Responsible, Safe, and Effective Use of AI for Complex Evaluation Tasks

Prof. Liming Zhu
Research Director, CSIRO's Data61
Conjoint Professor, UNSW

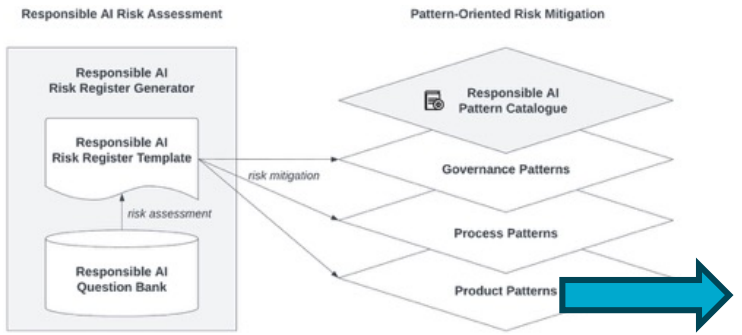
Expert in Working Groups

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

Responsible & Effective AI for Evaluation Tasks

- **Use Case**: Tender, Grant, Proposal, Paper evaluation based on pre-defined criteria
- Design AI-human collaboration for trustworthy and scalable eval.
- Focus on marginal risk and AI reasoning faithfulness
- Apply responsible AI best practices and adhere to gov policies, frameworks and standards

