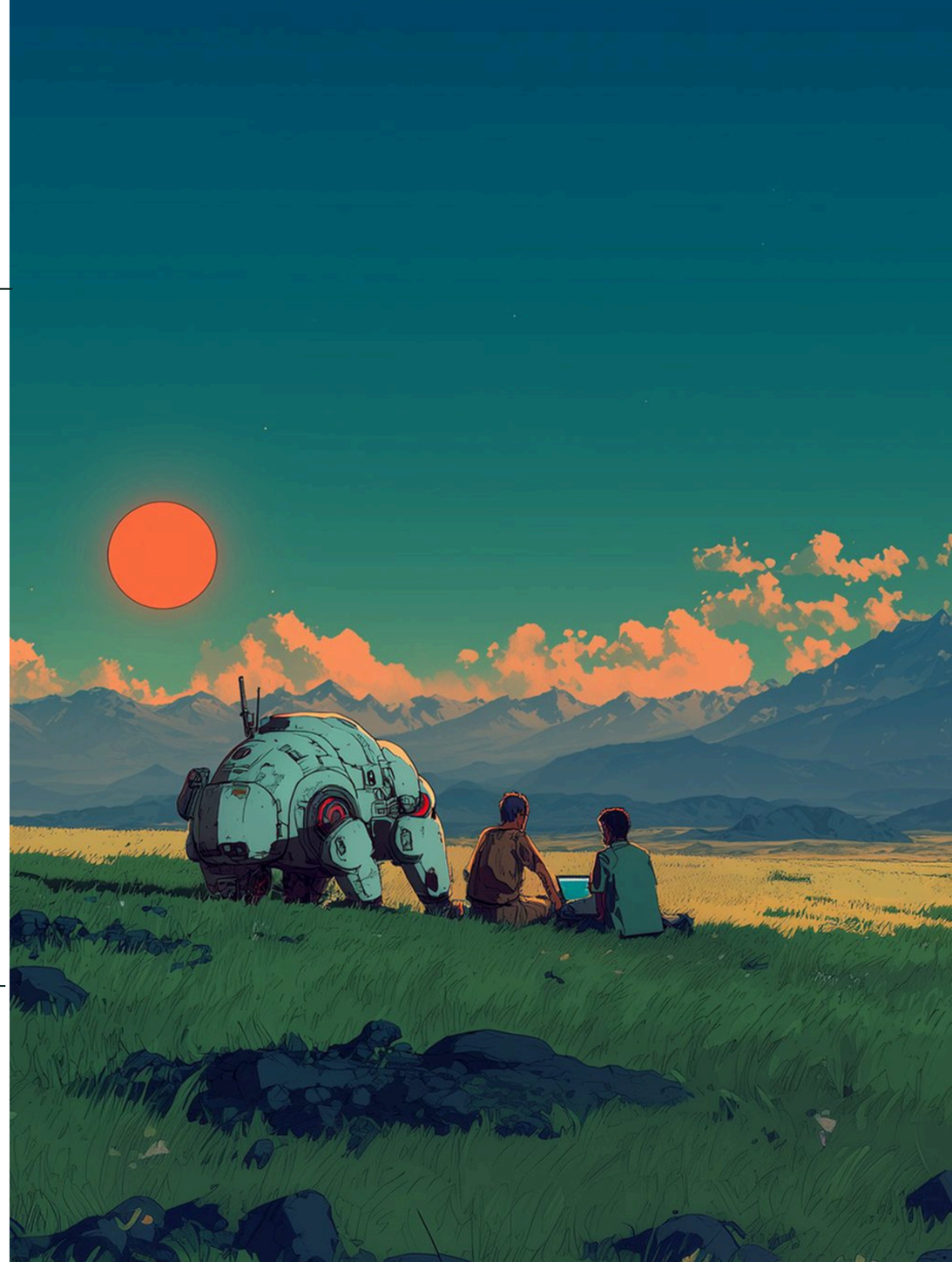
Cooperative AI
Research Fellowship

INAUGURAL COHORT · FEBRUARY - MAY 2026

Seeding a new cooperative AI ecosystem



COOPERATIVE AI RESEARCH FELLOWSHIP

An AI Safety South Africa Initiative

Cohort & Mentors

The program brought 10 researchers from 8 countries to South Africa to work on Cooperative AI for 3 months in early 2026. Fellows were selected from 1100+ applicants across a five-stage application process and were paired with mentors from institutions such as Google DeepMind and MIT. The program started with a week-long retreat, which was foundational in developing project ideas and cohort bonding.

