arc training centre for **information resilience**

Bridging the Gap: Human in the loop for Information Resilience

This Centre is funded by the Australian Government through the Australian Research Council Industrial Transformation Research Program in partnership with the University of Queensland and Swinburne University of Technology.

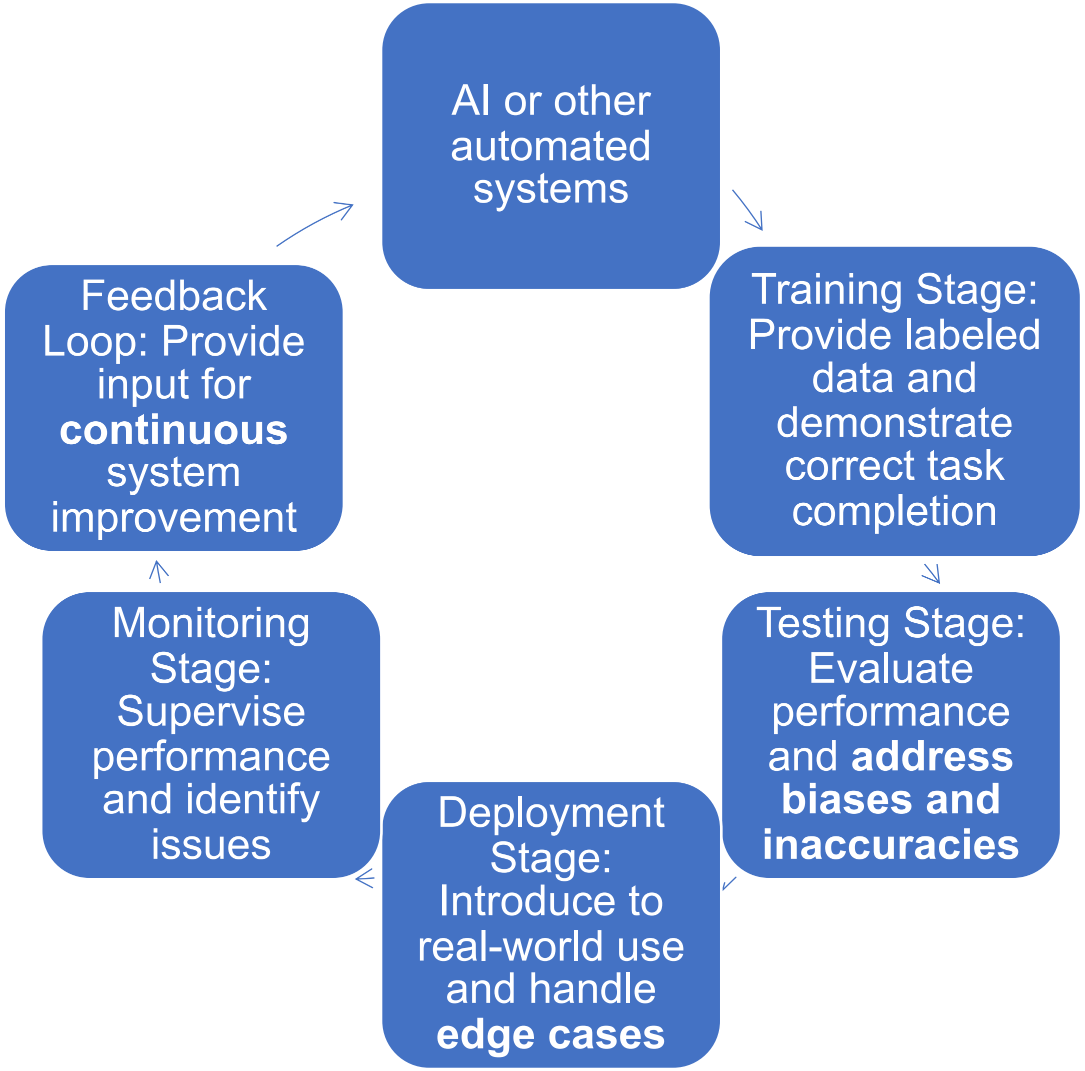
Introduction to Human-in-the-Loop (HITL)

- Definition: combining human judgment and expertise with automated systems to improve the **accuracy, reliability, and trust** of information processing and decision-making

- Example: **'Gender Shades'** project in 2018 shows three major gender classification algorithms, including those from IBM and Microsoft, **performed worst** on darker-skinned females, with error rates up to 34% higher than for lighter-skinned males [1]

- HITL helps **identify and address such biases**

HITL Process



AI or other automated systems

Training Stage: Provide labeled data and demonstrate correct task completion

Testing Stage: Evaluate performance and **address biases and inaccuracies**

Deployment Stage: Introduce to real-world use and handle **edge cases**

Monitoring Stage: Supervise performance and identify issues

Feedback Loop: Provide input for **continuous** system improvement

Humans are involved in every stage of AI development and operation.

