
BlueDot Impact Connect: A Comprehensive AI Safety Community Platform

Elise Racine
University of Oxford

Aliya Koishina
New York University

Aishwarya Gurung
University of Bath

Tatenda Mawema
IMM Graduate School

Haihao Liu
Harvard University

In collaboration with **Apart Research** and **BlueDot Impact**¹

Abstract

The AI safety field faces a critical challenge: while formal education resources are growing, personalized guidance and community connections remain scarce, especially for newcomers from diverse backgrounds. We propose BlueDot Impact Connect, a comprehensive AI Safety Community Platform designed to address this gap by creating a structured environment for knowledge transfer between experienced AI safety professionals and aspiring contributors, while fostering a vibrant community ecosystem. The platform will employ a sophisticated matching algorithm for mentorship that considers domain-specific expertise areas, career trajectories, and mentorship styles to create meaningful connections. Our solution features detailed AI safety-specific profiles, showcasing research publications, technical skills, specialized course completions, and research trajectories to facilitate optimal mentor-mentee pairings. The integrated community hub enables members to join specialized groups, participate in discussions, attend events, share resources, and connect with active members across the field. By implementing this platform with BlueDot Impact's community of 4,500+ professionals across 100+ countries, we anticipate significant improvements in mentee career trajectory clarity, research direction refinement, and community integration. We propose that by formalizing the mentorship process and creating robust community spaces, all accessible globally, this platform will help democratize access to AI safety expertise while creating a pipeline for expanding the field's talent pool—a crucial factor in addressing the complex challenge of catastrophic AI risk mitigation.

Keywords: AI safety education, mentorship systems, knowledge transfer, community building, career development, expertise matching, global accessibility, research guidance, discussion forums, collaborative learning

¹ Research conducted at the Women in AI Safety Hackathon, 7–10 March 2025

1. Introduction

a. Problem Statement

AI safety education faces a significant challenge—newcomers often struggle to integrate into the field due to a lack of **personalized guidance** and **community support**. This struggle is especially pronounced for individuals from underrepresented backgrounds, populations, and/or regions. This creates three interrelated problems that hinder the field’s ability to address AI risks effectively:

1. **Knowledge transfer inefficiency:** Crucial knowledge (e.g., research directions, technical methodologies, career paths) remains siloed within established networks.
2. **Community fragmentation:** Newcomers face barriers to engaging meaningfully with existing AI safety communities. Without clear entry points or networking structures, they face difficulties in collaboration, mentorship, and career progression.
3. **Talent pipeline limitations:** The field cannot efficiently identify, nurture, and retain promising contributors. Without mentorship, career guidance, and institutional support, many talented individuals—especially those outside major AI hubs—struggle to transition from education to impactful work in the field.

To address these gaps, we propose **BlueDot Impact Connect**, an accessible mentorship and community platform designed specifically for AI safety education, knowledge sharing, and career development. The platform will **formalize mentorship**, **foster structured knowledge exchange**, and **create truly global community spaces**. In doing so, it seeks to not only democratize AI safety expertise but also cultivate an inclusive AI education ecosystem that advances human flourishing.

b. Background and Motivation

Existing AI safety education primarily focuses on formal knowledge exchange (e.g., courses, workshops, seminars, conferences). While these programs provide invaluable foundations, they can lack personalized guidance for long-term skill development and career navigation. Informal mentorship exists within research organizations but remains largely inaccessible to newcomers, particularly those outside traditional academic and industry pathways. However, research in STEM education underscores the critical role of mentorship and peer learning in improving retention, skill development, and career outcomes—especially for underrepresented groups (Zuckerman et al., 2024; Antinluoma et al., 2021). As AI development accelerates, the absence of established mentorship and community infrastructure limits the field’s ability to cultivate diverse talent and ensure AI safety knowledge is effectively adapted across global contexts.

