AIS Collab Germany Report - 2023 Winter & 2024 Summer (public, still in progress)

TL;DR (2–3 Sentences)

The AIS Collab Germany program introduces students and professionals to AI Safety through discussion groups, readings, and optional projects. Around half of the participants remained engaged throughout, with about a quarter indicating plans to pursue a long-term career in AI Safety. While satisfaction was solid, several improvements (e.g., mandatory exercises, more consistent materials) could increase participant commitment and overall impact.

Executive Summary

Audience:

- **Internal Organizers and Instructors:** For refining course design, moderation strategies, and resource allocation.
- **Potential Partners/Funders (e.g., BlueDot, ENAIS, Universities):** To understand the program's impact, efficiency, and scalability.
- **Prospective Participants:** To gauge the program's intensity, relevance, and expected outcomes.

Course Target Group:

- Early-career students, graduates, and professionals interested in exploring AI Safety, Governance, and Alignment.
- Individuals new to AI Safety who seek structured, community-supported learning with a moderate weekly time commitment (2–3 hours).

Key Highlights:

- About 60 active participants per iteration, with attendance declining over time.
- Counterfactual impact: Approximately 25–50% of participants would not have engaged with AI Safety otherwise.
- Around 25% of participants report a long-term career interest in AI Safety.
- **Organizational effort per successful participant is roughly 9–10 hours, similar to other intro programs.**

- Potential improvements: more rigorous structure (mandatory exercises, final projects, consistent reading materials), better participant screening, and enhanced facilitator support.

Further Reading and Resources

For additional materials and a detailed breakdown of the report:

- AI Safety Collab Facilitator Hub
- Detailed Report Document

Contact Information

For feedback, questions, or to request materials, please contact:
aisafetycollabgermany@gmail.com.

Introduction

AIS Collab Germany aims to provide a structured introduction to AI Safety topics, including Alignment and Governance. Through weekly discussion groups, curated readings, and opportunities for project work, participants gain foundational knowledge, network with peers, and explore career directions in the field. This report covers two cohorts:

- **Winter 2023**
- **Summer 2024**

It evaluates their impact, participant feedback, and cost-effectiveness to inform future improvements and guide stakeholders in understanding the program's value.

Methodology and Data Collection

Data Sources:

- Attendance logs and participation metrics were recorded weekly.
- Feedback from participants and moderators was collected via anonymous surveys at the end of the course and optionally after each session.
- Participation in project work and career-related outcomes were self-reported in end-of-course questionnaires.

Data Limitations:

- Feedback response rates (Winter: ~18/63; Summer: ~8/60) were moderate, reducing representativeness.
- Results should be considered indicative rather than definitive.

Participation and Performance

Winter 2023

- **Applicants:** 118 (95%+ confirmed)
- **Active participants (≥1 session):** 63
- **No-shows:** 55

Weekly Attendance: Started at ~90%, dropping to ~40% by week 8.

Completion: 30 “successful” completions (15 Governance, 15 Alignment)
Projects Submitted: <5
Moderation: 16 moderators, ~10 online groups, 2–4 in-person groups

Key Feedback:

- Recommend to a peer (1–10): ~7.8
- Material depth (1–5): ~3.9
- Prep time: ~135 minutes/week (aligned with stated 2–3 hours)
- Session value (1–5): ~4.2
- Moderator quality (1–5): ~4.3
- Community feeling (1–10): ~8.9

Impact:

- **Counterfactual Engagement:** Approximately 50% of participants indicated that they would not have engaged with AI Safety content or programs (e.g., BlueDot) without this course. This highlights the program’s role as a gateway to new audiences.
- **Career Interest:** Around 25% of participants expressed plans to pursue AI Safety-related career paths (e.g., internships, theses, or long-term shifts).

Summer 2024

- **Applicants:** 83
- **Accepted:** 82 (~98.8%)
- **Active participants (≥1 session):** 60

Weekly Attendance: Started ~78%, declined to ~38% by week 6–7, and ended around 48%.

