

Finding Failure Cases in Multi-Agent AI Systems

Co-author research opportunity
with Cooperative AI Foundation

🏆 **\$2,000 in prizes**

With the Cooperative AI Foundation
alignmentjam.com/jam/multiagent

Schedule for the weekend

See [the website](#) for the schedule in your local time

- **Friday:** Keynote [19:00 UTC] and virtual mixer [20:00 UTC]
- **Saturday:** Project discussions [14:00 UTC] and technical talks w/ Joel Leibo and Christian Schroeder de Witt [17:00 UTC]
- **Sunday:** Project discussions [14:00 UTC], final presentation [19:00 UTC], and submission deadline [↗ 2:00 UTC]
- **Afterwards:** Anonymous submissions are evaluated and contacted for collaboration



Friday 19:00 (your local time): Keynote talk with Lewis Hammond and Esben Kran

Saturday 14:00 (your local time): Project discussions with Lewis Hammond

Saturday 19:00 (your local time): Two technical talks from Joel Leibo, Christian Schroeder de Witt and Sasha Vezhnevets with a 15 minute break in-between

Sunday 14:00 (your local time): Project discussions with Lewis Hammond

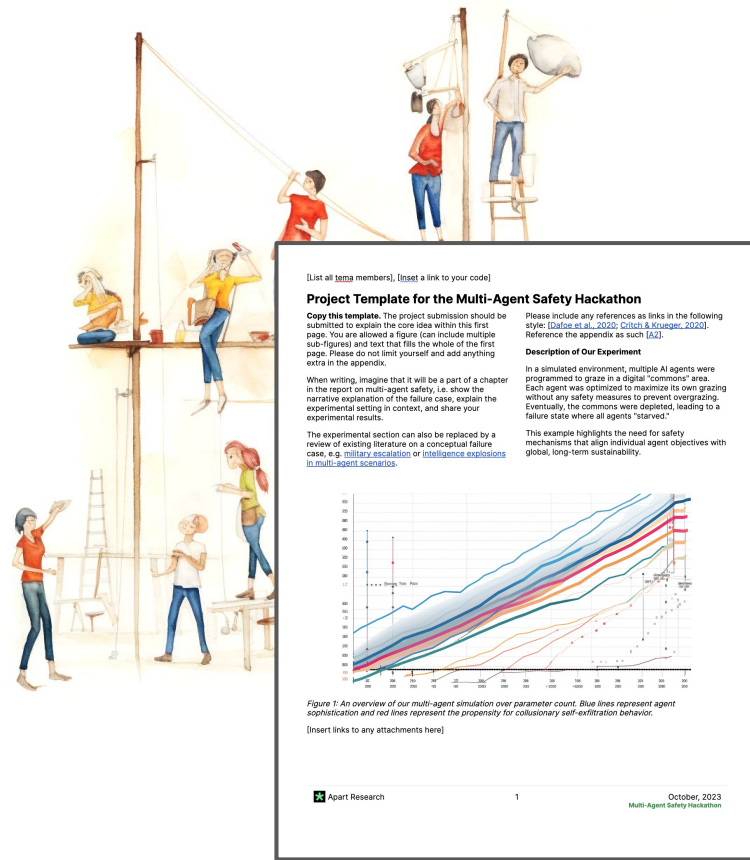
Sunday 21:00 (your local time): Ending session with Lewis Hammond

Monday 4:00 (your local time): Submission deadline

Your goal for this weekend

Search for the most concerning failures specific to systems of multiple AIs.

- Join in groups of 1-5 to experimentally showcase a concerning failure
- PDF submission by Sunday
 - Instructions on the submission page
 - Use [the template](#): One page of text and a visualization along with appendix
 - Summary, image, link to your code...