FELLOW	MENTOR(S)
Bhavyesh Sajja	Max Kleinman-Weiner (University of Washington) & Tan Xhi Xuan (National University of Signapore)
Joseph Low & Oscar Duys	Michiel Bakker (MIT, Google DeepMind) & Lewis Hammond (Cooperative AI Foundation, University of Oxford)
Akash Kundu	Vincent Conitzer & Emanuel Tewolde (Foundations of Cooperative AI Lab, Carnegie Mellon University)
Omer Kamal Ali Ebead	Joel Z. Leibo (Google DeepMind)
Qi Guo	Lewis Hammond (Coooperative AI Foundation, University of Oxford)
Pramod Kaushik	Sahar Abdelnabi (ELLIS Institute Tübingen & MPI-IS, COMPASS Lab)
Yves Bicker	Zhijing Jin & David Guzman Piedrahita (University of Toronto, Jinesis Lab)
Van Quynh Thi Truong	Zhijing Jin & David Guzman Piedrahita (University of Toronto, Jinesis Lab)
Mariana Meireles	Zhijing Jin & David Guzman Piedrahita (University of Toronto, Jinesis Lab)

Key Outcomes

FELLOWS

10

Paired with mentors at MIT, Oxford, DeepMind and more.

FELLOW NPS

9.1/10

Participant recommendation score. No fellow rated the program below 7.

STAKEHOLDER NPS

9.1/10

Mentor and partner recommendation score. No respondent rated below 8.

MENTORSHIP RATING

8.5/10

The highest-rated program component rated by fellows.

NEURIPS SUBMISSIONS

4+

Additional submissions at ICML.

CONTINUED FUNDING

6/10

Fellows with secured continuation funding post-fellowship.

PRODUCT OUTCOMES

1

Habermolt launched, paper to arXiv, acquired by Change.org mid-fellowship.

