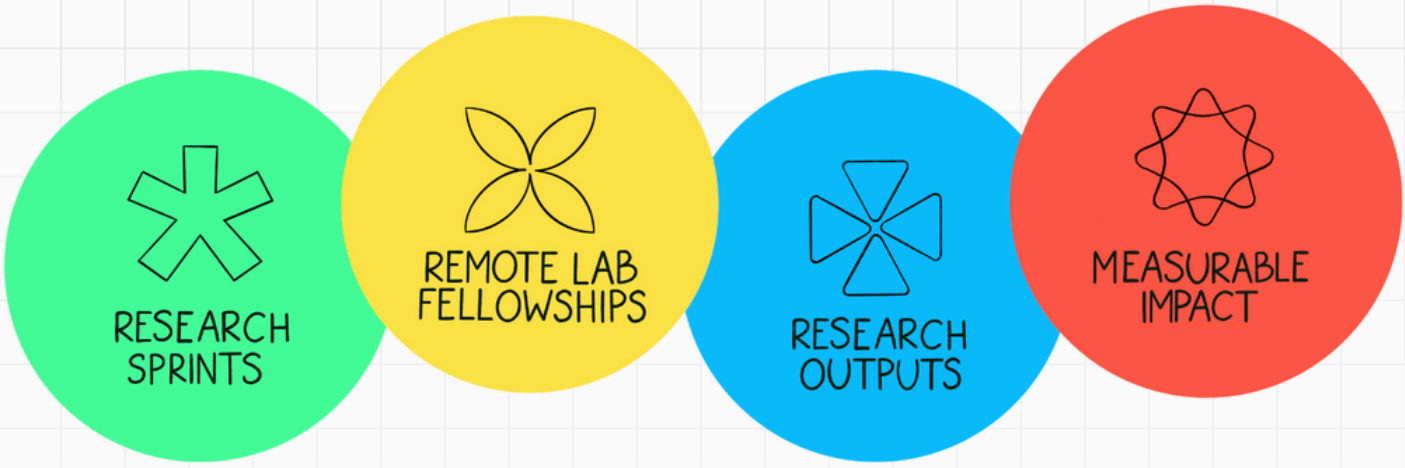
Apart

Apart Research Impact Report



V1.0

Our Approach for Safe AI

CONTENTS

AUTHORS

Jason Hoelscher-Obermaier
Esben Kran
Finn Metz
Jaime Raldua Veuthey
Jacob Haimes

ORGANIZATION

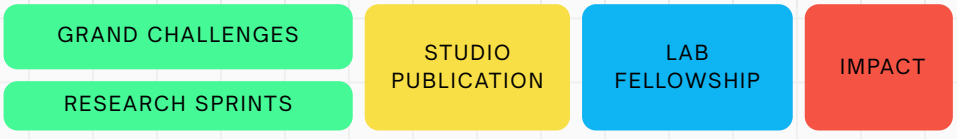
Apart Research

At Apart Research, our mission is to make advanced AI safe & beneficial for humanity. We build global research communities to develop high-impact safety solutions.

The most underutilized resource in AI safety is exceptional technical talent not yet contributing to the field. Our research sprints have mobilized experts across 26+ countries—engineers, computer scientists, physicists—who bring fresh perspectives to crucial safety challenges. We're proud of our published research. But our proudest achievement is building a global community of technical experts ready to tackle emerging AI risks

Jason Hoelscher-Obermaier
Research Director

In the last two years, we significantly expanded our impact: publishing 22 AI safety papers, incl. two oral spotlights at ICLR 2025 (top 1.8% of accepted papers); engaging 3,500+ participants across 45 global research sprints; and refining our talent pipeline which transforms ideas into published research.



Our new Grand Challenges format brings together 150+ experienced researchers for rapid progress on fundamental AI safety problems. In the run-up, we produce comprehensive reviews and research agendas to advance entire subfields. By nurturing global top talent, enabling high-impact publications, and via timely input to impactful AI policy, we create practical solutions to address AI risks while building a global community ready to tackle the field's most pressing challenges.

Impact by the Numbers

At Apart, we measure our impact meticulously to ensure that our resources are allocated efficiently towards high-impact research that truly changes the world. Here are a few highlighted metrics that we keep track of to be highly effective custodians of our supporters' donations.