CSIRO's Best practice catalogue

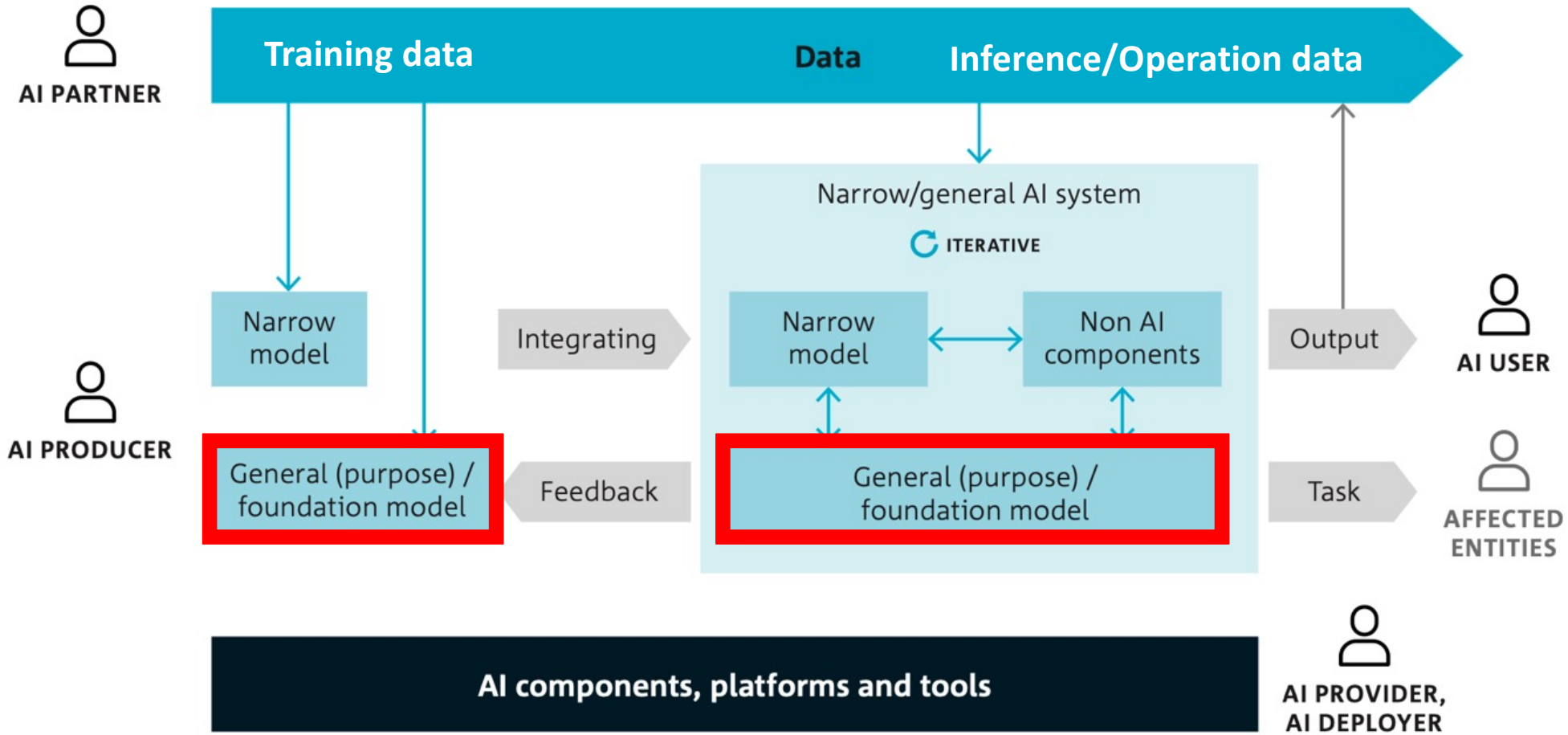


deployer v1
developer v2 coming

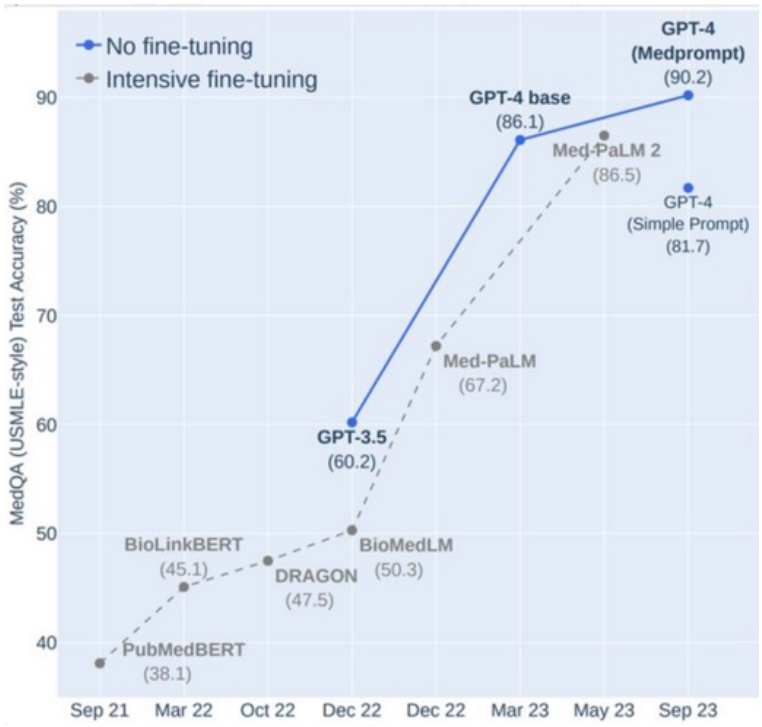
CSIRO

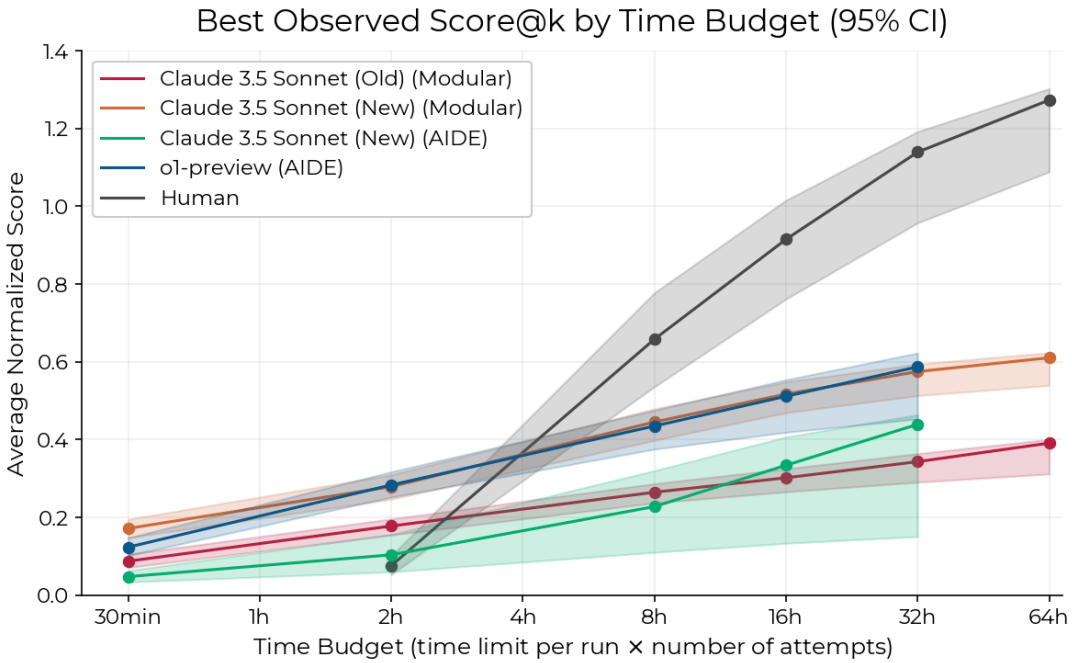
Trends & Challenges

AI Model vs. AI System



General AI vs Specific AI vs Human – Who Wins