Completion: 38 (6 Governance, 32 Alignment)
Projects Submitted: 2
Moderation: 9 moderators, 8 online groups, 1 in-person group

Key Feedback:

- Recommend to a peer (1–10): ~7
- Material depth (1–5): ~4
- Prep time: ~145 minutes/week
- Session value (1–5): ~4
- Moderator quality (1–5): ~4.125
- Community feeling (1–10): ~9.125
- **Counterfactual Engagement:** Approximately 25% of participants would not have participated in AI Safety initiatives without this program.
- **Career Interest:** Around 25% plan to explore career trajectories related to AI Safety.

Spring 2025

Participation and Performance

- **Applicants**: 372 (minimal outreach, primarily through established local groups)
- **Accepted**: 312 (~83.9%)
- **Active participants** (assigned to groups): 236
- **Weekly Attendance**: [Data to be added based on final records]
- **Completion**: [Data to be added once program concludes]
- **Projects Submitted**: [Data to be added once program concludes]
- **Moderation**: [Number of moderators and groups to be added]

Locations

Participants for the AI Safety Collab program in Summer 2025 represent diverse global regions, totaling 234 individuals from various locations.

The top regions by participant count include:

- London, UK: 32 participants
- Paris, France: 15 participants
- Berlin, Germany: 12 participants
- Moscow, Russia: 10 participants
- Bengaluru, India: 9 participants

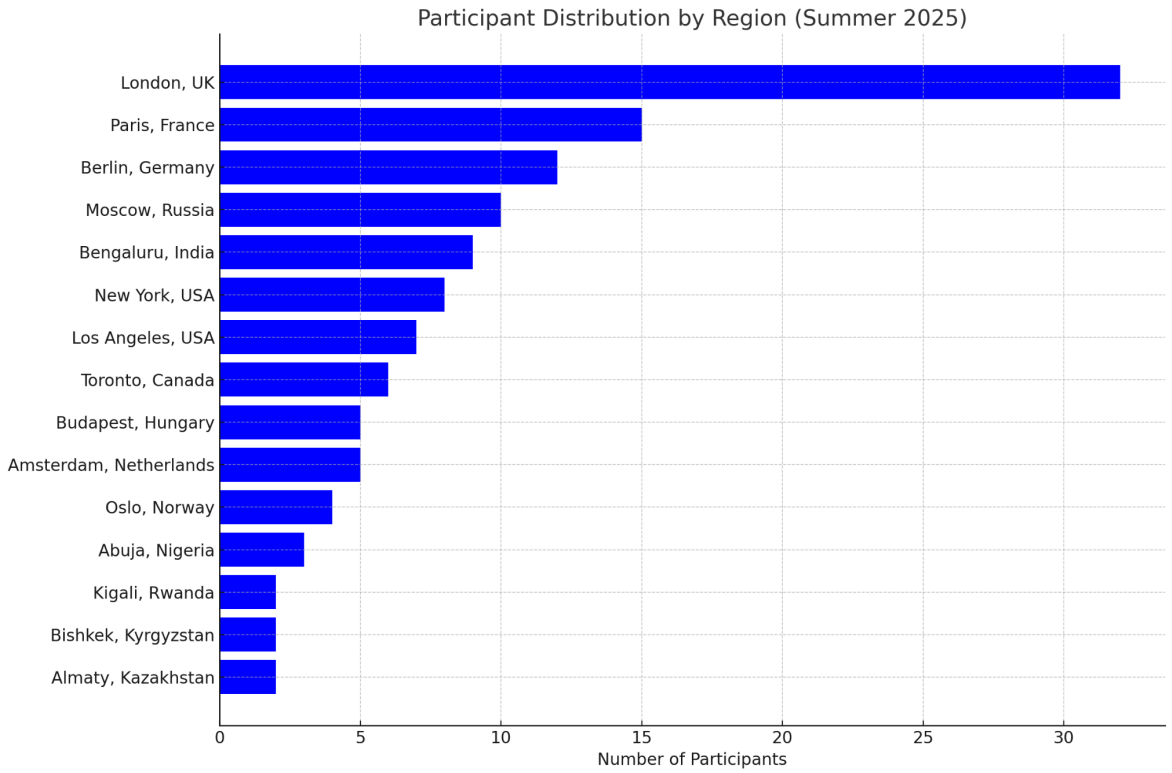
Additional notable regions:

