UBC AI Safety

# Session 2: Scheming and Deception

Location: IKB 263 Date: October 8th, 2025 Time: 5-7PM

Today's session will feature a deep dive on **scheming**: when AIs deliberately conceal misalignment in order to achieve their goals. We'll examine a few papers that ask whether and how LLMs exhibit this behavior.

## Session Activities

1. **Icebreaker (5 minutes)**
    - What's your name, what area do you work/study in?
    - What is a TV show/movie/podcast you have been enjoying recently?
    - What's your roommate horror story? (if you have one)

2. **Topic Overview (25 minutes)**
    - Read this primer (pages 6-10)
        - While you read, drop any questions that arise for you in the table below.
    - Discussion Questions:
        - Is this a behavior you'd expect today's frontier models to exhibit?
            - With what probability?
            - Under what circumstances?
            - If they did, how concerning would this be?
        - Is it more important to detect scheming or to design AIs that don't scheme?

| Name | Questions |
|---|---|
| | |
| | |
| | |
| | |

3. **Speed-Reading + Critical Analysis (30 minutes)**
   - **Paper Selection (5 minutes)**
     - Groups choose from the following list:
       - Frontier Models are Capable of In-context Scheming Ricky
       - AI Sandbagging: Language Models can Strategically Underperform on Evaluations (Belal)
       - Detecting Strategic Deception Using Linear Probes Britton
       - Detecting and reducing scheming in AI models (links to a blog post summarizing the paper)
   - **Reading (20 minutes)**
     - Groups can split the reading however works best - by sections, methods/results, etc.
     - Focus on understanding: main claims, methods, key results
   - **Partner Explanations (5 minutes)**
     - Each person explains their section/findings to their group
   - **Presentation Prep (15 minutes)**
     - Together, prepare a short verbal presentation covering at least:
       - Main contribution/claims of the paper
       - One methodological weakness or questionable assumption
       - Three questions you have about the paper

| Paper | Notes |
|---|---|
| Frontier Models are Capable of In-context Scheming | |
| AI Sandbagging: Language Models can Strategically Underperform on Evaluations | |
| Detecting Strategic Deception Using Linear Probes | |

| | |
|---|---|
| Detecting and reducing scheming in AI models | |

4. **Discussion + Mapping (40 minutes)**

    - **Paper Presentations (30 minutes)**
        - Each group briefly explains their paper to the rest of the group
        - As groups present, we will discuss them as a group
    - **Funding Vote (10 minutes)**
        - Which paper would you fund for follow-up work (and why)?

5. **Wrap-up (5 minutes)**

    - Next session logistics
    - Contact info sharing for those interested

## Resources & Additional Readings

- Technical Reading Group Background Readings - For those new to reading AI safety research.
- *"Scheming AIs: Will AIs fake alignment during training in order to get power?"*: Summary / Full Report (*very* thorough) - Explains reasons for and against expecting scheming to emerge.