# Inter-machine Existential Risk Triage

Andrii Shportko

The AI safety field is both talent and resources constrained, so we need an effective way to triage the risks to defend against. Multi-agent risks are among top concerns among AI experts (Saeri et al., 2026), but to this day, there has been no systematic study related to which inter-AI risks are to be prioritized and what risks are the most realistic. I would like to develop a dynamic Bayesian protocol that would help us to update the top risks based on evidence. I would also like to create a hazard scale related to the inter-AI security failures, similar to the Torino scale.

*Dynamic* is a crucial component here. Risk priorities should be updated based on the evidence available. For example, the reason why we started caring more about AI safety is because such technology became available. Typically, we update our risks post-hoc, but I would like to develop a pro-active protocol that would (de-)prioritize specific AI control directions based on evidence.

In late July, an OpenAI model left notes to its future instances on how to bypass containment. How does this event update our research priorities? The Cooperative AI Foundation's *Multi-Agent Risks from Advanced AI* (Hammond et al., 2025) provides a taxonomy of these risks by identifying three key failure modes (miscoordination, conflict, and collusion) based on agents' incentives, as well as seven key risk factors (information asymmetries, network effects, selection pressures, destabilising dynamics, commitment problems, emergent agency, and multi-agent security). However, it remains unclear what issues should be addressed first. For example, the machines may learn to communicate with each other in a way that bypasses monitors – they may learn to collude steganographically, in plain sight (Mathew et al., 2024). At the same time, we have little-to-no evidence that machines can naturally acquire a propensity for such sophisticated communication mechanisms, let alone reason in cipher (Guo et al., 2025).

Additionally, what does it take to trust an untrusted monitor (Gardner-Challis et al., 2026)? And what kind of protocols are collusion-resistant to the realistic threat models (Nguyen et al., 2026)? While risk assessment remains my top interest, I am also very interested in predicting when a certain capability becomes available parametrized over full automation of AI R&D timelines.

By the end of the program we will deliver a dynamic Bayesian protocol for prioritizing inter-AI risks that ingests concrete evidence (capability-eval scores, control-eval safety numbers, incident reports) and updates a ranked risk list as new evidence arrives. Alongside it we will produce a Torino-style hazard scale for inter-AI security failures and an initial evidence-backed ranking of the risks from the Hammond et al. taxonomy to give the field a first defensible answer to which inter-machine threats warrant attention now.

I am looking for enthusiasts in long-termism, Bayesian statistics, AI governance, AI control, or mech interp.