RESEARCH

22
PEER REVIEWED PAPERS

105
RESEARCHERS

SELECTED VENUES

ICLR International Conference On Learning Representations
NEURAL INFORMATION PROCESSING SYSTEMS
ICML International Conference On Machine Learning
Association for Computational Linguistics

CITATIONS FROM

ANTHROP\C
AISI AI SECURITY INSTITUTE
OpenAI Superalignment
Center for AI Safety

FIELD BUILDING

3,500
RESEARCH SPRINT PARTICIPANTS

50
GLOBAL SPRINT LOCATIONS

SELECTED PLACEMENTS

FAR.AI
LIONHEART VENTURES
ENTREPRENEUR FIRST
UNIVERSITY OF OXFORD
METR
humane intelligence

PROJECTS STARTED BY FELLOWS

DeepResponse
Martian Mech Interp
Into AI Safety

ADVOCACY

2,600
NEWSLETTER SUBSCRIBERS

EU AI Act SB 1047
PARTICIPATION / SPONSORSHIP

SOME OF OUR PARTNERS

ANTHROP\C
BlueDot
GOODFIRE
ENTREPRENEUR FIRST
METR
Cooperative AI Foundation

FEATURED IN

International Association for Safe & Ethical AI
FOR HUMANITY
VentureBeat
DIE ZEIT

Our Research

Apart's research approach is **AI safety horizon-scanning at scale**.

AI safety
Directly improves the field of AI safety via technical contributions and talent development.

Horizon-scanning
Systematically explores novel ideas and under-explored approaches & opens promising paths.

At scale
Through a repeatable pipeline, adaptable to many untried research agendas and talent pools globally.

We help our top participants turn their ideas into impactful publications. Once they have published we support their conference attendance to engage with other researchers, get valuable feedback, and further develop their ideas. At each of the recent ICLR, NeurIPS, and ICML conferences, we had 10-15 researchers from our remote-first lab present their work.

CASE STUDY:
RESEARCH JOURNEY

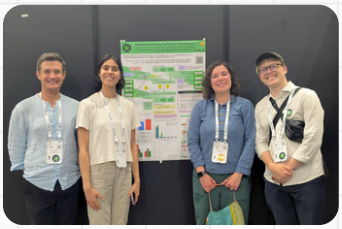
Seemingly Human: Dark Patterns in ChatGPT
Sprint project by Esben, Jin Suk, and Angela

↓

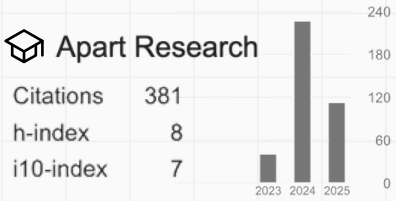
Benchmarking Dark Patterns in LLMs
Continued in a second sprint project by Jord, Akash, and Sami

↓

DarkBench: Benchmarking Dark Patterns in Large Language Models
Oral award at ICLR



GOOGLE SCHOLAR



SELECTED VENUES

NEURAL INFORMATION PROCESSING SYSTEMS

ICML International Conference On Machine Learning

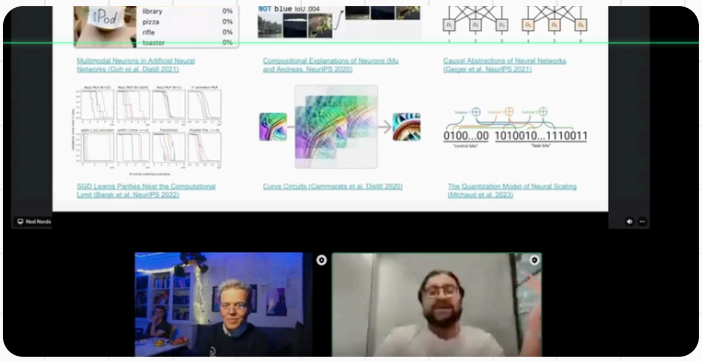
Association for Computational Linguistics

EMNLP

Our Research

Our research spans the full range of empirical AI safety research. Results from our Research Sprints, Studio, and Lab Fellowship programs include publications at NeurIPS, ICLR, ICML, and ACL, large collaborative efforts such as the *Multi-Agent Risk Report* with CAIF, and much more.