Addressing these gaps requires moving beyond content delivery to build structured mentorship and learning communities that connect theory with practical application. Integrating formal education with relationship-based learning environments will only strengthen knowledge retention, implementation capacity, and equitable participation in AI safety. By expanding access to expertise and support networks, we can empower a broader range of people to actively shape the future of AI, ensuring it serves humanity equitably and responsibly. See *Appendix A* for an expanded version of this Background and Motivation.

c. Threat Model and Safety Implications

The development of transformative AI poses potentially catastrophic risks to humanity. According to BlueDot Impact’s educational materials, NeurIPS and ICML authors estimate a median 5% chance of human extinction-level risk from AI (*Hilton & the 80,000 Hours team, 2024*). Addressing these risks

requires both technical solutions (alignment) and governance approaches, both of which depend on a robust pipeline of skilled professionals and effective coordination mechanisms.

Moreover, research shows the Global Majority is systematically excluded from all stages of the AI lifecycle (i.e., design, development, deployment, research, and governance; Racine, 2024; Okolo, 2023). This exclusion not only creates a dangerous homogeneity in AI safety work but also introduces critical blind spots that weaken safety outcomes.

BlueDot Impact Connect targets three specific safety challenges:

1. **Knowledge dissemination:** Technical alignment and governance knowledge must be efficiently transferred to the next generation of researchers and practitioners.
2. **Diversity of approaches:** A homogeneous AI safety community risks overlooking critical perspectives and pressing safety concerns. Community-building and mentorship can help bridge this gap by incorporating diverse epistemologies, cultural frameworks, and non-Western perspectives, leading to stronger, context-aware safety solutions.
3. **Retention risk:** Without clear career pathways, guidance, and community support, promising talent may leave AI safety, reducing the field's capacity to mitigate existential and systemic risks.

By facilitating high-quality mentorship relationships and fostering dedicated community spaces, BlueDot Impact Connect strengthens the human component of AI safety work. In doing so, it helps ensure that AI safety research, governance, and alignment efforts benefit from a more diverse, well-supported, and globally representative talent pool.

2. Methods

a. Approach

Our proposed methodology centers on four principles that we have identified as key for fostering meaningful connections and addressing the above threat model.

1. **Domain specificity:** A platform designed specifically for AI safety, incorporating field-specific terminology, expertise areas, and career pathways.
2. **Worldwide accessibility:** A platform that not only serves BlueDot Impact's community of 4,500+ professionals from 100+ countries, but expands access and cultivates a truly global AI safety community.
3. **Structured matching:** A sophisticated matching algorithm for mentorship that considers multiple factors beyond basic interests.
4. **Community integration:** Interconnected community spaces that align with the field's key focus areas and learning needs.

With these in mind, we have developed two key components for BlueDot Impact Connect—the **Mentorship System** and the **Community Hub**.

Mentorship System

These factors are weighed by the matching algorithm to produce robust compatibility scores. For instance, mentees' skill development needs are matched with mentors' areas of expertise while accounting for the relative scarcity of mentor capacity in certain specializations.