Hacking across the Globe

- [Code of Conduct](#)
- Locations
 - EffiSciences at ENS Ulm
 - SAIA at Stanford University
 - Center on Long-Term Risk in London
 - LAIA at EPFL in Switzerland
 - Apart in Copenhagen
- Country-specific channels on Discord
- Write any questions in the [#? | questions](#) channel! Lewis and others will be monitoring this channel



Multi-Agent Safety Hackathon at Center on Long-Term Risk (London)

CLR offices are hosting a hackathon site from 10am-10pm UK time on Saturday and Sunday.

[VISIT EVENT PAGE](#)



Center on Long-Term Risk

Multi-Agent Safety Hackathon with EffiSciences (Paris)

The French AIS hub host the hackathon in Paris, 45 rue d'Ulm (moving rooms during the week-end)

[VISIT EVENT PAGE](#)



ENS Ulm, Paris

Stanford AI Alignment Multi-Agent Hackathon

Stanford AI Alignment is hosting a local site in the Gates CS building all day Saturday and Sunday! Watch the SAIA email list/Slack if you're already on it, or contact gmukobi@stanford.edu to join!

[VISIT EVENT PAGE](#)



Stanford University

Multi-Agent Safety Hackathon Hub at EPFL

On campus - Exact location TBD but probably Rolex or DLL

[VISIT EVENT PAGE](#)



EPFL

Copenhagen Multi-Agent Safety Hackathon

Join us in Copenhagen to contribute to exciting research within multi-agent safety research. We will provide food and drinks for your research journey!

[VISIT EVENT PAGE](#)



Copenhagen, the Apart Offices

Alignment Jam

- ~Monthly hackathons on research and governance related to AI safety
- We appreciate [your feedback](#)
- Upcoming research challenges
 - [Delft Entrepreneurship AI Safety Hackathon](#) w/ EntrepreneurFirst (November)
 - AI governance research sprint in December / January
- See projects [here](#) from previous finished challenges [here](#)
- Non-profit project



Exploring the Robustness of Model-Graded Evaluations of Language Models

Simon Lermen, Ondřej Kvapil | Safety Benchmarks hackathon

Relating induction heads in Transformers to temporal context model in human free recall

Ji-An Li | Interpretability hackathon

Data Taxation

Joshua Sammet, Per Ivar Friborg, William Wale | AI Governance Hackathon hackathon | March 26, 2023

Automated Sandwiching: Efficient Self-Evaluations of Conversation-Based Scalable Oversight Techniques

Sophia Pung, Gabriel Mukobi | ScaleOversight hackathon | February 12, 2023

We Discovered An Neuron

Joseph Miller, Clement Neo | Mechanistic Interpretability Hackathon hackathon | January 22, 2023

Discovering Latent Knowledge in Language Models Without Supervision - extensions and testing

Agatha Duzan, Matthieu David, Jonathan Claybrough | AI Testing hackathon | December 18, 2022

Alignment Jam

We organize monthly hackathons to engage people across the world (all 7 continents!) in technical AI safety. Browse our upcoming events here.

[See all events](#)

[Subscribe to the calendar](#)

[Join our community](#)

[See the top research entries](#)

Finding Failure Cases in Multi-Agent AI Systems

Co-author research opportunity with Cooperative AI Foundation

 **\$2,000 in prizes**
With cooperativeai.org
alignmentjam.com/jam/multiagent

Currently happening

Global

Multi-Agent Safety Hackathon

As AI systems proliferate and become increasingly agent-like, they will interact with each other and with humans in new ways. These new multi-agent systems will create entirely new risk surfaces.

WHERE

EPFL
Center on Long-Term Risk
ENS Ulm, Paris
...and more

WHEN

September 29, 2023
October 1, 2023

PRIZES

 \$2,000

[Register Now](#)

Safeguarding Human Agency Against A(GI)

With agencyfoundations.ai
alignmentjam.com/jam/agency

Global

Ended

September 8, 2023

Agency Foundations Challenge

[See the results](#)

Global

Ended

August 25, 2023

Distillation Write-a-thon 2.0

[See the results](#)

With collaborators and previous keynote speakers from...



