DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

Mental Model Mismatch in Human-AI Interactions

Prof. Mohamed Abdelrazek

Background and problem statement

As we move AI from co-pilot to decision-support and autonomous agents, failures increasingly come from human-AI interaction breakdowns, not just model accuracy. A recurring **root cause is mental model mismatch:** humans form beliefs about how the AI works, what it "knows," and when it will fail, but those beliefs diverge from the AI's actual behaviour. This leads to:

- Over-trust (misuse), under-trust (disuse), and fragile reliance
- Poor handoffs and "automation surprise"
- Incorrect attribution of causality, intent, and accountability.
- Unsafe outcomes in high-stakes workflows (health, finance, cyber, aviation, incident response).

Aim

This project proposes a rigorous, measurable study of mental model mismatch and interventions that reduce it. Identify the dominant mismatch types (causality vs correlation, confidence miscalibration, explanation illusions, objective-function misunderstanding, mode confusion). Test interventions that align human expectations with AI behaviour. Deliver demos for simple trustworthy human-AI teaming.

Tasks

- Mismatch types vs misuse/disuse, automation surprise, and downstream errors
- Quantify human mental models of AI and compute mismatch against the AI's actual decision logic/behavioural patterns
- Strategies to improve calibration

How will this opportunity benefit you?

Skill development in the following areas:

- Machine Learning
- Data analysis
- Programming

Other details

- Located on-campus at Burwood
- 14 weeks in duration (Unpaid)

- Undertaken as 3 days per week for a total of ~320 hours
- To be commenced within the 12-week trimester period