Transcript Analysis Project Proposal

Overview

SaferAI and a partner collaborate on a transcript-evaluation framework to assess adversarial cybersecurity capability by analyzing transcripts from evals and real-world environments. This could mean extracting some structured data from transcripts like token efficiency, tool usage, or refusal circumstances. It could also extract less structured data like action classification (i.e. lateral movement vs execution) or highlight high-risk actions that may increase the likelihood of detection. The goal is to extract novel insights about cyber capabilities and tendencies of agentic AI systems that are hard to obtain based on benchmark scores alone. These could be useful on their own, and perhaps even serve as inputs to our quantitative risk models, once a rigorous way of turning these insights into numerical values is determined.

Rationale

- The SaferAI risk modelling team is looking for new inputs to our risk models given our view that benchmarks are weak (and rapidly saturating) proxies for operational capability in live, defended cyber environments. One input we are considering is using open ended evaluations (like cyber ranges) or real-world red team data.
- In order to make use of these kinds of evaluations in risk models, we need to figure out how to process their outputs - they are generally much less structured than benchmarks. Transcript analysis could be a good option for this, and seems to be a direction some folks in the field are moving towards. It also could easily be repurposed for agentic benchmark runs (like CyBench), or more structured cyber ranges which could extend the utility of already saturated benchmarks.
- One of the important functions of a risk model is to provide a substrate for integrating many types of data, drawing operational conclusions by integrating information from diverse sources into a common quantitative “language”. We have not yet explored the potential of risk models to use any sorts of data beyond benchmarks, and transcripts are a natural first step in that they are relatively plentiful, generally applicable, easily adapted to reflect red team data beyond evals data, and provide higher resolution signal than a binary “pass / fail” metric.
- We are one of less than 5 organizations with a fully functional evals team which can provide cyber-relevant eval transcripts as needed. In addition, we have cybersecurity partners who can provide live-environment transcripts as well. Taking advantage of this data for this kind of work is high-impact.

Deliverables

- **Methodology paper** — describes the transcript-evaluation framework: annotation schema, extraction pipeline, etc. Also outlines how they may fit into our risk models. Targets the same audience as our existing methodology paper, maybe useful for other Safety orgs looking to make sense of transcripts.
- **Applied blog post** — applies the methodology to CyBench transcripts and presents insights. Ideally we are able to outline which offensive skills are operational, and set a baseline for model competencies in operational security. This communication can be targeted at the operational cyber, policy, or evals communities.

Resource Commitment

From SPAR Participant(s)

- Transcript-analysis methodology design and implementation
- Drafting and publication of outputs

From SaferAI

- Cybersecurity expertise — Jack Kengott as cyber security domain expert
- Evaluation expertise — May Dixit as evals domain expert
- Risk-modeling expertise — Jakub Kryś as risk modelling lead
- General mentorship - Advising on draft and publication of outputs (review from SaferAI team + external review).

Timeline

- **Weeks 1-6:** Develop modelling approach for transcript analysis, answering “how is transcript data turned into a number that can be used in a risk model”
- **Week 6:** Proof of concept on CyBench data, augmenting existing OC3 risk model structure.
- **Milestone:** Internal report on methodology and initial CyBench results. Can be shared externally for closed review as well.
- **Week 7:** Testing and development on live data, integrating above approach for more open-ended purposes
- **Week 8-9:** Integration of both benchmark and live data transcript analysis into a novel risk model that aligns with updated inputs. Experimentation to determine transcript analysis’ impact on overall risk estimations. Gathering necessary data for writing.
- **Week 9-11:** Scaffolding and writing blog and methodology paper
- **Week 12:** Handoff and retrospective.

How does this project advance AI Safety?

- **SaferAI's Risk Modelling Theory of Change:** We develop models with the goal of identifying rapidly advancing and/or highly consequential offensive capabilities, and communicate these capabilities to policy makers and the cybersecurity community.
- **Project Goal:** Current risk models don't have enough data to resolve detailed questions about capability with high uncertainty, so we need to find ways to improve our input data. Transcript analysis is a good option, and will improve our understanding of capabilities, and be useful to the AI Safety field at large.
- **Audience(s):** Policy Makers and Operational Security Community, potential extension to AI Safety researchers
    - **What it does for Policymakers:** Direct funding towards research areas which counteract identified developing capabilities. This could result in updated vulnerability disclosure practices, patching requirements for government/critical infra, use of open source projects, etc
    - **What it does for the Operational Security Community:** We provide insight into emerging dynamics and how to adapt defense strategies (would be ideal to partner with security orgs for these kinds of recommendations). The private defense market has incentives to develop defensive capabilities, but likely needs a better signal on where to direct this funding.
    - **AI Safety Researchers:** Loads of evals coming out with a common theme, "how do we make this make sense to decision makers". Transcript analysis provides a rich interface between eval runs and real-world impact.