Submitting your project

- All information should be in [the template](#)
- One-pager for judging
- Appendix for any extra content
- Submit under the "Submit" tab

Several projects have ended up as publications! See them [here](#)



Submit your project for the hackathon!

Title of your project

Upload your PDF report

Upload your slideshow

Upload a project image

Link to a recording of your presentation (optional)

Link to your code (optional)

Link to other material (optional)

Short summary of your project

Team information

Team name

Full names of your team members

All team email addresses (separate by comma, private)

Discord username

Where are you participating from?

Virtually

Are you interested in publishing this project?

Yes

Are you interested in publishing this project?

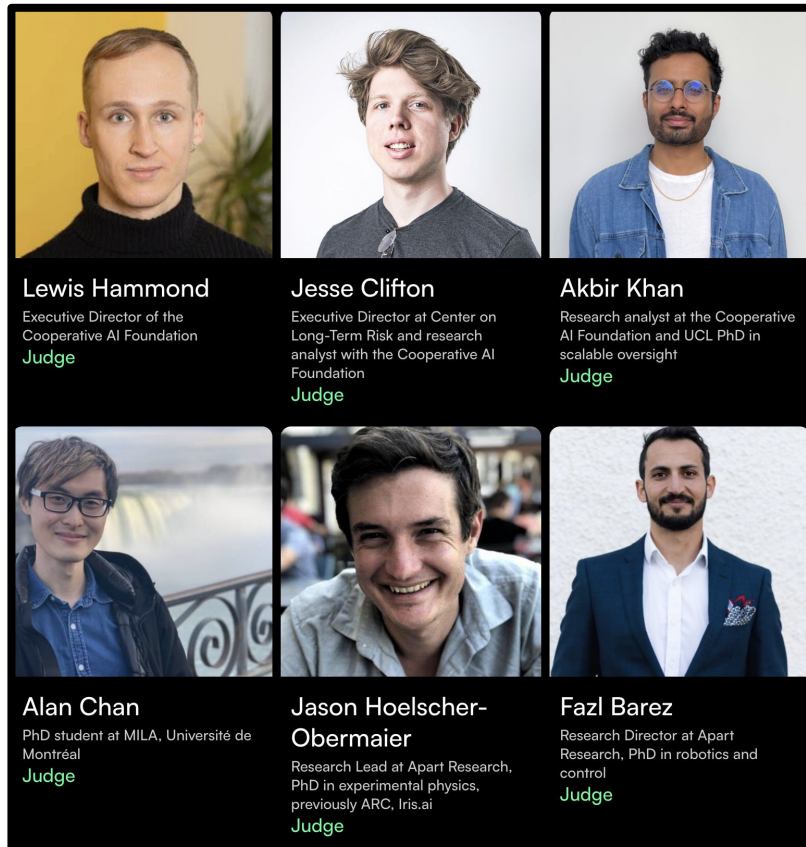
☒ By submitting, I agree to Alignment Jam to share my project and its results.

☒ If I do not personally wish to continue the project, I allow others to continue work on the project.

Submit your project

Prizes and judging

- Winners receive \$1,000!
 - 2nd place: \$500
 - 3rd place: \$250
 - 4th place: \$150
 - Random prize: \$100
- Judging criteria
 - Compelling Narrative & Easy to Understand
 - Focus on Multi-Agent & Frontier Systems
 - Easily Replicable
- Co-authoring the report on Multi-Agent Risk
- Thank you to our judges!



Hackathon Spirit

- Remember to have fun!
- Chance to learn and develop your ideas with concrete experiments
- You'll be surprised (see previous project [here](#))
- Coordinate with your team: All-nighters with energy drinks, constant online co-working, chill learning?



Feedback & images

- Respond to our [feedback survey](#) and get a free book about AI safety!
- Submit your images of you hacking away under the "images" tab



Finding Failure Cases in Multi-Agent AI Systems

Co-author research opportunity
with Cooperative AI Foundation



\$2,000 in prizes

With the Cooperative AI Foundation
alignmentjam.com/jam/multiagent