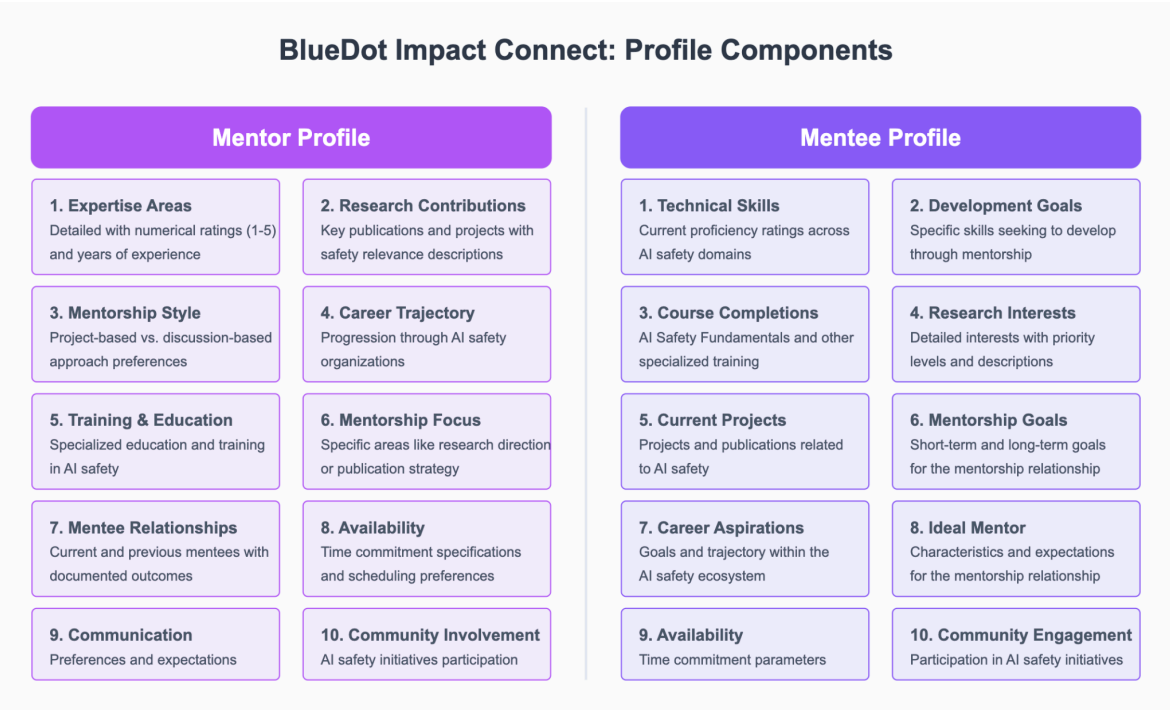


Figure 1. BlueDot Impact Connect Profile Components

Community Hub

The Community Hub comprises five interconnected elements.

