🧠 AI Safety & Alignment – Final Project Presentations

📅 **Date**: 23 April 2025
🕜 **Time**: 1:30 PM – 4:20 PM
📍 **Location**: FC006
🧑‍🏫 **Facilitator**: Prof. Peter Henderson, Teaching Assistants - Amaya Dharmasiri, Lucy He

---

🎯 Presentation Guidelines

- Each group will have **12 minutes total**:
    1. **9 minutes** to present
    2. **3 minutes** for Q&A
- Presentations should include:
    1. **Project Motivation** – What problem are you addressing? Why is it important in the context of the topics discussed in class?
    2. **Approach** – Methods, models, or theoretical framing used
    3. **Key Findings** – Results, insights, or progress made
    4. **Discussion** – Limitations, open questions, or ethical reflections
    5. **Future Work** – Where this could go next
- **Slides are required**. Please send a PDF version of your slides to the TAs by 9.00 am on 23 April 2025.
- Keep the audience in mind: Your peers are technically literate but may not know your specific sub-area.
- We will keep time strictly.

📝 Evaluation Criteria (Total: 10 Points)

Category	Points	Description
Clarity of Presentation	2	Clear explanation, structure, and visuals
Technical Depth	3	Sophistication of methods and analysis
Alignment Relevance	2	How clearly the project relates to AI safety/alignment challenges
Engagement	2	Handling of Q&A, audience connection
Originality	1	Creativity in approach or framing

⏱️ Presentation Schedule (12 Groups)

Time	Project	Group members		
1:30 – 1:32	Opening & Intro			
1:32 – 1:44	Adversarial Reinforcement Learning with Stackelberg Actor-Critic: A Multiagent Approach for Autonomous Systems	Jarod Wille	Justin Wang	
1:44 – 1:56	Analyzing Prompt Landscapes and Elicitation Probabilities in Large Language Models	Zerui Cheng	Zeyu Shen	Siqing Zou
1:56 – 2:08	Investigating Shallow Preference Signals in LLM Alignment	James Zhang	Cathy Di	Jiahao Qiu
2:08 – 2:20	Jailbreaking Large Language Models with Low Cost	Abhi Vellore	Yihan Wang	Tanvi Haldiya
2:20 – 2:32	Evaluating the Cybersecurity Risk of Self- Improving Agent	Boyi Wei	Benedikt Stroebl	
2:32 – 2:44	Emergent Alignment: Exploring LLM Alignment Through Narrow Fine-tuning on Behavioral Games	Aabid Ismail	Leo Stepanewk	
2:44 – 2:54	Break			
2:54 – 3:06	A therapist by any other name: investigating user social relationships with ChatGPT	Patty Liu	Kara Schechtman	Kylie Zhang
3:06 – 3:18	Measuring Variations in Moral Foundations of Large Language Models	Peter Kirgis	Harish Krishnakumar	Donggeon Oh
3:18 – 3:30	FairTune: When Can Fine-tuning on Benign Data Break Fair- ness?	Tanvi Namjoshi	Jane Castleman	
3:30 – 3:42	AI for Mental Health Counseling: Evaluating LLM Models and Prompt Strategies	Alexandra Maxwell	Maria Morales-Salgado	Michael Chapman
3:42 – 3:54	Reducing Hallucination in Scientific Question Answering	Mahsa Bastankhah		
3:54 – 4:06	Sequential Processing in Vision Language Models Improves Relational Reasoning	Udith Haputhanthri		

4:06 - 4.18	Does Tree-of-Thought Deliberation Morally Align Large Language Models?	Gili Karni	Max Gupta	Akash Selvakumar
4:18 – 4:20	Wrap-Up, Final Remarks, Buffer Time			