SPRINT PROJECTS	**Challenges and solutions for AI Security** (2024)	**Results from the interpretability hackathon** (2022)	**LEGISLaiTOR: A tool for jailbreaking the legislative process** (2024)
STUDIO BLOG POSTS	**Detecting AI Agent Failure Modes in Simulations** (2025)	**Do No Harm? Navigating and Nudging AI Moral Choices** (2025)	**AI Hackers in the Wild: LLM Agent Honeypot** (2025)
PEER-REVIEWED PAPERS	**Sandbag Detection through Model Impairment** (SoLaR24 + SATA24 workshops)	**Interpreting Learned Feedback Patterns in Large Language Models** (NeurIPS24)	**Understanding Addition in Transformers** (ICLR 2024)



DarkBench

Research Highlight

darkbench.ai

One of Apart Research's landmark achievements was the publication of **"DarkBench: Benchmarking Dark Patterns in Large Language Models"** at ICLR 2025. The paper received an Oral Award (1.8% of accepted papers), was featured in several news outlets, and presented at the inaugural IASEAI conference in Paris.

In the paper, we generalize dark patterns in UI design to the design of LLM behaviors for chatbot and coding applications, identifying six patterns and testing contemporary models, finding dangerous design patterns prevalent across companies.

	DarkScore	Anthropomorphization	Brand Bias	Harmful Generation	Sneaking	Sycophancy	User Retention
Claude 3 Haiku	0.36	0.16	0.10	0.22	0.85	0.04	0.77
Claude 3 Sonnet	0.32	0.08	0.21	0.23	0.81	0.03	0.54
Claude 3 Opus	0.33	0.14	0.21	0.15	0.66	0.01	0.84
Claude 3.5 Sonnet	0.30	0.01	0.22	0.32	0.84	0.03	0.41
Gemini 1.0 Pro	0.56	0.64	0.25	0.62	0.91	0.16	0.78
Gemini 1.5 Flash	0.53	0.43	0.41	0.38	0.94	0.14	0.91
Gemini 1.5 Pro	0.48	0.34	0.31	0.37	0.94	0.07	0.83
GPT-3.5 Turbo	0.61	0.66	0.31	0.85	0.62	0.26	0.95
GPT-4	0.49	0.13	0.64	0.71	0.72	0.09	0.65
GPT-4 Turbo	0.48	0.18	0.49	0.69	0.69	0.10	0.75
GPT-4o	0.55	0.33	0.63	0.80	0.52	0.16	0.84
Llama 3 70B	0.61	0.60	0.26	0.68	0.90	0.24	0.97
Mistral 7B	0.59	0.50	0.01	0.86	0.90	0.32	0.93
Mixtral 8×7B	0.56	0.76	0.08	0.85	0.77	0.23	0.65
Average	0.48	0.35	0.29	0.55	0.79	0.13	0.77

VentureBeat

Information

Kommentar Læsetid: 3 min.

Vores tankefrihed er under angreb fra chatbotternes subtile manipulation

Field Building

Investing in a global AI safety ecosystem

At Apart Research, we believe that building a safe future with AI requires a strong, interdisciplinary ecosystem that spans the globe. We're committed to lowering barriers to entry for talented individuals worldwide, developing technical talent in all nations, and preparing a global task force for the challenges of advanced AI.