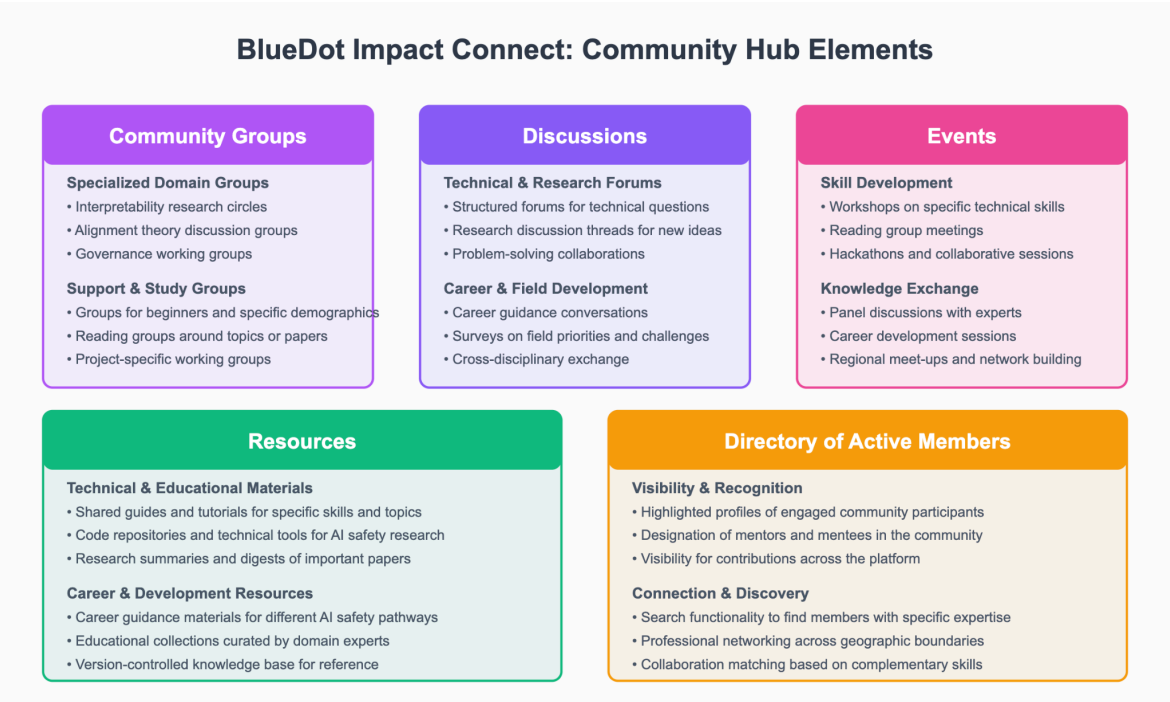


Figure 2. BlueDot Impact Connect Community

Apart Research Hackathon Submission

b. Implementation

We propose to implement BlueDot Impact Connect as a responsive web application consisting of the following architecture.

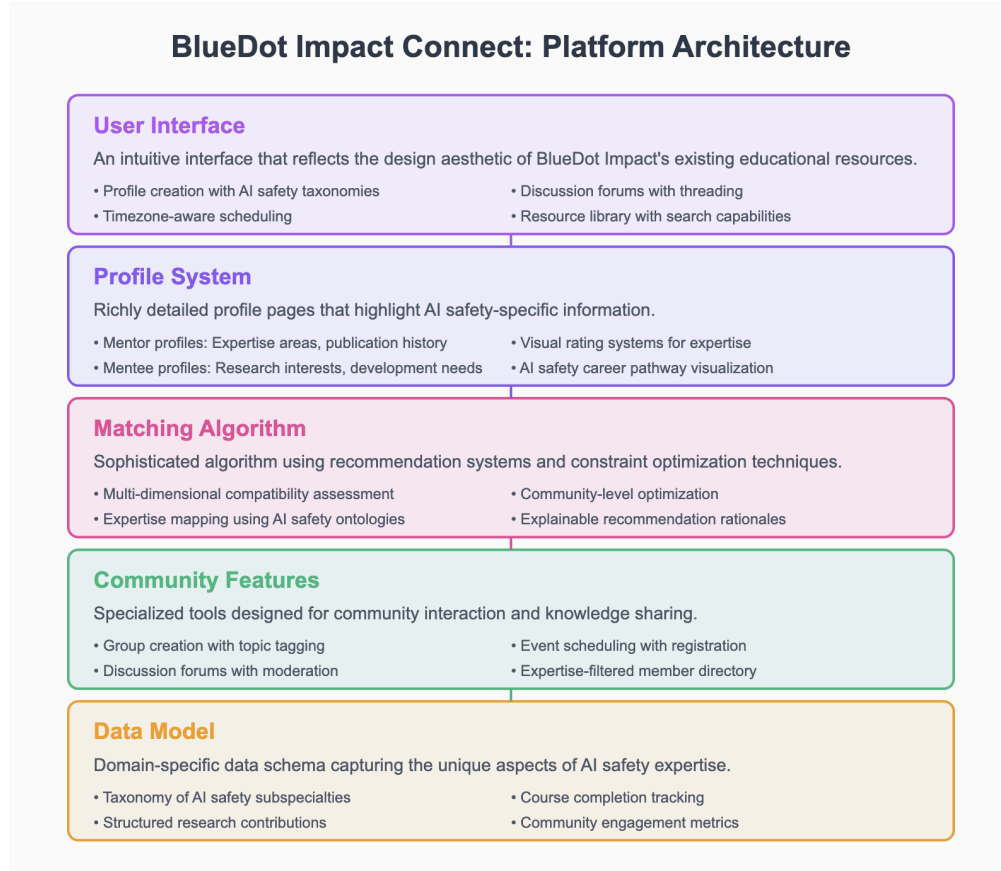


Figure 3. BlueDot Impact Connect Platform Architecture

This technical stack will prioritize reliability, security, and accessibility, using proven technologies that balance innovation with stability. The platform will be designed to scale with the growing AI safety community, with particular attention to international / diverse user needs. See *Appendix B* for screenshots.

