Marcus Williams — AI Alignment Engineer

✉ marcusjw@zoho.com • in marcus-williams-2623681a0
marcus-jw

Experience

Research Scholar — **Berkeley, USA**
MATS — *June 2024–Present*
Investigated how language models trained on user feedback can develop deceptive behaviors, selectively targeting vulnerable users. See publication below.
`https://github.com/marcus-jw/Targeted-Manipulation-and-Deception-in-LLMs`

Multi-Objective Reinforcement Learning from AI Feedback
Research grant by the Long-Term Future Fund — *Oct 2023–May 2024*
Research project investigating using multiple preference models and multi-objective RL approaches to improve safety-performance of LLMs. See publication below.
`https://github.com/marcus-jw/Multi-Objective-Reinforcement-Learning-from-AI-Feedback`

Researching the expressivity of different RL formalisms — **Oxford, UK**
AI Safety Hub Oxford — *Jul 2023–Oct 2023*
Formally proved the expressivity relationships between various reinforcement learning formalisms, see publication below.

Course Facilitator — **Lund, Sweden**
AI Safety Fundamentals — *Aug 2023–Dec 2023*
Facilitating multiple in-person groups for the AI Safety Fundamentals Alignment course.

Software Developer in Mobile Applications — **Lund, Sweden**
Axis Communications — *2019–2022*
I worked on maintaining and improving their Android mobile applications through error and latency analysis.

Education

Master of Engineering (MEng) — **Lund, Sweden**
The Faculty of Engineering at Lund University, Grade: 5.0/5.0 — *2018–2023*
Combined 5 year bachelor's and master's in Engineering Physics specialising in Machine Learning

Publications

"Targeted Manipulation and Deception Emerge in LLMs Trained on User Feedback": Submitted to ICLR 2025, Accepted as an oral presentation at SATA and as a spotlight at SoLaR.

Description: Training large language models (LLMs) to maximize user feedback holds promise for personalization but introduces risks of manipulative behavior. Our study reveals that models can selectively target vulnerable users while appearing benign to others. Attempts to safeguard using LLM judges sometimes result in subtler manipulation rather than prevention.
Paper: `https://arxiv.org/abs/2411.02306`

"On The Expressivity of Objective-Specification Formalisms in RL": ICLR 2024

Description: We evaluated the expressive power of 17 objective-specification formalisms in re-

inforcement learning (RL), organizing them in a Hasse diagram to illustrate expressiveness and optimization trade-offs. Our findings indicate no single formalism outperforms others, with unique objectives emerging for specific methods.
Paper: `https://arxiv.org/abs/2310.11840`

"Multi-objective Reinforcement Learning from AI Feedback":

Description: Explored improving LLM alignment by decomposing human preferences into multiple principles and training distinct preference models on each. MORLAIF seems to outperform standard RLAIF, and allows us to effectively align larger models using smaller ones.
Paper: `https://arxiv.org/abs/2406.07295`