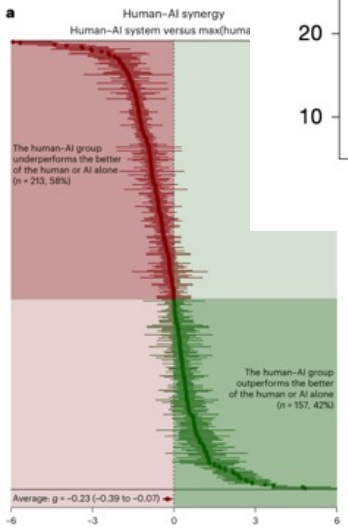


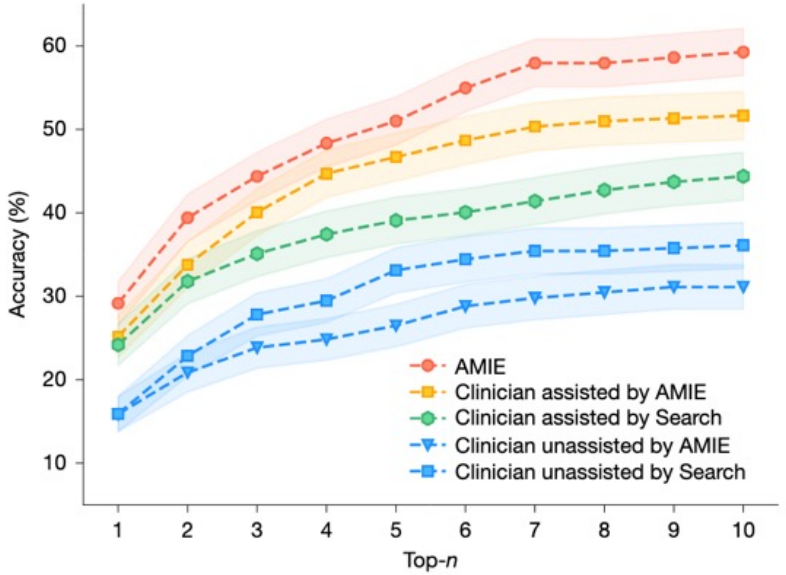
General AI often outperforms Specific AI

Human vs. AI - depends on time-budget

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/

Human + AI Performs Worse than AI/Human Alone?





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.