The comprehensive profile system and technical self-assessment framework serve as the backbone of the platform. This structured approach reduces the inefficiencies of informal networking and enables precise matching between mentors and mentees based on the profile components identified in *Figure 1*. The self-assessment survey helps mentees evaluate their technical and research competencies, providing mentors with a standardized framework to understand mentee capabilities and needs. Due to limited time, we were only able to create a sample self-assessment for AI Safety Research and Engineering, but we plan to develop similar assessments for AI Governance, Philosophy, Economics, Journalism, and other fields relevant to AI Safety. These additional self-assessments will help expand mentorship opportunities beyond technical research, ensuring that individuals from diverse disciplines receive structured guidance and career development support. See *Appendix C* for the detailed mentor and mentee profile questionnaires and *Appendix D* for the comprehensive self-assessment survey example that evaluates AI Safety Research and Engineering competencies at beginner, intermediate, and advanced levels.

3. Results

a. Platform Development

BlueDot Impact Connect is currently in the design and early development phase. We have established the foundational architecture and core components necessary for addressing the identified challenges in AI safety education. Our user interface design reflects BlueDot Impact's existing educational resources aesthetic, creating a seamless experience that integrates with current offerings while providing the specialized functionality needed for effective mentorship and community building.

b. Implementation Plan

Development will follow an iterative approach with regular user testing and feedback cycles to ensure the platform meets the specific needs of the AI safety community. The implementation process will begin with initial prototype testing involving a small group of mentors, mentees, and community members selected to represent diverse backgrounds and expertise levels. This will be followed by more structured usability studies with various user segments to identify potential barriers to adoption and areas for improvement. Throughout this process, user feedback will be systematically collected through targeted surveys designed for different engagement stages—including initial registration, one-week check-ins, and in-depth interviews with active community members. See *Appendix D* for the complete set of survey instruments developed for these different engagement touchpoints.

Features will be deployed incrementally based on feedback priority, allowing for adaptation to user needs before full-scale implementation. From prototype to full deployment, continuous evaluation against key performance metrics will guide development decisions and ensure the platform effectively addresses the knowledge transfer, community fragmentation, and talent pipeline challenges identified in our problem statement.

c. Impact Assessment

The impact of BlueDot Impact Connect will be assessed through a combination of quantitative and qualitative measures focused on three key dimensions: knowledge transfer effectiveness, community integration, and talent pipeline development. Quantitative metrics will include mentorship relationship duration and satisfaction ratings, user engagement patterns across platform features, and career progression tracking for mentees. Qualitative assessment will involve in-depth interviews with participants to capture improvements in research direction clarity, sense of belonging within the AI safety community, and access to previously unavailable expertise or opportunities.

4. Discussion

BlueDot Impact Connect directly addresses the three key challenges identified in our problem statement. While existing AI safety education provides strong foundations, mentorship and community-building remain essential for deeper engagement and long-term development. To enhance knowledge transfer, the platform's sophisticated matching algorithm and detailed AI safety-specific profiles create structured pathways for sharing insights that might otherwise remain siloed within established networks. The mentor-mentee relationship enables personalized guidance on research methodologies, career navigation, and field-specific expertise, complementing formal education with individualized support. To strengthen community integration, the platform's Community Hub—with specialized groups, discussion forums, and events—offers clear entry points for newcomers and sustained engagement opportunities for practitioners across diverse geographic regions. To expand the talent pipeline, mentorship and peer learning provide structured support

for identifying and nurturing promising contributors, ensuring they have the guidance and resources needed to advance in the field.