- New York, USA: 8 participants
- Los Angeles, USA: 7 participants
- Toronto, Canada: 6 participants
- Budapest, Hungary: 5 participants
- Amsterdam, Netherlands: 5 participants
- Oslo, Norway: 4 participants

Participants from traditionally underrepresented regions include:

- Abuja, Nigeria: 3 participants
- Kigali, Rwanda: 2 participants
- Bishkek, Kyrgyzstan: 2 participants

- Almaty, Kazakhstan: 2 participants



Key Operational Highlights

- Program running smoothly overall with significant operational improvements:
  - Automated several administrative processes
  - Enhanced quality of documentation and resources
  - Streamlined workflows to increase operational efficiency
  - Better time allocation for strategic program development

Preliminary Observations

- Increased participant intake represents approximately **5× growth** from Summer 2024 cohort
- Improved acceptance-to-active participation ratio (76% vs. 73% in Summer 2024)
- [Additional observations to be added as program progresses]

Initial Feedback Themes

- [To be completed based on mid-program feedback]

Ongoing Improvements

- Continued refinement of program materials
- Enhanced facilitator support and training
- More efficient group formation and management

- [Additional improvements to be added]

Next Steps

- Complete comprehensive end-of-program survey
- Analyze attendance patterns and completion rates
- Evaluate counterfactual impact and career interest indicators
- Prepare detailed recommendations for future iterations

Note: This section is in progress and will be updated as the Spring 2025 cohort concludes and final data becomes available.

Recommendations

Increase Rigor and Participant Investment:

- Mandatory exercises, final quizzes, or exams to ensure knowledge retention.
- Make final projects obligatory with small incentives (e.g., a prize) to boost engagement.
- Stricter requirements for certificates to raise the program's perceived value.

Improve Materials and Structure:

- Use stable, well-curated materials (e.g., BlueDot Impact) until the custom textbook matures.
- Integrate more interactive components (role-plays, debates, breakout discussions).

Better Group Formation and Communication:

- Collect availability upfront and form groups based on common time slots.
- Hold mid-course moderator meetups to share lessons learned.
- Offer optional additional sessions to enhance network building.

Refined Promotion and Screening:

- Clearly communicate time commitment and expectations to attract committed participants.
- Expand outreach to non-EA networks and academic fields like CS or political science.

Long-Term Monitoring and Guidance:

- Conduct a 3-month follow-up survey to track sustained impact.
- Provide a "personal action plan" template to guide participants' next steps in AI Safety.

Future Steps

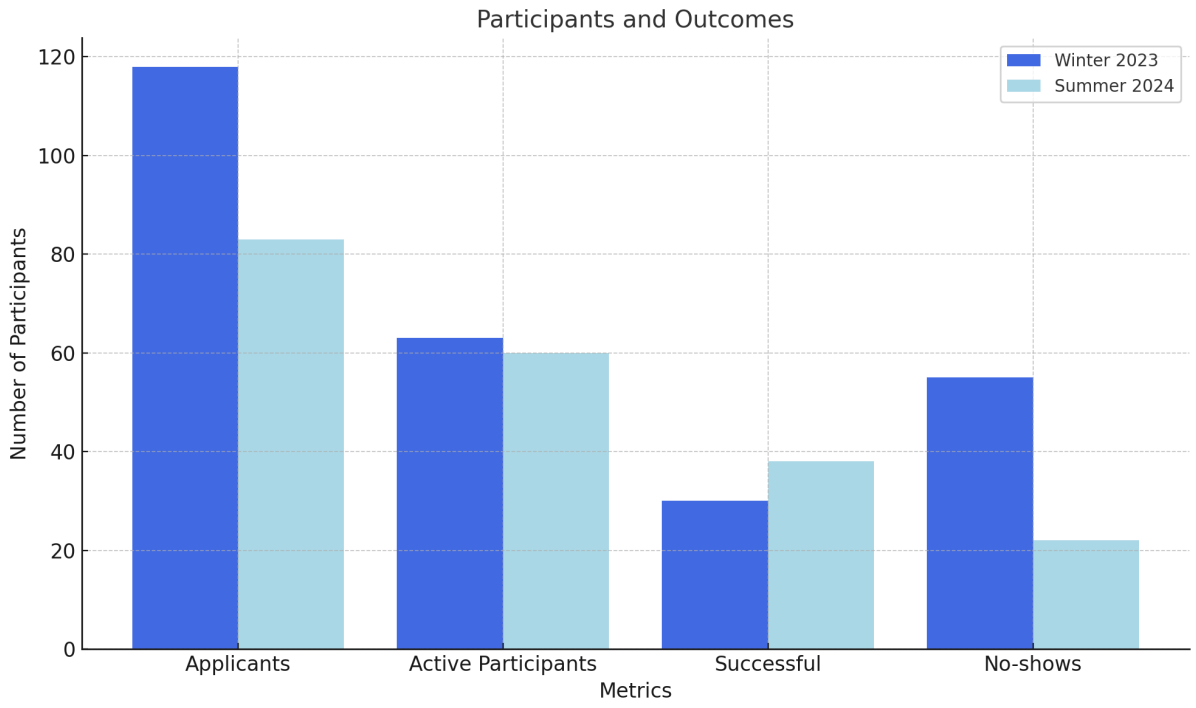
- **International Expansion:** Collaborate with AI Safety Collab, BlueDot, and ENAIS to scale internationally while preserving regional contexts.
- **Regional Customization:** Introduce Germany/EU-specific governance readings and materials.
- **Iterative Improvement:** Implement recommended changes, then re-measure outcomes to assess progress.