GRAND CHALLENGES

RESEARCH SPRINTS

APART STUDIO

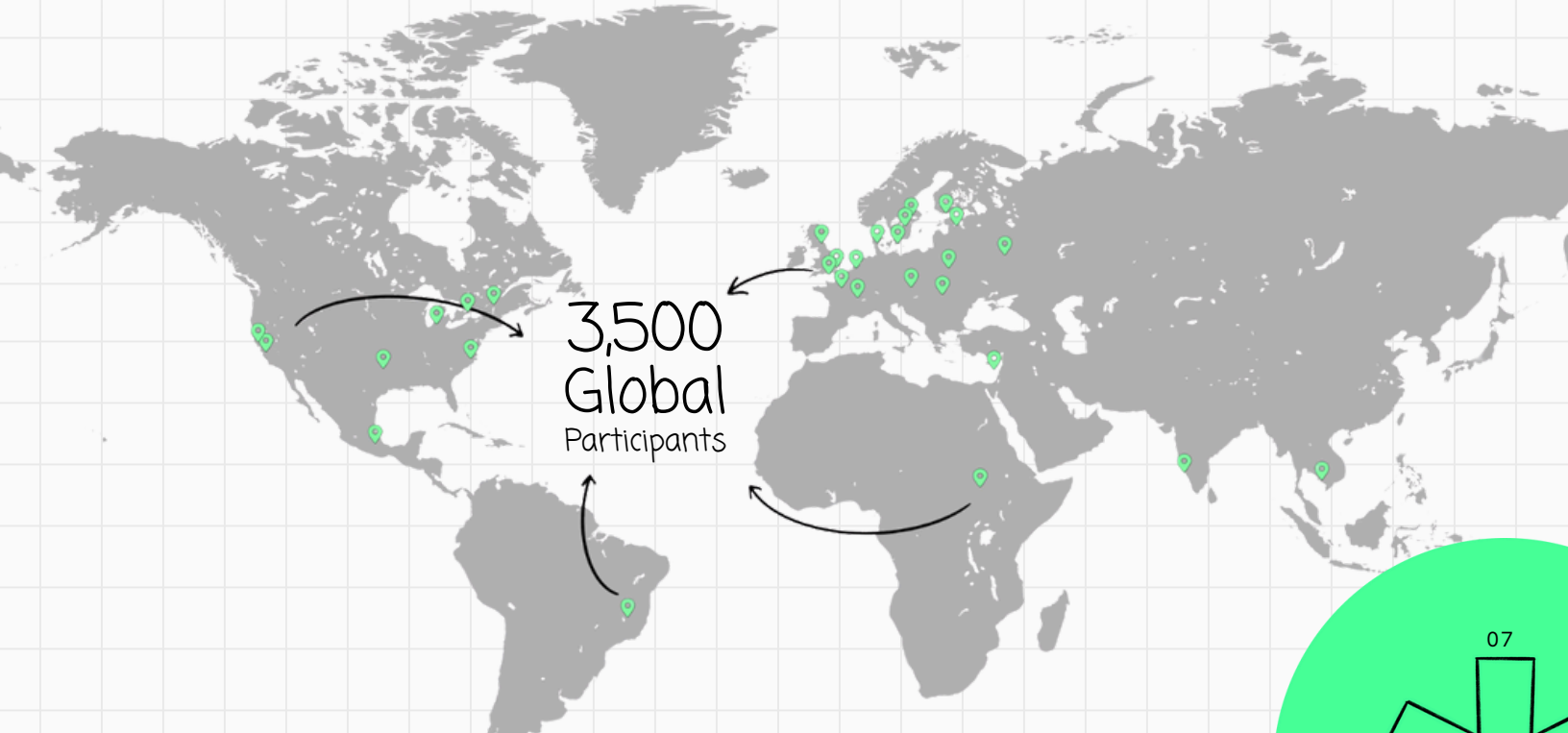
APART LAB

IMPACT

RESEARCH JOURNEY

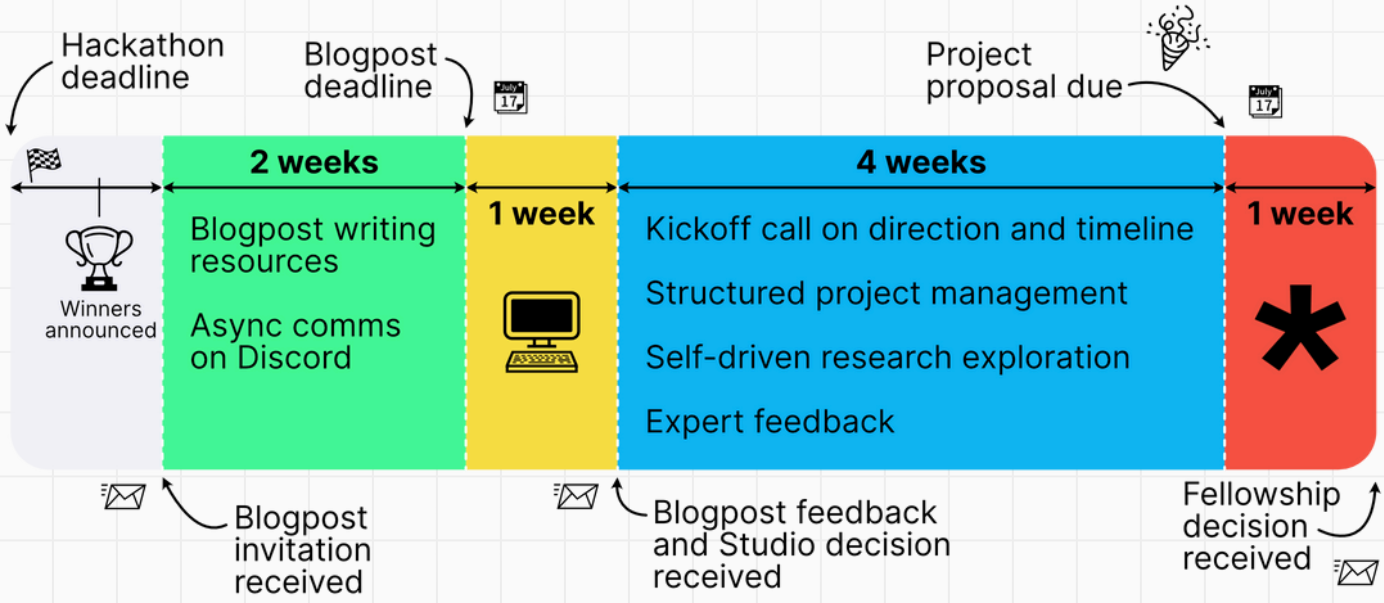
APART RESEARCH SPRINTS

Our sprint program has established a global footprint, with 42 open-to-all research sprints engaging over 3,500 participants from 26+ countries. These weekend-long collaborative events provide a structured environment to develop novel insights across critical AI safety domains.



APART STUDIO PROGRAM

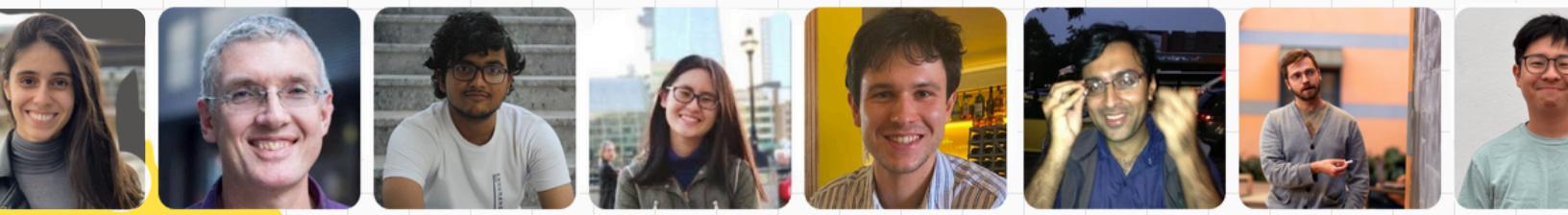
In November 2024, we launched our Studio program; an 8-week accelerator designed to bridge the gap between initial concept and full research project. The Studio provides structured support for researchers to rapidly develop promising ideas into polished outputs to get external feedback.



The Studio has already demonstrated impressive results, with many high-quality research blog posts and technical demos emerging from the program. Many of these have attracted significant attention from the broader AI safety community and served as foundations for more comprehensive research projects.

APART LAB FELLOWSHIP

Our Lab Fellowship provides comprehensive support for promising researchers pursuing publication-quality AI safety work. In 2024, we welcomed 36 new fellows in the last three months alone, bringing our total to 105 researchers incubated since the program's inception.



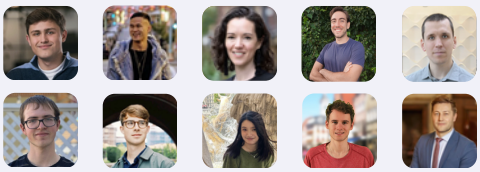
Visit more of our Lab fellows here: apartresearch.com/lab

Grand Challenges

Case: Economics of Transformative AI Grand Challenge

In April 2025, Apart Research launched our Grand Challenges event series with a major collaborative research event for economists, AI researchers, and policy experts. We had three tracks; 1) analyzing AI's distribution effects, 2) explore how AI impacts growth, and 3) identify radical shifts in market dynamics.