But while the platform presents a promising approach to addressing these challenges, several considerations must be taken into account. First, the platform's effectiveness depends on securing sufficient mentor participation, particularly from those with expertise in high-demand areas. Mentor availability may vary, potentially creating constraints in knowledge-sharing. Second, technological barriers such as internet access, device availability, and digital literacy may limit participation in some regions, requiring thoughtful design to ensure broad accessibility and avoid reinforcing existing disparities in AI safety education. Third, cultural differences in mentorship expectations and communication styles may also shape relationship dynamics, emphasizing the need for careful refinement of the matching algorithm and guidance materials to support effective cross-cultural collaboration.

Thus, in recognition of these considerations and to maximize the platform's impact, we propose several enhancements for future development. A tiered mentorship model, where intermediate practitioners mentor beginners while receiving guidance from senior experts, would help expand mentorship capacity and create a sustainable pipeline for knowledge exchange. Developing offline-capable features and low-bandwidth alternatives would improve accessibility for individuals in regions with infrastructure constraints, ensuring equitable participation. Providing cultural sensitivity training for mentors and developing region-specific mentorship resources could help facilitate more effective and inclusive mentorship experiences. Additionally, incentive structures such as professional development opportunities and public recognition of mentor contributions / mentee achievements could encourage long-term mentor participation and sustain community involvement.

By expanding access to personalized guidance and community support beyond traditional academic and industry pathways, the platform helps democratize participation in shaping AI's future. This broader participation is crucial for addressing catastrophic risks, as it brings diverse perspectives to identifying potential failure modes and developing culturally appropriate safeguards. A globally inclusive approach to AI safety ensures that solutions are not only technically robust but also contextually relevant, accounting for diverse societal needs and governance structures. By strengthening mentorship and fostering collaboration, BlueDot Impact Connect helps build the human infrastructure necessary to support both technical and policy advancements in AI safety.

5. Conclusion

Ultimately, BlueDot Impact Connect represents a valuable step toward ensuring that AI safety work benefits from diverse expertise and sustained engagement. As AI development accelerates, the need for well-supported, globally connected talent is more urgent than ever. Structured mentorship improves knowledge transfer, making knowledge and research more accessible beyond academic and industry silos. It also strengthens community integration, offering clear entry points for newcomers, especially underrepresented groups. By formalizing knowledge transfer pathways, strengthening professional networks, and supporting career development, the platform contributes to a stronger, more resilient AI safety ecosystem. Through careful implementation and continuous evaluation, BlueDot Impact Connect has the potential to significantly enhance talent development and ensure that AI safety efforts reflect a wide range of perspectives. In doing so, it could be an important step forward in safeguarding humanity's future.

6. References

- Antinluoma, M., Ilomäki, L., and Toom, A. (2021). Practices of Professional Learning Communities. *Frontiers in Education*, 6, 617613. <https://doi.org/10.3389/feduc.2021.617613>
- Hilton, B. and the 80,000 Hours team. (2024, July). Preventing an AI-related catastrophe. *80,000 Hours*. <https://80000hours.org/problem-profiles/artificial-intelligence/>
- Gillespie, C. H. (2024, December 27). *Navigating the Global-Local Divide in Education*. <https://ren-network.com/navigating-the-global-local-divide-in-education/>
- Okolo, C. T. (2023, November 1). AI in the Global South: Opportunities and challenges towards more inclusive governance. Brookings.
- Racine, E. (2024). *Reparative Algorithmic Impact Assessments: A Decolonial, Justice-Oriented Accountability Framework for AI and the Global Majority*. In L. Belli & W. B. Gaspar (Eds.), *AI from the Global Majority* (pp. 137–149). United Nations Internet Governance Forum, Data and Artificial Intelligence Governance Coalition.
- Zuckerman, A. L., Lo, S. M., & Juavinett, A. L. (2024). Mentorship for transfer student success in STEM research: Mentor approaches and reflections. *CBE—Life Sciences Education*, 23(2). <https://doi.org/10.1187/cbe.23-08-0156>