CAREER IMPACT

10/10

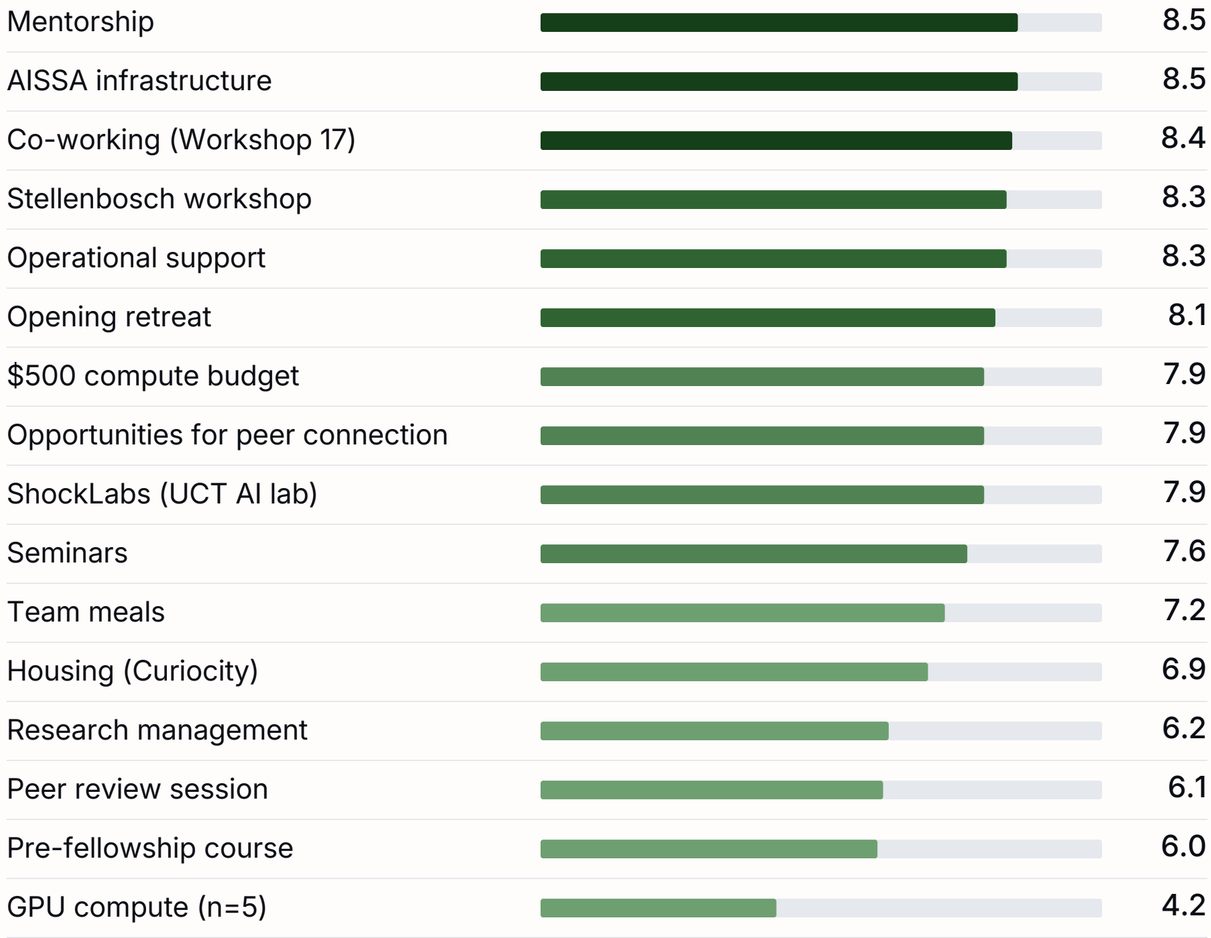
All fellows reported the program confirmed, redirected or accelerated their trajectory.

Research Outputs

FELLOW	PROJECT	HEADLINE OUTCOME
Bhavyesh Sajja	Evaluating Rationality in Natural-Language Contracting	NEURIPS SUBMITTED
Joseph Low & Oscar Duys	Delegating Deliberation to AI Agents (Habermolt)	LAUNCHED · ACQUIRED · ARXIV
Akash Kundu	Similarity as a Signal	NEURIPS SUBMITTED · EXTENSION FUNDED
Omer Kamal Ali Ebead	Contextual Integrity in Multi-Agent LLM Systems	EXTENSION FUNDED
Qi Guo	Inter-Agent Influence	NEURIPS SUBMITTED · HIRED BY CAIF
Van Quynh Thi Truong	Social Strategies for Cooperation in Multi-Agent Societies	COLM / NEURIPS TARGET
Yves Bicker	MoralGym: Training Cooperative LLM Agents	RAPID GRANT · ICLR TARGET
Pramod Kaushik	Strategic Indirect Speech in LLMs	EXPERIMENTS COMPLETE
Mariana Meireles	Toward Collective Intelligence	EXPERIMENTS COMPLETE

Fellow Ratings

Program components, scored out of 10.



FELLOW NPS

9.1/10

Peer bonds were the most-cited value

They're all super sharp and generous and I've learnt an enormous amount from them on everything from technical questions to how to think about my career. Three months is a short time but I genuinely believe these are friendships for life.

FELLOW

Having people to lean on, did have a huge impact on me, especially during the middle when there are no quick results to look at. I think they helped me push further. Moreover, I was amazed by how diligent everyone was at work. They somehow passed on their discipline which made me push slightly harder than I would have, because everyone around me, was.

FELLOW

Mentor & Stakeholder Feedback

10 respondents from MIT, NUS, DeepMind, CAIF, PIBBSS, CMU, Oxford, UCT, UW and Lambda. Aggregate NPS 9.1.