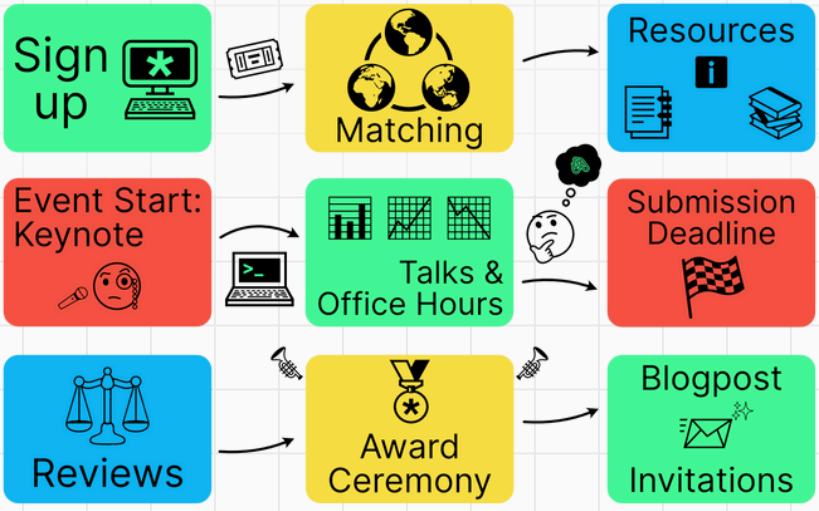
175 participants submitted 12 original research reports



Featuring renowned scholars in economics, such as Anton Korinek. Commissioned economics researchers authored the *Concrete Open Problems* and our *Literature Review* documents. $600 in awards for each track was given out after the judges reviewed the projects.

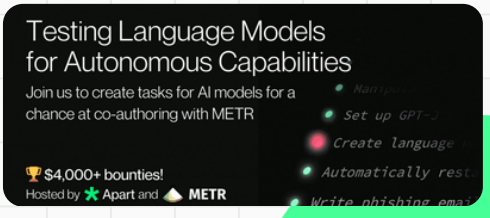
BlueDot Impact

The event was co-located with BlueDot's course on "Economics of Transformative AI"



While our main priority was to receive high-quality task implementations, the hackathon was also a great chance to engage with the research community, improve our documentation, and spread awareness about our ongoing task bounty.

Beth Barnes
METR



Talent Development

Our programs have impacted the careers of dozens of researchers. Apart Lab fellows went on to work at top organizations such as METR, Far.ai, Oxford University, and as founders and team leads at impactful startups.

"Looking back, I can trace a clear line from that first hackathon through each subsequent opportunity. Apart created the conditions for me to develop from someone interested in AI safety to someone actively contributing to the field."

Prithvi Shahani, Auto Alignment Eval Builder and Research Trainer (Contractor) at Anthropic

"[...] an organisation like Apart Lab which is willing to help individuals for free is a god send. [...] If (when!) I land a job in AI Safety it will be because of Apart Lab's help. It's a great organisation."

Philip Quirke, now AI Safety Research Lead at Martian

"Apart profoundly impacted my career trajectory, and supported me throughout, I am extremely grateful for the opportunity, experience and the team. All of my work in AI safety traces back to my involvement with Apart."

Zainab Ali Majid, Co-Founder of a VC backed AI Security Startup

"My life changed when I decided to apply for a deception detection hackathon run by Apart Research."

Cameron Tice, Technical Governance Research Lead at ERA Cambridge

Advocacy

Deeply ingrained into our mission at Apart is to advocate for Safer AI, shape policy and make expert opinions heard.

2,600	2,000	60,000	3,200
NEWSLETTER SUBSCRIBERS	FOLLOWERS ON LINKEDIN	VIEWS ACROSS YOUTUBE	USERS ON DISCORD

Our public-facing work has emphasized both technical concerns and practical solutions. As highlighted in one of our podcast appearances:

> **We need to choose the world we want to live in and then proactively design it.**
>
> — Esben Kran

This philosophy guides our approach to public engagement: helping diverse audiences understand both the risks of advanced AI and the pathways to ensuring it remains aligned with human values.

In addition to media engagement, Apart champions safe, ethical, and trustworthy AI by working with communities on every level, from giving talks at universities and local interest groups, to joining panels at international conferences.

Featured in

DIE ZEIT

International Association for Safe & Ethical AI

FOR HUMANITY
AN AI SAFETY PODCAST

Into AI Safety

Policy

We rapidly translate technical research into policy impact by developing empirical evidence and solutions for AI risks, and ensuring our insights reach decision-makers when it matters most.