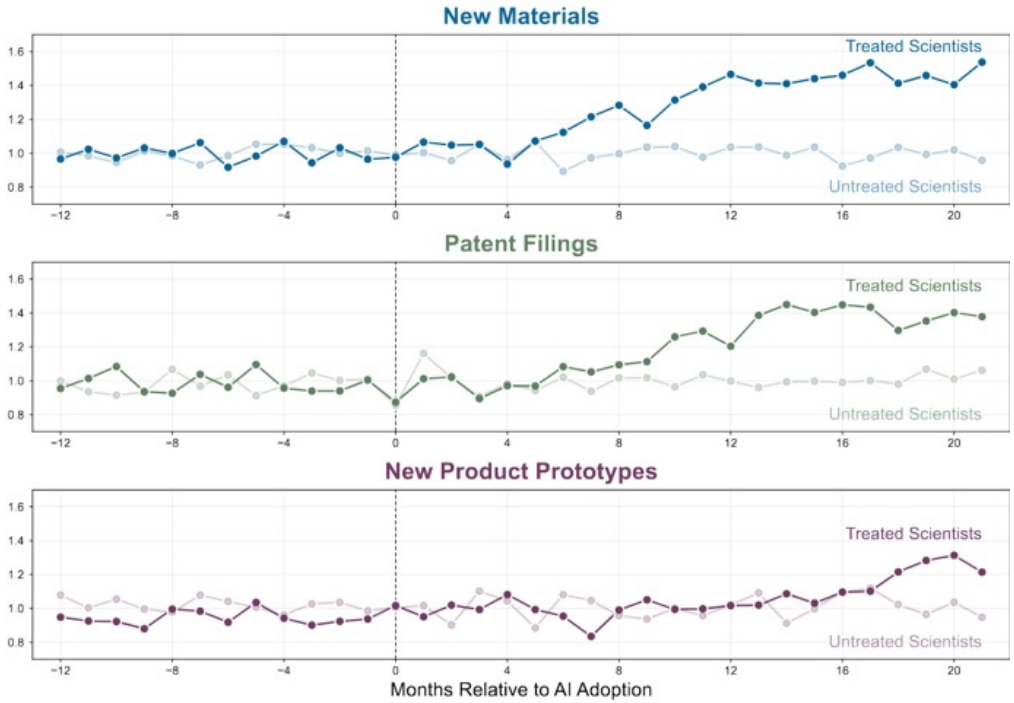
Vaccaro, M., Almaatouq, A. and Malone, T. (2024) 'When combinations of humans and AI are useful: A systematic review and meta-analysis', *Nature Human Behaviour*, pp. 1–11. https://doi.org/10.1038/s41562-024-02024-1

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***

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

Challenges in Evaluation: Before and After AI

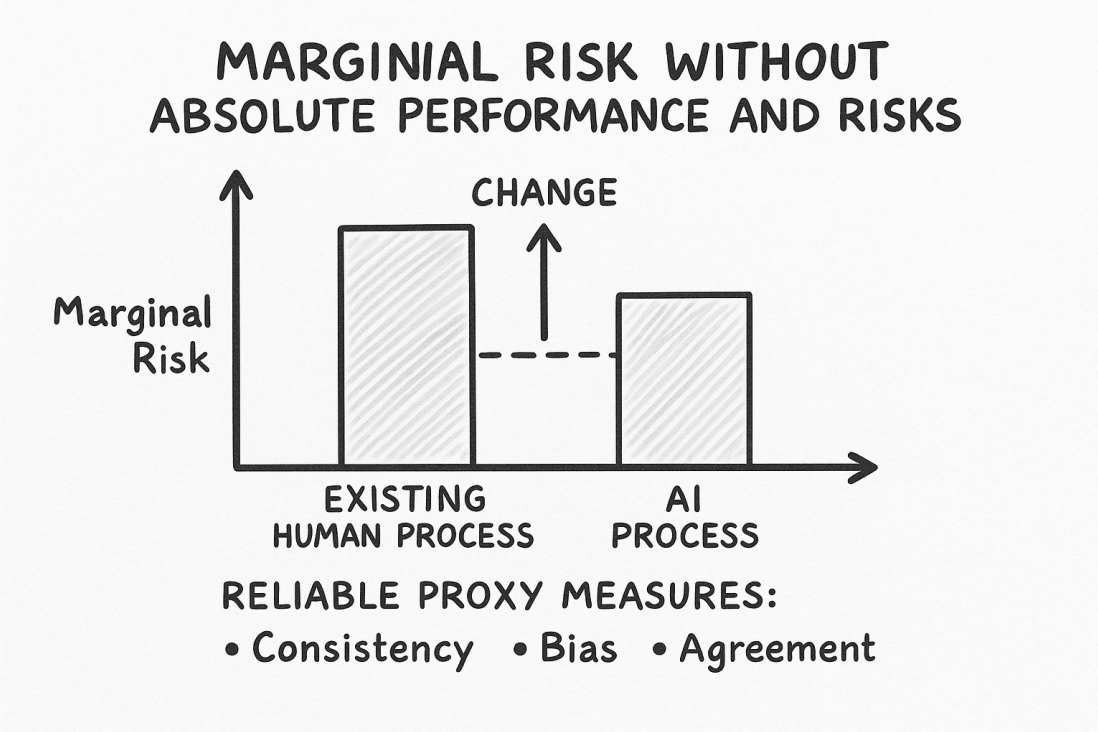
Human Evaluator: subjective, slow, inconsistent across reviewers

