrustworthy Parts



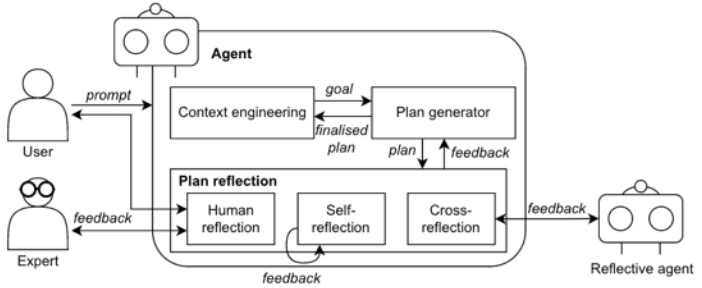


Figure 11: Plan reflection pattern.

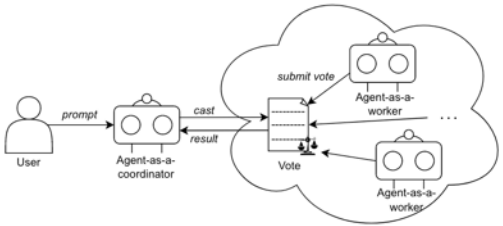


Figure 12: Voting-based cooperation.

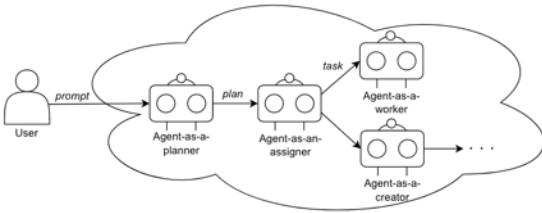
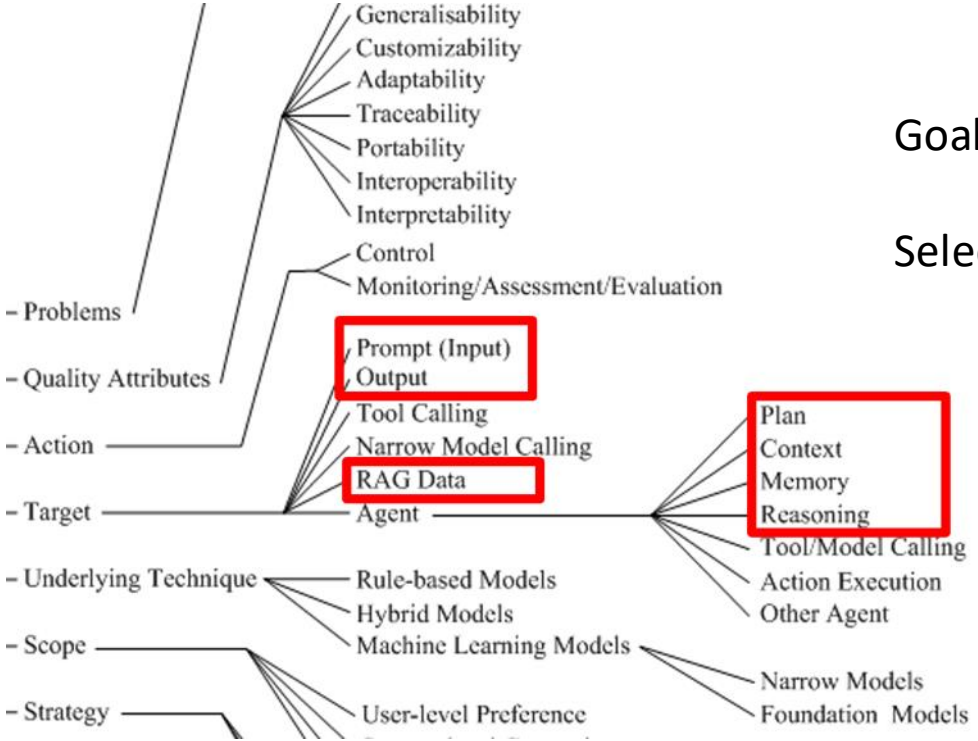


Figure 13: Role-based cooperation.

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

Runtime Oversight: Guardrails & Process Monitoring



Goals and reasoning traces, not just output

Selective cases, not everything

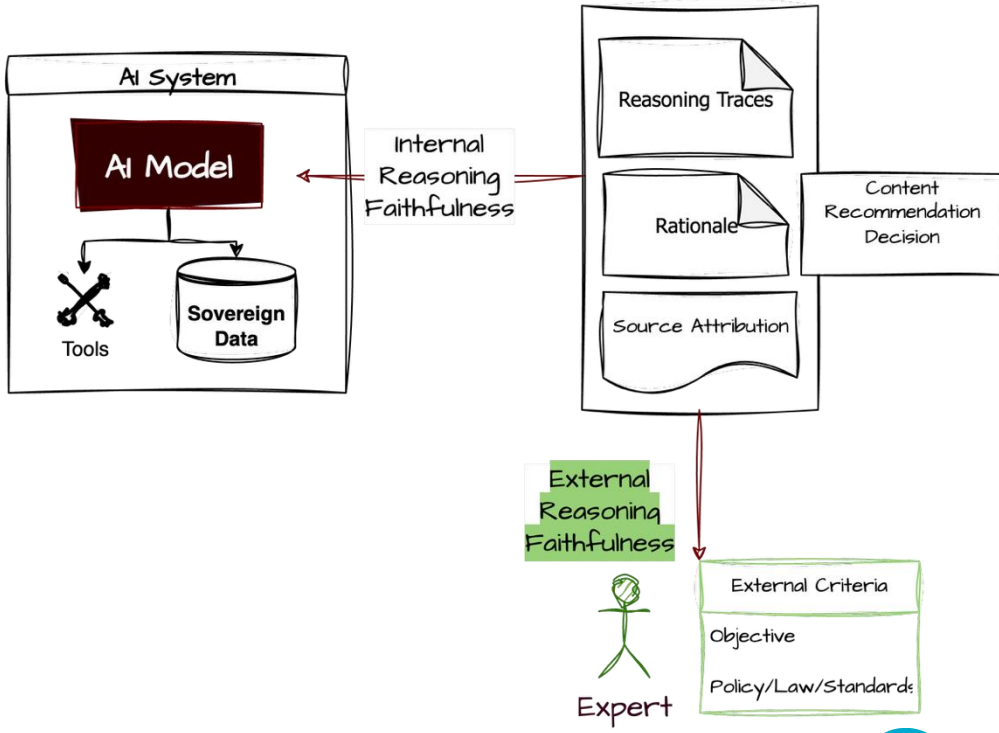
Data61 Work: Shamsujjoha, M. *et al.* (2024) 'Towards AI-Safety-by-Design: A Taxonomy of Runtime Guardrails in Foundation Model based Systems'. arXiv. Available at: https://doi.org/10.48550/arXiv.2408.02205.

