Building Bridges for AI Safety: Proposal for a Collaborative Platform for Alumni and Researchers

Aisha Gurung
University of Bath

The AI Safety Society is a centralized platform designed to support alumni of AI safety programs, such as SPAR, MATS, and ARENA, as well as independent researchers in the field. By providing access to resources, mentorship, collaboration opportunities, and shared infrastructure, the Society empowers its members to advance impactful work in AI safety. Through institutional and individual subscription models, the Society ensures accessibility across diverse geographies and demographics while fostering global collaboration. This initiative aims to address current gaps in resource access, collaboration, and mentorship, while also building a vibrant community that accelerates progress in the field of AI safety.

1. Problem overview: The Need for Centralized Alumni Support

AI safety programs have successfully trained talented individuals, but many face challenges after program completion due to limited resources and infrastructure. These challenges include:

- **Resource Gaps**: Alumni lack access to critical tools like academic journals, research software, and computational resources.
- **Fragmentation**: Collaboration is hindered as alumni lack a unified platform, often leading to duplicated efforts and inefficiencies when multiple individuals unknowingly work on the same problems.
- **Mentorship Deficit**: Structured guidance and peer mentorship opportunities are scarce, making it difficult for researchers to navigate career paths effectively.
- **Talent Retention**: Without adequate support, many promising individuals transition out of AI safety research.

Institutional collaboration can mitigate these issues by pooling resources, expertise, and infrastructure. By building a global, interdisciplinary platform, the AI Safety Society bridges these gaps, providing the necessary foundation for sustained impact. Research highlights the importance of informal learning and professional networks in developing expertise and fostering innovation. The Society addresses these needs by:

- **Expanding Academic Opportunities**: Members can access specialized programs, joint degrees, and cross-disciplinary learning opportunities.
- **Enhancing Networking**: Strong professional connections increase access to advice, job openings, and collaborative opportunities. Networking is a trainable skill, and the Society plans to host workshops to help members refine this ability.
- **Fostering Soft Skills**: Communication, teamwork, adaptability, and emotional intelligence are vital in advancing a novel field like AI safety.

Workshops and training sessions will help members develop these competencies.

- **Encouraging Good Community Members**: As Chris Leong (2024) emphasizes, attributes like intellectual humility, emotional support, and the ability to provide constructive feedback are invaluable. The Society cultivates these qualities to create a collaborative and supportive environment.

2. Proposed Solution

Vision

1. Collaborate with universities and institutions with active AI safety groups to strengthen academic ties.
2. Implement a dual subscription model: an institutional subscription for AI safety programs and a tiered individual subscription for independent researchers, tailored to geographic and demographic diversity.
3. Provide members with access to shared resources, mentorship opportunities, and global collaboration tools to enable impactful work.

Key Features

1. **Resource Access**:
    - Institutional subscriptions to journals and research software.
    - Discounts on essential tools, materials, and conference tickets.
    - Access to shared databases, simulation tools, and other advanced AI research tools.
    - Opportunities to join global academic and professional consortia.
    - Subsidized memberships for relevant professional organizations.
2. **Community Building**:
    - A global platform with local hubs to foster collaboration.
    - Pro bono career counseling and tailored mentorship programs.
    - Networking opportunities to address shared challenges and support career growth.
3. **Collaboration Facilitation**:
    - Centralized directories of open problems, ongoing projects, and alumni expertise.
    - Streamlined mentee-mentor connections through curated profiles.
    - Tools to enhance collaboration, reduce duplication of efforts, and improve talent retention.
4. **Knowledge and Event Sharing**:
    - Regular workshops, speaker events, and panel discussions.
    - An updated calendar for grant deadlines, conferences, and other opportunities.

3. Pilot Phase

During the pilot phase, the Society aims to:

1. Develop a low-code prototype of the platform with key features.

2. Gather feedback from alumni and researchers to refine the design.
3. Test functionalities such as resource sharing, mentorship matching, and project directories.

Key Deliverables

1. A functional prototype with basic features.
2. Initial feedback and survey data to guide improvements.
3. A roadmap for scaling the platform post-pilot.