Applications in Information Resilience

- **Expert input**: Meta's **expert-driven content moderation** during 2023 Israel-Hamas conflict to adapt new threats [2].
- **Crowdsourcing**: Twitter(Now X)'s Birdwatch (Now Community Notes) for misinformation to improve accuracy and **context understanding** [3].
- **End User Feedback**: End users identified historically inaccurate images generated by Google's Gemini AI tool, highlighting the **need for ongoing human oversight** [4].

Benefits:

- **Experts** provide **specialized knowledge** in critical situations
- **Crowdsourcing** leverages **diverse perspectives** for complex and multifaceted issues
- **End user feedback** catches **real-world errors** and guides improvements

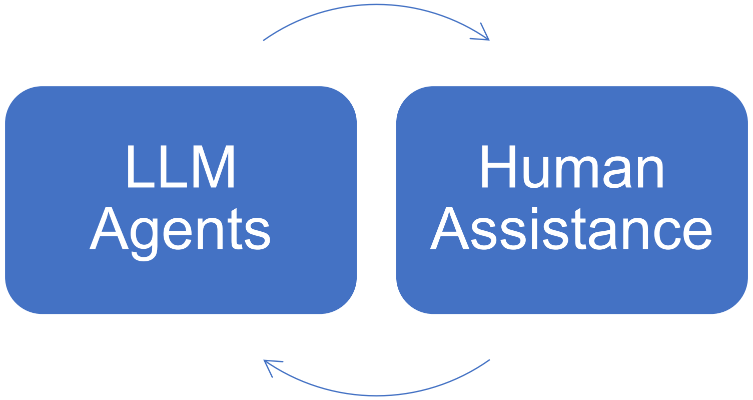
HITL in LLM

LLM Development [5]:

- OpenAI's GPT models use **Reinforcement Learning from Human Feedback** (RLHF) for **alignment and safety improvements**
- Human evaluators essential in testing and refining models
- RLHF trains models using **human preferences as a reward signal**
- Results in **better performance and alignment with user instructions**

Human-In-The-Loop LLM Agents [6]

- HITL tool allows agents to **reach out to humans** when existing tools are insufficient
- Improves on traditional chatbot models by allowing multiple iterations of human support, transparent to the end-user
- **Pushing the boundaries** of what AI can achieve while ensuring human judgment remains a crucial part of the process



Challenges and Solutions in HITL Approaches

- **Balancing efficiency** with **human involvement**: **Adaptive** systems based on **task complexity**
- **Training** and **maintaining** expert reviewers: Comprehensive training programs and clear career paths
- Ensuring **consistency** across human judgments**: Clear guidelines** and regular **calibration** sessions
- **Human Errors** Solution: Robust quality control and **cross-checking** procedures
- **Higher Expenses** Solution: Thorough cost-benefit analyses and **resource optimization**

Australian Government
Australian Research Council

Conclusion

HITL is crucial for creating **trustworthy, adaptable** AI systems

- Sources: Experts, Crowdsourcing, End Users
- Stages: Training, Testing, Deployment, Monitoring
- Benefits: **Specialized knowledge, Diverse perspectives, Real-world error detection, Continuous improvement**

Thank you !

Australian Government
Australian Research Council

1 http://gendershades.org/overview.html

2 https://about.fb.com/news/2023/10/metas-efforts-regarding-israel-hamas-war/

3 https://help.x.com/en/using-x/community-notes

4 https://www.theguardian.com/technology/2024/mar/08/we-definitely-messed-up-why-did-google-ai-tool-make-offensive-historical-images

5 https://openai.com/index/instruction-following/

6 https://cobusgreyling.medium.com/human-in-the-loop-llm-agents-e0a046c1ec26