Introducing AI Evaluator:

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

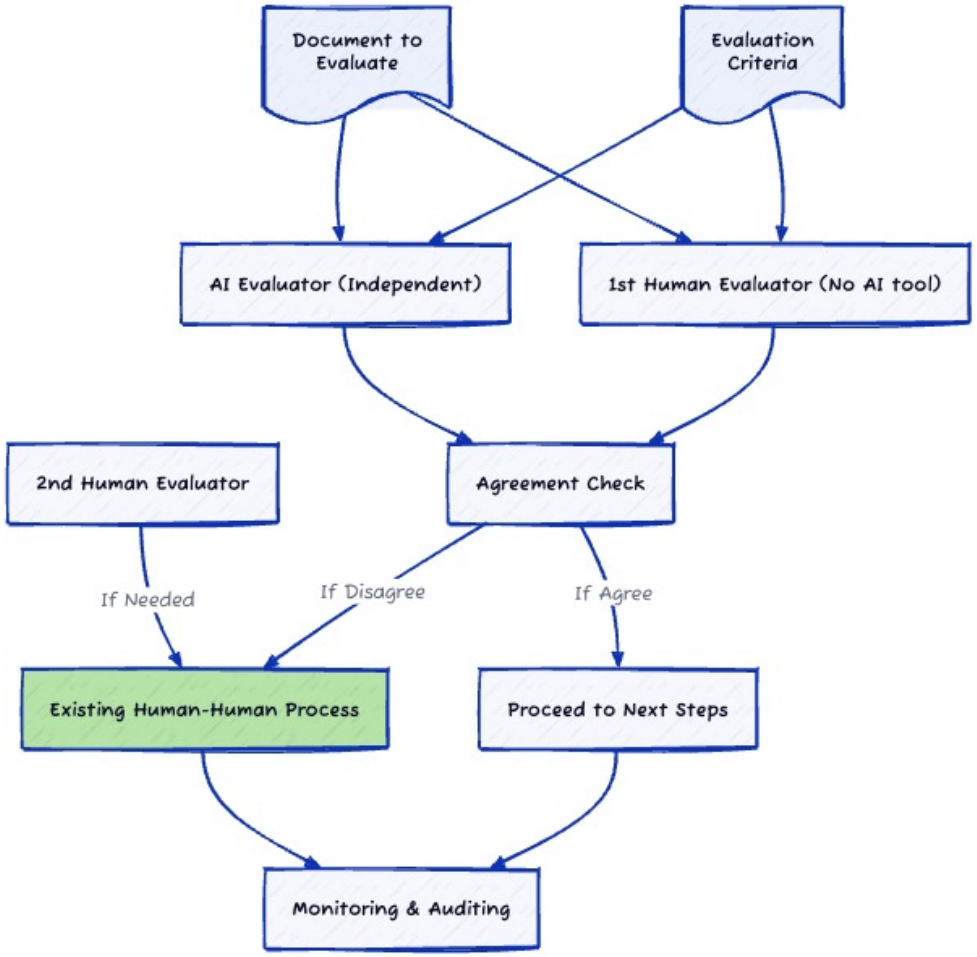
Solution 1: Practical Marginal Risk Assessment