4. Process

Timeframe	What will you do?
Next 3 months	Do a comprehensive SWOT Analysis. Work on a Theory of Change. Create a detailed budget, one-year plan. Focus on running market research and identifying collaborators. Do
2025	Finalize the platform prototype. Engage 50-100 beta testers from existing alumni. Apply for additional funding. Build partnerships with key AI safety programs. Launch the full platform with subscription models.
2026	Expand mentorship programs and initiate local hubs. Refine services based on user feedback. Introduce advanced features like mentorship pairing and collaborative project tools.
2027	Achieve full-scale implementation with a robust alumni base, rich resources, and sustainable funding models.

5. Impact on AI safety & key risks

The AI Safety Society represents a bold experiment in addressing the systemic challenges faced by the AI safety community. While the risks are real, the potential to create a transformative, supportive, and collaborative ecosystem makes this initiative worth pursuing. This unified effort has the potential to transform AI safety into a more connected, innovative, and effective domain.

Metrics for Success

- Number of active users and institutional partnerships.
- Volume and quality of collaborative projects.
- User satisfaction and retention rates.
- Increased diversity and inclusion within the community.

Potential Failure Modes

1. **Lack of Adoption**: The platform might struggle to attract a critical mass of users, either due to limited awareness or skepticism about its value.
2. **Fragmentation of Efforts**: Instead of uniting the AI safety community, the platform might inadvertently compete with existing initiatives, leading to further division.
3. **Financial Sustainability**: Balancing the costs of maintaining the platform with affordable subscription fees could prove challenging, particularly for independent researchers in underrepresented regions.
4. **Overcomplexity**: Attempting to address too many needs might dilute the platform's focus and effectiveness, resulting in an overwhelming or underutilized resource.
5. **Cultural and Geographic Barriers**: Differences in norms, priorities, and resources across regions could hinder the creation of a cohesive global community.

Despite the risks, the potential benefits of the AI Safety Society justify its exploration:

- **Accelerating Progress**: By connecting alumni and researchers globally, the platform reduces fragmentation, fosters collaboration, and improves resource access, accelerating progress in AI safety.
- **Accessibility and Inclusivity**: Tailored subscription models ensure diverse participants, including independent researchers from underrepresented regions, have access to essential resources and opportunities.
- **Fostering Innovation**: Interdisciplinary and collaborative projects encourage creative solutions to complex AI safety challenges.
- **Talent Retention**: Structured mentorship and support systems help retain skilled individuals within the AI safety field.
- **Scalability and Growth**: A centralized hub provides a scalable solution to meet the growing needs of the community, filling critical gaps in mentorship, networking, and resource availability.
- **Learning and Experimentation**: Even with setbacks, the initiative generates insights into effective strategies for global collaboration and platform development.
- **Building Momentum**: A vibrant community can catalyze the creation of new organizations, projects, and initiatives, amplifying the overall impact on the field.

Adoption challenges can be potentially addressed through a user-friendly platform design and active community engagement to build trust and encourage participation. Funding constraints can be mitigated by diversifying revenue streams, including institutional subscriptions, individual memberships, sponsorships, and grants. Scalability issues will be tackled with a modular platform design and phased implementation, ensuring the system can grow and adapt to meet the increasing needs of the AI safety community.

6. Appendix

1. **Market research questionnaire:** To ensure the platform effectively addresses the needs of its stakeholders, we designed two tailored questionnaires: one for organizers of AI safety programs and another for participants. These tools aim to gather insights on current barriers, unmet needs, and potential areas where the Society could provide value. The structure of these forms is inspired by the AISRP (AI Safety Research Pathways) initiative, which has successfully gathered similar data. AISRP would also be a key organization to engage for feedback and collaboration.

    a. **Questionnaire for Organizers of AI Safety Programs**

    Section 1: Program Overview

    1. What is the primary focus of your program? (e.g., AI alignment, governance, technical safety, policy, other)
    2. How long has your program been running?
    3. How many participants do you admit per cohort or year?
    4. What criteria do you use to select participants for your program?

    Section 2: Goals and Outcomes

    5. What are the primary goals you hope participants achieve by the end of your program?
    6. How do you measure the success of your program? (e.g., participant feedback, post-program outcomes, published research, etc.)
    7. Do you track alumni progress after program completion? If yes, how?

