Overarching Structure

- **Session 1** focuses on setting out the groundwork for discussions around AI control research. I think safety cases are an important framework to know about, particularly in the context of control evaluations, and they offer a good non-technical introduction.
- **Session 2** pivots to looking at existing technical work on control evaluations, introducing the different settings and protocols currently in the literature. This makes up some of the most important and actionable work in control.
- **Session 3** highlights research that is being done on control-relevant scheming and sabotage capabilities. Capability (and propensity) evaluations are important for two reasons: they help shape our timelines; and they help us design red-team affordances to improve our control evaluations.
- **Session 4** delves into two specific technical areas of interest in chain-of-thought monitoring and steganography. The former depends on the faithfulness of CoT in frontier models, and the latter examines one way in which these models might evade CoT monitoring.
- **Session 5** (not yet compiled!) is an optional bonus session digging into some alternative avenues within control, including the Games for AI Control paper that I (Louis) worked on. We could also look into the MONA paper. If anyone has suggestions or requests for this bonus session, let me know!