JAC2023 Agenda

Saturday March 11

9:30 - 9:45 Opening Talk - Connor Leahy

- Background / contextualizing information and an overview of the conference agenda. The first half of Saturday is focused on creating common knowledge about AI safety and alignment. The second half of Saturday is focused on deepening technical discussions.

9:45 - 10:30 Fireside Chat - Jaan Tallinn and Ryota Kanai

- Open conversation on AI safety in the West vs. Japan.

10:30 - 11:00 Alignment Introduction - Beren Millidge

- A presentation on AI alignment vs AI safety, and discussion of major research directions.

11:00 - 11:30 Overview of Japanese AI Safety Research

- Presentations of some research directions from Japanese participants:

11:30 - 12:30 Small Group Breakouts

- Small, 5-person groups to further discuss AI alignment and answer any unaddressed questions. The aim is to conclude this section with enough common knowledge to have more substantive and technical conversations about promising research directions.

11:30 - 12:30 Online Group Breakout

- Small, online groups to further discuss AI alignment and answer any unaddressed questions. The aim is to conclude this section with enough common knowledge to have more substantive and technical conversations about promising research directions.

12:30 - 2:00 Lunch and Break

- Lunchboxes will be provided, vegetarian options included.

2:00 - 2:15 Regroup and Recap

- A quick recap of the breakout conversations, and an overview of the afternoon session.

2:15 - 3:00 Q&A with Connor Leahy and Eliezer Yudkowsky

- A short talk on the various risks of AI systems, followed by Q&A and discussion.

3:00 - 4:30 Thematic Breakout Groups

- Participants will split into four groups to discuss the AI alignment theme of their choice. Groups will be loosely moderated, and larger groups will be split into smaller subgroups. There will be suggested readings for each theme. Themes include:
    - Interpretability: How interpretable are current models? What techniques have shown the most success so far? What are the limitations of interpretability?
    - Conceptual Alignment and Philosophy: What is alignment in theory? What kind of assumptions do/should we make about AI systems? How much can we deduce about alignment without experimentation?
    - AI Applications and Alignment: What are the current limitations of alignment techniques? How strong are capabilities? Are current AI alignment techniques sufficient compared to these capabilities?
    - Policy, Governance and Fieldbuilding: What policy and governance mechanisms may help with AI alignment? What other coordination efforts may help? What does the AI safety scene in Japan look like, and how can it be strengthened?

4:30 - 5:00 Regroup and Recap

- Groups will share highlights from their conversations. Sunday's agenda will be discussed. Participants will be invited to add to the agenda any additional lightning talks or discussions they would like to have.

5:00 - 5:30 Break

- Short break to refresh before traveling to the dinner venue.

5:30 - 8:00+ Travel and Dinner

- Gather at the venue entrance for travel.

Sunday March 12

9:30 - 9:45 Opening Talk

- Welcome back and an overview of Sunday's agenda. The first half of Sunday is focused on participant-driven conversations and networking. The second half of Sunday is focused on creating concrete alignment research directions and plans for the future of AI alignment in Japan.

9:45 - 12:30 Lightning Talks / Open Discussion

- The conference has two rooms. Room 302 will support mingling and open discussion.
- Room 301 will support 15 - 30 minute lightning talks and discussion from participants. The talks will be sourced based on participant interest.

12:30 - 2:00 Lunch and Break

- Lunchboxes will be provided, vegetarian options included.

2:00 - 4:30 Breakout Discussions - Getting Concrete

- Participants will split into groups to discuss various actionable next steps. Groups will be loosely moderated, and larger groups will be split into smaller subgroups. Details on groups to be determined based on Satuday's conversation, but will roughly cover:
    - Open brainstorming of direct research proposals
    - Discussion of AI safety in Japan and how it could better address alignment
    - Creation of communication channels to share information or collaborate
    - Proposals to make AI alignment more of a priority in Japan / globally
    - Career paths and funding for alignment research
    - Future of AI safety in Japan

4:30 - 5:45 Regroup: Takeaways and Next Steps

- Participants will regroup to have a large conversation and final brainstorm on specific implementable things people at the workshop can do / follow up on. E.g., write a memo or letter about the conference, formalize a scientific steering committee, etc.

5:45 - 6:00 Closing Remarks - Ryota Kanai

- Closing words and formal conclusion of the conference.

6:30 + Closing Drinks

- Post-conference drinks.