    Section 3: Barriers and Challenges

    8. What are the biggest challenges you face in running the program? (e.g., funding, staff capacity, lack of resources, etc.)
    9. What gaps do you observe in the resources available to support alumni post-program?
    10. How do you address participants' needs for mentorship, collaboration, and career guidance?

    Section 4: Collaboration and Resources

    11. Would you be interested in a shared platform to collaborate with other AI safety programs? Why or why not?

12. What resources or features would you value most in a centralized platform? (e.g., shared access to journals, mentorship databases, event organization, etc.)
13. Are there any specific areas where you feel a centralized platform could alleviate existing challenges in your program?

Section 5: Future Opportunities

14. How can a global alumni network support your program's goals?
15. Would you be willing to adopt a subscription model for your program's alumni to access centralized resources? What pricing or tiers would be appropriate?

b. Questionnaire for Participants of AI Safety Programs

Section 1: Background

1. Which AI safety program(s) have you participated in? (e.g., SPAR, MATS, ARENA, etc.)
2. What motivated you to join the program(s)?
3. How would you describe your current involvement in AI safety research?
   1 = Learning fundamentals | 3 = Actively contributing to a project | 5 = Publishing or leading research efforts
4. What are your career aspirations within the AI safety field?

Section 2: Resources and Support

5. What resources have been the most valuable during your AI safety journey? (e.g., access to mentors, funding, collaboration opportunities)
6. How much do the following factors influence your ability to engage in AI safety research? (Use a sliding scale or rank order)
   - Access to mentors
   - Career capital
   - Stable income
   - Peer support
   - Availability of research tools or software
   - Clarity on promising research directions
   - Confidence in personal impact
7. What are the biggest challenges or unmet needs you have encountered in pursuing AI safety research?

Section 3: Networking and Collaboration

8. How do you currently connect with peers, mentors, or collaborators in the field?
9. Would a centralized platform for alumni and researchers be useful for you? Why or why not?
10. What features would you prioritize in such a platform? (e.g., mentorship connections, shared tools, job postings, event organization)

Section 4: Post-Program Support

11. What kinds of support do you currently lack after completing an AI safety program?
12. Do you feel adequately prepared to pursue a career in AI safety after completing your program? Why or why not?
13. Have you faced difficulties in accessing post-program resources such as academic journals, funding, or computational tools?

Section 5: Feedback and Recommendations

14. What suggestions do you have for improving the experience of AI safety programs?
15. Is there anything else you would like to add regarding how a collaborative platform could support your goals in AI safety?

2. Team and Fit

The ideal team structure for the pilot phase and real-world implementation includes:

Pilot Phase:

- Understanding of AI Safety Programs and Communities: Team members should be familiar with AI safety research, tools, and the landscape of existing programs.
- Product Design Expertise: Ensures the prototype aligns with user needs and delivers a functional, user-friendly experience.
- Familiarity with Academic and Research Tools: Supports the integration of resources such as journals, software, and other scholarly tools.
- Basic Web Development Skills: Proficiency in low-code or no-code platforms to rapidly create and test a functional prototype.

Real-World Implementation (Post-Hackathon):

- Operations Expertise: For efficient platform management and scaling operations.

- Market Research Skills: To identify and address user needs, gaps, and opportunities in the AI safety ecosystem.
- User Experience Design: Ensures the platform remains intuitive and valuable for users.
- Community Building and Management: Facilitates collaboration and engagement across the global AI safety community.
- Fundraising Experience: Secures the financial sustainability of the platform through diverse funding sources.

Founder Bio: Aishwarya Gurung, LinkedIn

With over two years of experience in AI policy and governance, I am passionate about facilitating the development and implementation of ethical and human-centric AI solutions. I am currently a member of research group at the Center for AI and Digital Policy (CAIDP), where I contribute to developing analytical skills and conducting empirical research on national AI policies and practices, while also engaging in collaborative writing, monitoring AI policy news, identifying public voice opportunities, and participating in significant AI policy events. I previously founded Effective Altruism Nepal and collaborated on community-building initiatives, local research, and poverty alleviation efforts, which helps me bring a wealth of experience in addressing pressing social issues. My involvement with NGOs in Nepal, particularly in digitizing educational content for the Ministry of Education and developing interactive games to teach STEAM subjects to students from grade 1 to 8, showcases my experience in leading innovative education solutions.