STAKEHOLDER NPS

9.1/10

MENTOR RATINGS

- **Weekly meetings with fellows** — 9.3 / 10
- **Mentor matching experience** — 9.2 / 10
- **Program prepared mentee for career in coop AI** — 9 / 10
- **Operational support & comms** — 8.7 / 10
- **Pre-fellowship project development** — 8 / 10
- **Value from Research Manager (mentor-rated)** — 7.3/ 10

> *CAIRF is launching the next generation of scholars in Cooperative AI from around the world. Highly recommend this well run program.*
>
> MAX KLEIMAN-WEINER · UNIVERSITY OF WASHINGTON

> *I loved working with Oscar and Joseph, the two CAIRF fellows I mentored. Having everyone together in person in Cape Town (except for me, sadly) made a huge difference to how quickly things moved and what the fellows were able to achieve. I'm also very grateful for all the support from Claude Formanek and the rest of the team!*
>
> MICHIEL BAKKER · MIT

> *Well-structured, made tons of research progress, funding is providing for experiments, and it seems like my student benefited a lot from the community and support too :)*
>
> TAN ZHI XUAN · NUS

Outcomes vs. Aims

The aims we set when we launched, and where we landed by the end of the fellowship in May 2026.

6 MONTHS	12 high-quality research projects in cooperative AI	10 projects produced, two fellows removed mid-program. Quality strong (NeurIPS submissions, acquisition, ongoing research); target nearly met.	PARTIAL
6 MONTHS	3–5 fellows identified for permanent positions	6 fellows with continuing funded research — Qi at CAIF, Joseph & Oscar at Change.org, Omer & Akash on CAIRF extension, Yves on rapid grant.	SUCCESS
6 MONTHS	2–3 South African academics shifting research focus toward AI safety	Had many professors express interest in getting involved in our work, either by providing students for projects, attending events, or working on research together.	STARTED
1–2 YEARS	2–3 permanent cooperative AI research positions at UCT AI Safety Hub	Submitted a joint proposal with UCT for a multi-agent safety lab with Claude Formanek with the PI, and Prof. Shock as CO-PI	STARTED
1–2 YEARS	Established talent pipeline from Africa to global AI safety organisations	Qi Guo's hire by CAIF is the first concrete instance. Mentor network now reaches Jin lab, FOCAL, COMPASS, CoSI lab, Computational Minds & Machines, MIT and DeepMind.	STARTED

Going Forward

Six recommendations the team is committing to for the next iteration, synthesised from fellow, mentor and team feedback.

Co-working space

Customised office

Despite receiving a high rating out of 10, many written responses that highlighted improvements for the office setup. Future planning will prioritise securing a private, dedicated facility that separates quiet work zones from social or collaborative areas and caters to a wider variety of work styles and social needs.

Community Manager

Hire community manager

A community manager will serve as the dedicated bridge between the fellows and the AISSA team, so our presence is more visible and approachable. This role would also support preventing and resolving community health issues.

Research management

Clarify RM role & reporting

Provide a clearer brief to RMs and fellows about what is expected of them. Also, we would clarify reporting and feedback systems from the start of the fellowship. Furthermore, we think we could improve outcomes by scoping this role better in terms of hourly demond.

Compute

Expand the compute pool.

Combine GPU + API budgets into a single flexible per-fellow allocation. Raise default API budget for compute-heavy projects. Budget $1500-3500/fellow.

Timeline

Secure funding & housing earlier

Funding transferred 6 months prior, housing at least 4 months before fellowship start. This would prevent significant uncertainty that blocked further decisions early on in the fellowship.

Due Diligence

Screening & code of conduct

Contact candidate’s references and screen them through CEA. Require signing of code of conduct prior to start of fellowship and have a talk about norms and appropriate conduct on day 1.

Full recommendation set in internal team retrospective document.

Cooperative AI
Research Fellowship

END

Thank you

CONTACT

info@cai-research-fellowship.com

MORE ON THE PROGRAM

www.cai-research-fellowship.com

Cooperative AI | AI Safety South Africa | PRINCIPLES of INTELLIGENCE | AI Initiative | Lambda