Meaningful Explainability

Human for Reasoning/Rationale Evaluation

- **Challenges**: AI's plausible but flawed reasoning and justification

- **Solution**:
  - Flexible reasoning strategy patterns
  - Humans evaluate reasoning process and justification
  - AI judge for wider set of quality



Real World Case Study – Document Evaluation

Use Case: Tender, Grant, Proposal, Paper evaluation based on pre-defined criteria

Human Evaluator: subjective, slow, inconsistent across reviewers

Introducing AI Evaluator:

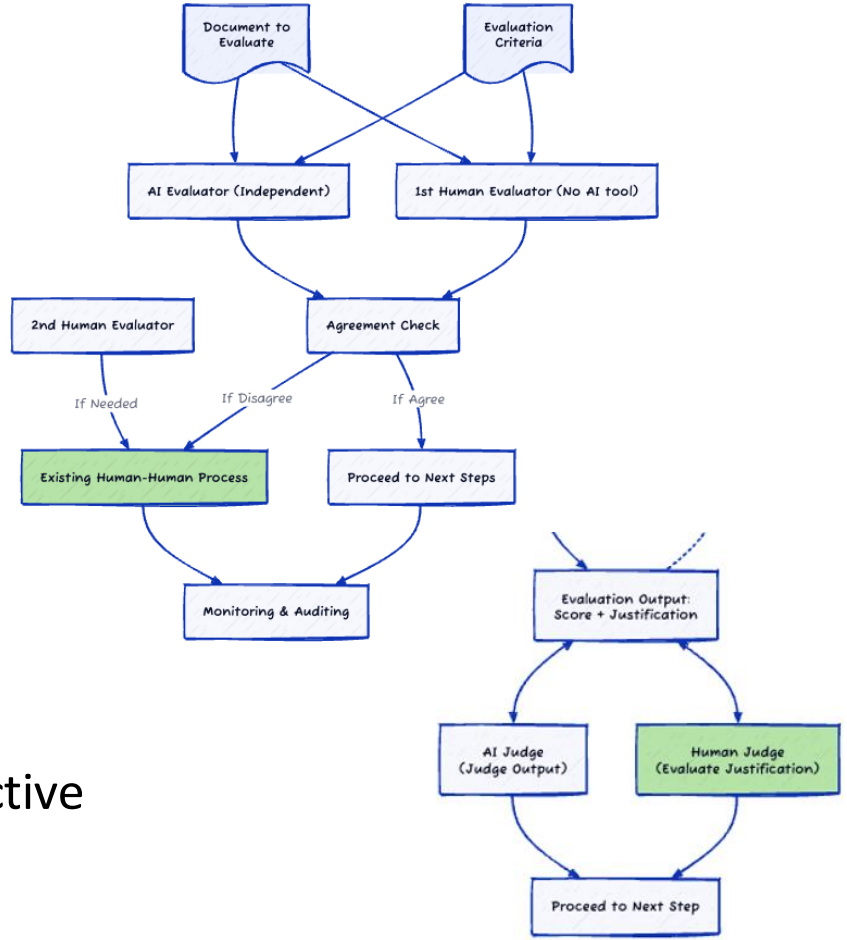
- Difficult to assess performance/risk without ground truth and baseline measurements
- Risks of human-AI interaction risk & reasoning faithfulness
- AI alone can be better than Human-AI: adverse attitudes towards automation

Meaningful Performance and Oversight

- System-level design
  - Independent human & AI (no human-AI)

- Evaluation
  - Marginal performance & risk assessment
  - Continuous compliance/conformance

- Meaningful Oversight
  - Design time: patterns, safety cases…
  - Runtime/Contest-time: guardrails, external reasoning faithfulness, monitoring and selective human audits



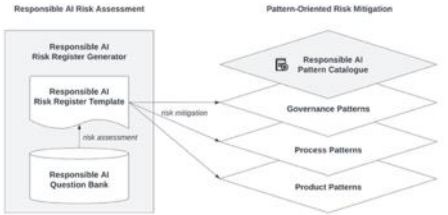
Innovations in Automated/AI-Enabled Decision Making

- Contentions & myths in AI/ADM

- Innovations

1. Evaluation-driven performance & risk
2. Marginal risk/performance evaluation
3. Design-time patterns
4. Runtime guardrails and process monitoring
5. External reasoning faithfulness
6. AgentOps: we are all managers now.

CSIRO's Best practice catalogue



deployer v1
developer v2 coming

Looking for Gov use cases and collaborators
Contact: liming.zhu@data61.csiro.au

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