In addition, I support the operations of ForHumanity, a non-profit organization that promotes AI safety, ethics, and transparency. I manage the social media accounts, the Slack community, and the ForHumanity University, as well as provide essential assistance to the Executive Director. I have a Bachelor's degree in Mechanical Engineering from Embry-Riddle Aeronautical University, a MicroMasters in Emerging Automotive Technologies from Chalmers University of Technology, and I am pursuing a Master's degree in Artificial Intelligence from the University of Bath. I have also obtained multiple certifications in AI from IBM and the University of Helsinki. I am fluent in English, and Nepali, and have working knowledge of Urdu and Hindi.

3. References and sources

Covarrubias, A. and Veld, S. (2024) *Ai Safety University Organizing: Early takeaways from thirteen groups*, *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/M4QLjhua5KmmePtvi/ai-safety-university-organizing-early-takeaways-from (Accessed: 19 January 2025).

Frazier, K. (2024) *The AI safety community exists, but its impact is uncertain*, *Tech Policy Press*. Available at: https://www.techpolicy.press/the-ai-safety-community-exists-but-its-impact-is-uncertain/ (Accessed: 19 January 2025).

Hobbhahn, M. (2023) *There should be more AI safety Orgs*, *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/dua879FhtLf9jqyJo/there-should-be-more-ai-safety-orgs (Accessed: 19 January 2025).

Kran, E. (2024) *AI safety needs to scale, and here's how you can do it* , *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/kc9igJKXeJbmfxsxq/ai-safety-needs-to-scale-and-here-s-how-you-can-do-it (Accessed: 19 January 2025).

Leong, C. (2024) *The bar for contributing to AI safety is lower than you think*, *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/dBxfzHX2bsXD9SAiR/the-bar-for-contributing-to-ai-safety-is-lower-than-you (Accessed: 19 January 2025).

Lundy, K. and Ladd, H. (2020) *Why collaboration is key to the future of Higher Education*, *EY Parthenon*. Available at: https://www.ey.com/en_gl/insights/strategy/strategies-for-collaborating-in-a-new-era-for-higher-education (Accessed: 19 January 2025).

Madsen, A. (2020) *Becoming an independent researcher and getting published in ICLR with spotlight*, *Medium*. Available at: https://andreas-madsen.medium.com/becoming-an-independent-researcher-and-getting-published-in-iclr-with-spotlight-c93ef0b39b8b (Accessed: 19 January 2025).

McLean, E. (2024a) *Big picture ai safety: Introduction*, *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/uJioXCz5Foo9eqpJ9/big-picture-ai-safety-introduction (Accessed: 19 January 2025).

McLean, E. (2024b) *What mistakes has the AI safety movement made?*, *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/tEmQrfMs9qdBPrGKh/what-mistakes-has-the-ai-safety-movement-made (Accessed: 19 January 2025).

McLean, E. (2024c) *What should AI safety be trying to achieve?* , *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/EsYRThpymeFvNLQQY/what-should-ai-safety-be-trying-to-achieve (Accessed: 19 January 2025).

National Academy of Sciences, National Academy of Sciences and Institute of Medicine (2005) 'Chapter: 7 The Role of Professional Societies', in *Facilitating Interdisciplinary Research*. Washington, DC: The National Academies Press, pp. 137–148. Available at: https://nap.nationalacademies.org/read/11153/chapter/9 (Accessed: 19 January 2025).

Oliveira, C. (2024) *Aspiring Jr.. AI safety researchers: What's stopping you?: Survey*, *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/FpG8H53aCRfpM6JPS/aspiring-jr-ai-safety-researchers-what-s-stopping-you-or-1 (Accessed: 19 January 2025).

Professional Development for Policy Careers (2024) *Emerging Technology Policy Careers* . Available at: https://emergingtechpolicy.org/tips/professional-development/ (Accessed: 19 January 2025).

Vail, E.A. *et al.* (2020) 'The role of professional organizations in fostering the Early Career Development of Academic intensivists', *Annals of the American Thoracic Society*, 17(4), pp. 412–418. doi:10.1513/annalsats.201908-573ps.

'Functions of professional bodies' (2025) *Understanding your sector*, 19 January. Available at: https://www.open.edu/openlearn/mod/oucontent/view.php?id=64020&section=1.1 (Accessed: 19 January 2025).