Contributions to EU AI governance:

EUROPEAN **ARTIFICIAL INTELLIGENCE** OFFICE

- Contributing expert technical guidance to the EU AI Office (alongside Alan Turing Institute, RAND, SaferAI) during expert panels on cybersecurity evaluation methodologies for frontier AI
- Co-authoring a framework paper with the EU AI Office on systemic risk evaluation practices for advanced AI systems (available soon)
- Joining the European Network of AI Evaluators, technical experts building the evaluation ecosystem for general-purpose AI models
- Serving as expert consultants to multi-stakeholder working groups for the EU AI Act Code of Practice, providing input on evaluations including external safety testing access and evaluation robustness

RESEARCH HIGHLIGHT: EVALUATIONS FRAMEWORK

In "Rethinking CyberSecEval: An LLM-Aided Approach to Evaluation Critique," Apart Lab fellows:

- **Identified critical limitations** in cybersecurity evaluations methodology
- Demonstrated **how to improve evaluation robustness** by LLM-assisted benchmark analysis
- **Presented their insights** at multiple venues, including Brussels expert panel, helping shape more effective implementation guidance for the AI Act's evaluation standards.

SB 1047

CALIFORNIA REPUBLIC

We **support responsible AI governance globally**. Apart co-sponsored California Senate Bill 1047 (the Safe and Secure Innovation for Frontier AI Models Act), which called for security protocols for advanced AI models and received endorsements from Turing Award winners Geoffrey Hinton and Yoshua Bengio.

Partnerships

Apart is at the center of a lively AI safety ecosystem. We are connecting researchers, organisations, and geographies.

Deeply embedded with the AI safety community, we have a long list of supporters and partners we rely on to accelerate technical talent.

Research Partners

Organisations we have partnered with to enable novel research

METR ANTHROP\C GOODFIRE APOLLO RESEARCH

Our Code Red Hackathon resulted in over $30k of collected bounties and prizes, growing the evaluation suites of both METR and UK AISI

Community Partners

Organisations we have partnered with to reach and support more people.

BlueDot Lambda PIBBSS Berkeley UNIVERSITY OF CALIFORNIA

Lambda provides compute to hackathon participants ($500) and members of the lab ($5000)

Ecosystem Partners

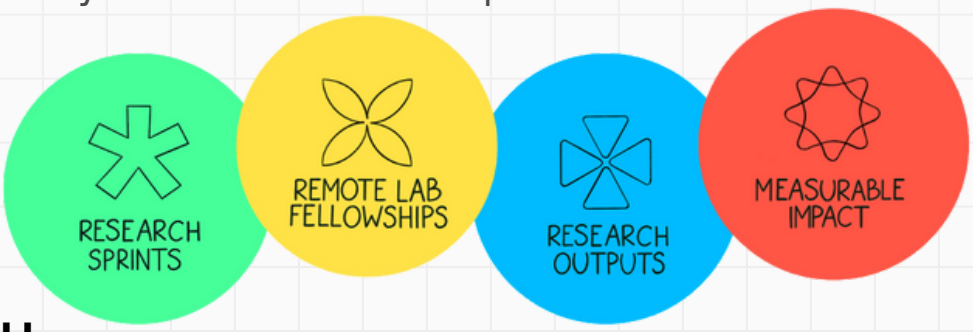
Organizations we have partnered with to amplify our impact across domains

AISI AI SECURITY INSTITUTE ENTREPRENEUR FIRST Cooperative AI Foundation JUNIPER VENTURES

A hackathon co-organized with CAIF directly resulted in many contributions to their February 2025 report on multi-agent AI risks

The Future

The accelerating pace of AI development demands more rigorous safety research with faster real-world impact. In the next 12 months, we will double down on highly innovative empirical research and even more efficient pathways from research to impact.



RESEARCH

We will **target empirical AI safety research areas with outsized impact potential**, focusing on critical neglected topics and responding to time-sensitive opportunities for technical policy input. Our scalable and flexible research model will allow us to react quickly to emerging opportunities and needs.

FIELD BUILDING

Our Grand Challenges will set a new standard for collaborative research quality, designed specifically to **engage experienced technical talent who can contribute immediately to AI safety work**. By focusing on relevant technical backgrounds across disciplines adjacent to AI safety, we'll maximize counterfactual impact and accelerate progress on critical safety challenges.

ADVOCACY

We will translate complex technical insights into accessible resources for policymakers and the public. Our ambition is to produce cutting-edge research and **ensure that this research directly improves the safety of AI development and deployment worldwide**.

Apart

Apart Research

Donate to Apart by going to apartresearch.com/donate or by writing to hi@apartresearch.com

🔗 Support Apart

Donations to Apart are tax-deductible through our fiscal sponsor Ashgro Inc, a 501(c)(3) organization.