- **Challenges**: No ground truth; stakeholder resistance; privacy issues
- **Solution**:
  - Marginal risk assessment using consistency, variance, bias
  - Use existing KPIs
  - Embed downstream human audits



Case Study 1 – Independent AI Evaluator

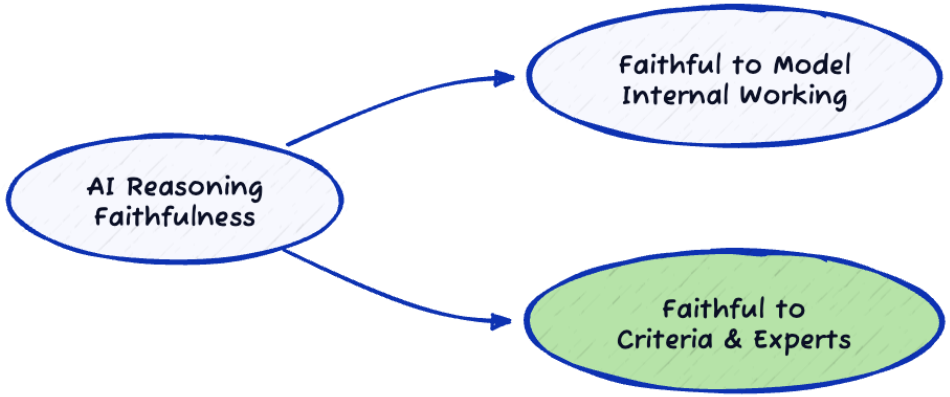
- Major gov agency: Application Evaluation
- Independent AI reviewer without influencing human judgment
- Disagreements escalated to second human evaluator
- Measured consistency, agreement and bias delta to validate risk
- Safeguards: escalation paths, clean human judgment



Solution 2 – Human for Justification 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

