AI Safety Workshop Programme
27th March 2026, 14.00-17.00
School for Data Science and Computational Thinking, Stellenbosch University (in person)

The Policy Innovation Lab and AI Safety South Africa are proud to present a series of talks on the real-world challenges of keeping AI safe, aligned, and secure. We will be discussing everything from the safety of delegating deliberation tasks to AI to the intricacies of moral frameworks needed to guide AI systems.

Event Schedule

14:00 – Welcome
14:10 – *Embedded Adversarial Exploitation of Information Sharing in Multi-agent LLM Systems*, Omer Kamal Ali Ebead
14:30 – *Mechanistic Interpretability in AI Safety and the Shift from Correlation to Causation,* Gray Manicom
14:50 – *ASIF: A Descriptive Moral Scaffold for Reducing Value Underspecification in AI Alignment*, Willem Fourie and Isabel Ray
15:10 – *MoralGym: RL Fine-Tuning of LLM Agents on Social Dilemma Games*, Yves Bicker
15:30 – Tea break
15:50 – D*elegating Deliberation to Agents,* Joseph Low and Oscar Duys
16:10 – *Similarity as a Signal: Do AI Agents Cooperate More When They Know They're Alike?*, Akash Kundu
16:30 – Thank you and goodbye
17:00 – Event Finished

Presentation Abstracts

Embedded Adversarial Exploitation of Information Sharing in Multi-agent LLM Systems, Omer Kamal Ali Ebead

This research embeds adversarial agents into cooperative multi agent LLM simulations and measures their ability to extract private user data through social engineering. Context priming is seen to increase behavioural compliance even when safety filters block actual data output. The project is currently establishing a new benchmark for multi-agent cooperation and privacy against embedded adversarials.

Mechanistic Interpretability in AI Safety and the Shift from Correlation to Causation, Gray Manicom

This talk explores the role of mechanistic interpretability in AI safety, comparing its focus on causal insights with more traditional, correlation-based methods. It also examines the localisation of AI models to African contexts through fine-tuning and distillation. The discussion aims to determine whether these adaptations represent fundamental changes to the network or remain merely surface-level.

ASIF: A Descriptive Moral Scaffold for Reducing Value Underspecification in AI Alignment, Willem Fourie and Isabel Ray

ASIF is a new theoretical framework designed to move AI alignment away from implicit rewards and toward explicit, auditable governance. By breaking down reward models into interpretable components like moral orientations and situational contexts, it addresses the issue of value underspecification. This approach allows for the inspection of moral trade-offs without forcing a commitment to a single normative doctrine.

MoralGym: RL Fine-Tuning of LLM Agents on Social Dilemma Games, Yves Bicker

This research investigates whether procedural RL fine-tuning of LLM agents on N-player social dilemmas can induce transferable pro-social dispositions that generalise to complex multi-agent governance settings. Building on prior moral alignment work, the study introduces a diverse curriculum of varied games and a rich opponent league, with the goal of learning social behaviour that transfers to richer settings such as commons resource dilemmas.

D*elegating Deliberation to Agents,* Joseph Low and Oscar Duys

This presentation examines the technical and safety implications of delegating complex deliberation tasks to AI agents. It explores how automated reasoning can be structured to remain reliable when operating independently. The session highlights potential risks and best practices for maintaining oversight in delegated systems.

Similarity as a Signal: Do AI Agents Cooperate More When They Know They're Alike?, Akash Kundu

This early-stage investigation explores whether perceived similarity between language models influences their cooperative behaviour in strategic settings. The talk shares preliminary experimental findings and open questions. The session highlights current research directions and welcomes feedback on these evolving observations.