Conclusion

AIS Collab Germany successfully introduces newcomers to AI Safety, catalyzing interest and some career shifts. However, opportunities exist to improve commitment, streamline course materials, and enhance overall impact. By adopting more rigorous requirements, refining materials, and strengthening group cohesion, the program can better retain participants, deepen learning, and further boost its counterfactual and career-related impact.

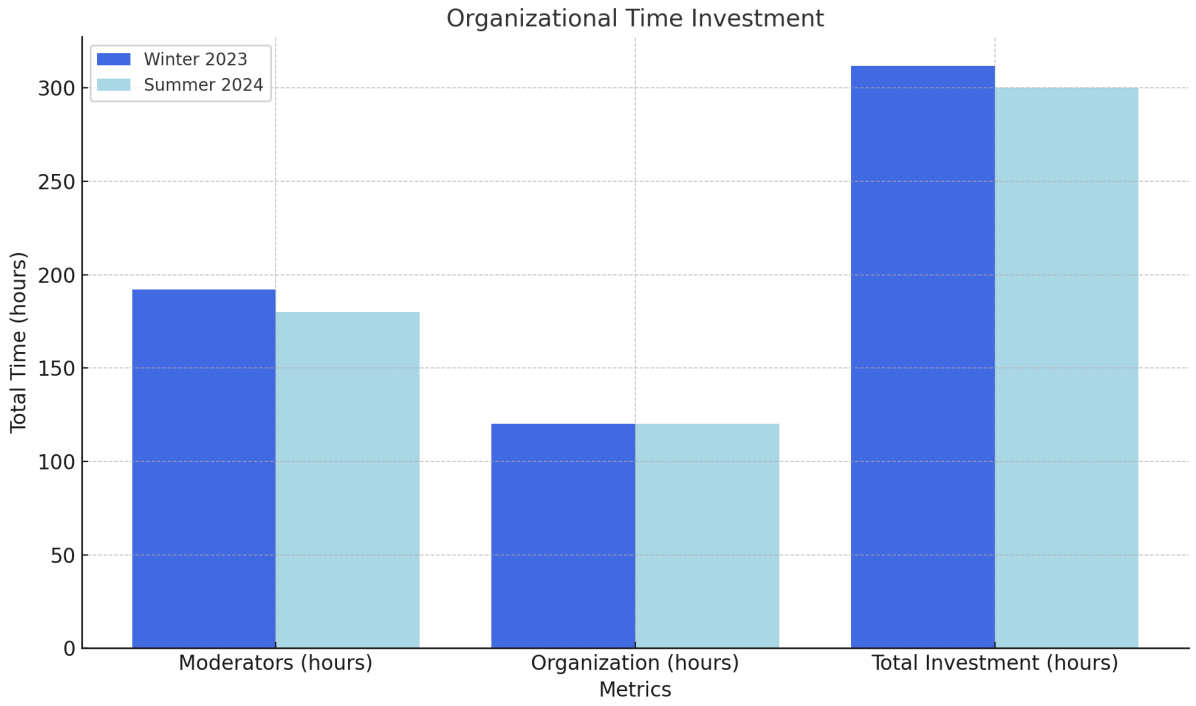
Visualizations

Participants and Outcomes



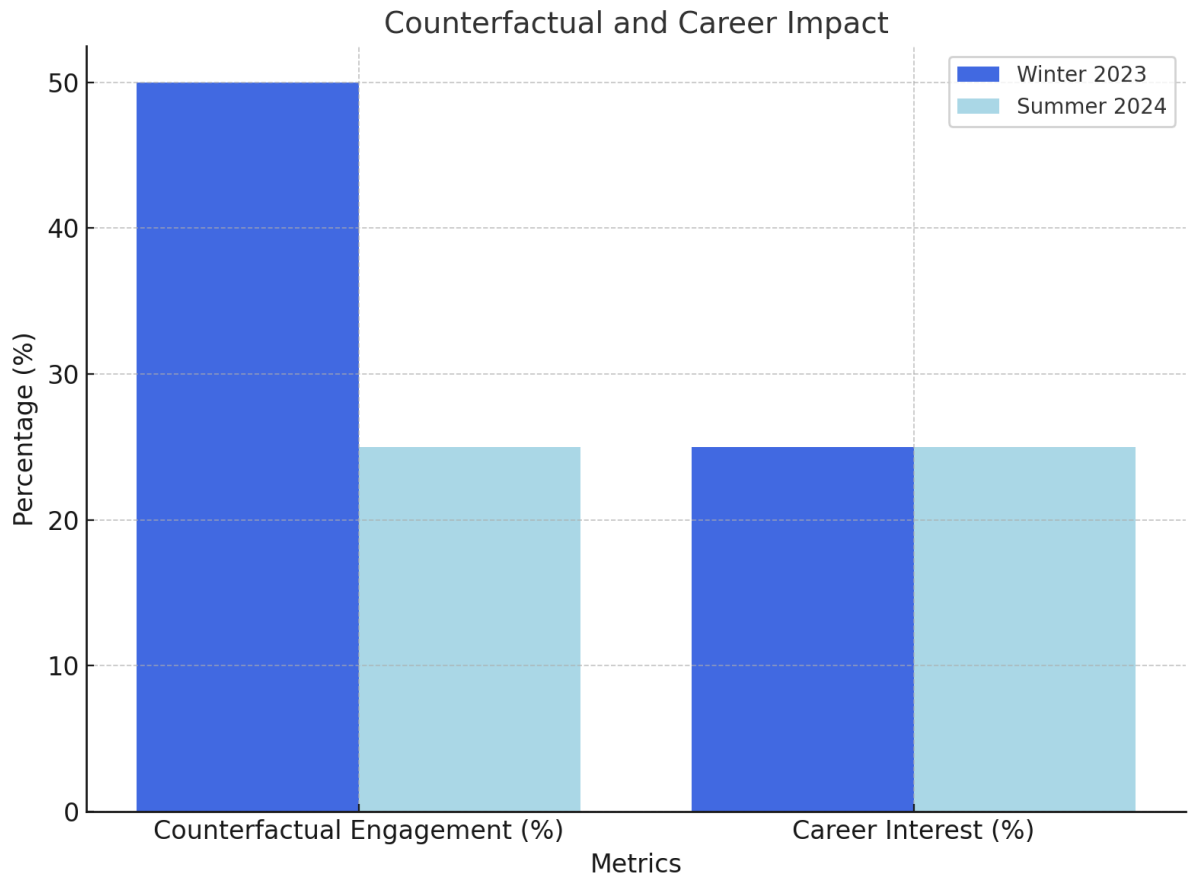
This chart illustrates the number of applicants, active participants, successful completions, and no-shows for both Winter 2023 and Summer 2024 cohorts.

Organizational Time Investment



This chart compares the total hours invested by moderators, organizers, and the combined total for both program iterations.

Counterfactual and Career Impact



This chart highlights the percentage of participants who counterfactually engaged with AI Safety and those expressing long-term career interest.

For feedback, questions, or to request materials, please contact:
aisafetycollabgermany@gmail.com