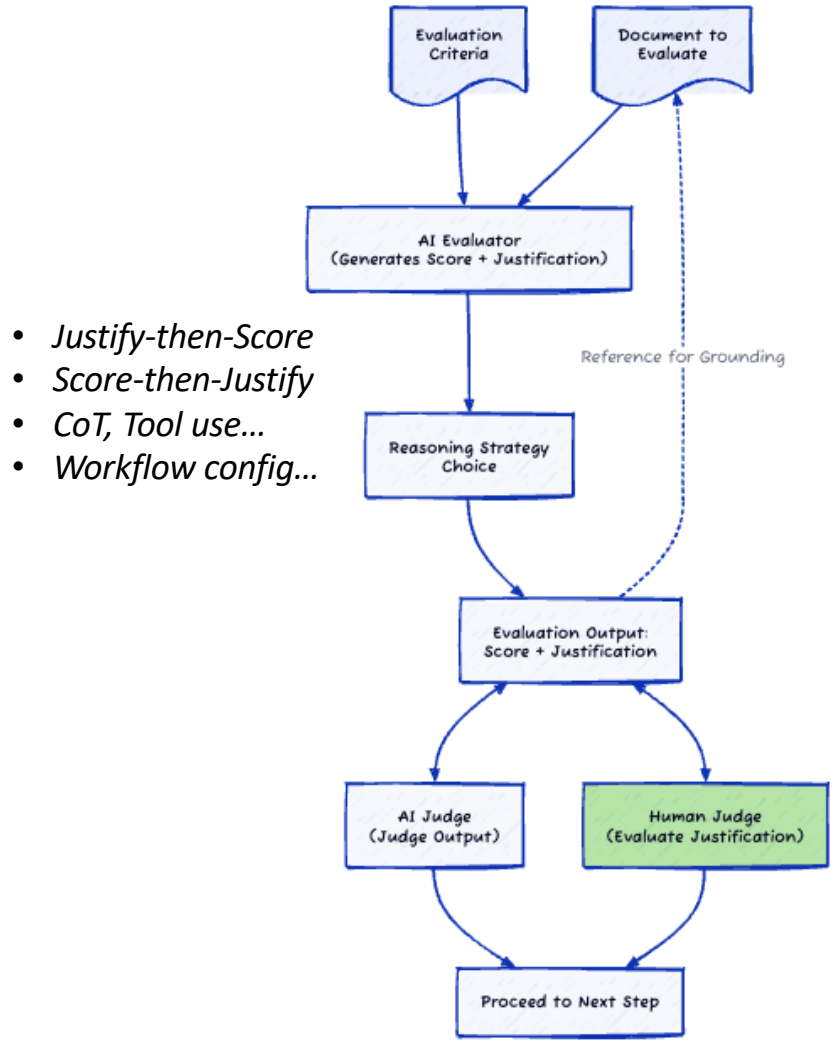


Case Study 2 – AI Peer Reviewer

- AI scores and drafts structured evaluation output with evidence links

- AI judge for wider set of quality
- Human judge for evaluating reasoning process and justifications
- Humans have final decision authority

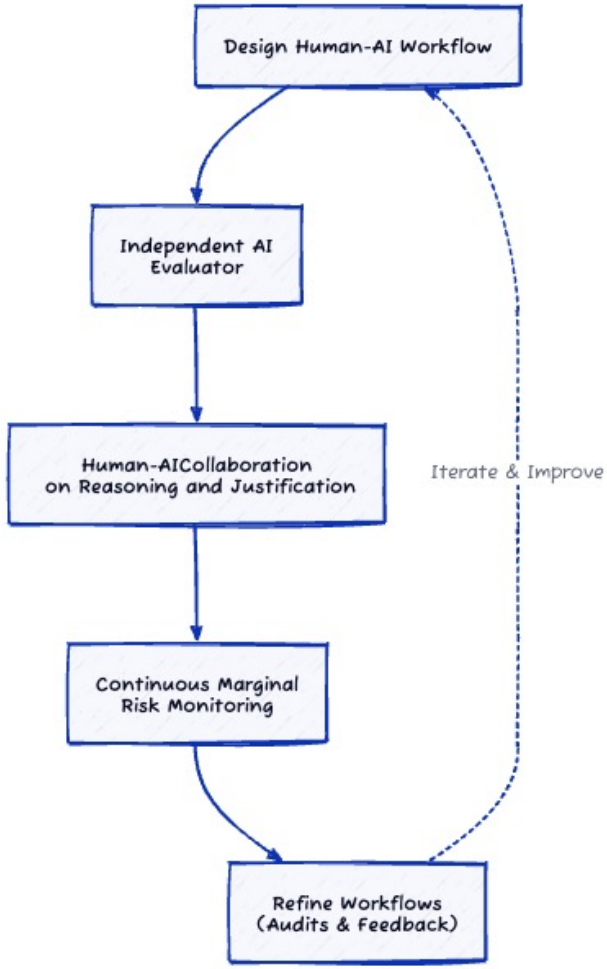
- Audits detect and correct marginal risks



- *Justify-then-Score*
- *Score-then-Justify*
- *CoT, Tool use…*
- *Workflow config…*

Unified Safe Human-AI Evaluation Design

- Start with independent AI evaluation to build understanding and trust with simplicity
- Introduce controlled human-AI collaboration on reasoning and justification
- Monitor "marginal risk" continuously
- Refine workflows based on audits and feedback



Responsible & Effective AI for Evaluation

Trends and Challenges

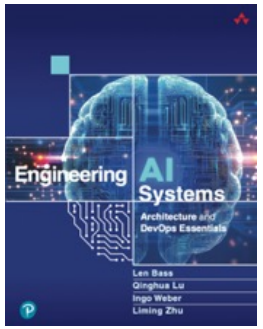
•AI alone may outperform human-AI collaboration
•Complex interactions and reasoning faithfulness
•Lack of ground truth complicates evaluation

Solutions and Case Studies

• Marginal risk assessment: Independent AI Evaluator
• Human validation of Reasoning: AI Peer Reviewer
• Iterative improvement and continuous risk monitoring

Responsible AI and AI Engineering

•Standards, policy, and framework alignment
•Applied engineering patterns for trustworthy AI



Looking for Gov and Industry use cases and collaborators

Contact: liming.zhu@data61.csiro.au

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