7. Appendix

Appendix A: Expanded Background and Motivation

Existing approaches to AI safety education have primarily focused on content delivery through structured courses like BlueDot Impact's AI Safety Fundamentals suite (e.g., Intro to Transformative AI, Economics of Transformative AI, AI Alignment, AI Governance). While these provide extremely valuable foundations, they can lack the personalized guidance necessary for more long-term, advanced skill development and career navigation, as well as the community infrastructure needed for ongoing engagement and collaboration beyond the formal parameters of the course. Informal mentorship does occur within established research organizations, but these relationships are largely inaccessible to newcomers, especially those outside traditional academic and industry pathways.

Studies in STEM education consistently show that mentorship significantly improves retention, skills development, and career outcomes—especially for underrepresented groups (Zuckerman et al., 2024). Similarly, research on learning communities shows that peer-to-peer interaction and community engagement are critical for knowledge retention and practical application (Antinluoma et al., 2021). As the AI safety field grapples with the need to rapidly expand its talent pool in response to accelerating AI development timelines—and the potentially catastrophic risks that accompany this—formalized mentorship and robust community infrastructure represent critical yet underdeveloped components of the field's education ecosystem.

To address these challenges comprehensively, educational initiatives must expand beyond content delivery to create implementation pathways across different global regions. As Okolo (2023) notes, concerns about AI's negative impacts "have primarily been concentrated on the West, excluding perspectives on how AI may affect countries in the Global South." This not only creates significant blindspots in safety work, but requires developing frameworks that help learners adapt knowledge to diverse cultural, economic, and technological contexts. Contemporary educational theory recognizes that effective knowledge transfer requires more than information access; it demands structured support systems that bridge the gap between abstract principles and practical application in specific regional environments (Gillespie, 2024).

The creation of structured mentorship and community platforms specifically designed for AI safety education represents a critical next step in the field's development. These structures are not merely

supplementary resources but essential components of a comprehensive educational ecosystem that facilitates tacit knowledge development, professional identity formation, and contextual application of concepts. By integrating formal educational content with relationship-based learning environments, we can enhance both knowledge retention and implementation capacity across global contexts.

In doing so, we not only build technical capacity but create pathways for meaningful participation in shaping humanity's technological future, regardless of geographic location. This approach recognizes that true safety—and, consequently, human flourishing—requires both technical expertise and the cultivation of human potential in all its diversity, ensuring that those most likely to be affected by advanced AI systems have agency in determining how these systems evolve.

Appendix B: Screenshots of the Platform *(see following pages)*

Appendix C: Mentorship Program Matching Profiles

Data Collection for Mentor Profile

1. Expertise Areas (1-5 Rating)

- Which AI Safety domains do you specialize in? (e.g., interpretability, adversarial robustness, governance)
- Rate your expertise in these areas (1 = beginner, 5 = deep expertise).
- Are you open to mentoring outside your primary expertise areas?

2. Research Contributions

- Could you list your key publications, projects, or research contributions in AI safety?
- How do you see your work contributing to the broader AI safety ecosystem?
- Do you prefer mentoring mentees on research-related work (e.g., paper development, grant writing)?

3. Mentorship Style

- How do you typically mentor? (Check all that apply)
 - Project-based mentorship
 - Discussion-based mentorship
 - Career guidance
 - Open-ended brainstorming
- How structured do you like your mentorship? (Scale of 1-5, from flexible to highly structured)

4. Career Trajectory

- What has been your career progression in AI safety? (Feel free to list organizations and roles)
- What were the most valuable experiences that helped you get to where you are now?

5. Training & Education

- What specialized training or education in AI safety have you completed?
- Have you taught any AI safety-related courses or workshops? If so, which ones?

6. Mentorship Focus

- What specific areas do you want to focus on while mentoring? (e.g., research direction, publication strategy, technical skills)
- What level of mentees do you prefer working with?
 - Undergraduate
 - Master's
 - PhD
 - Independent researchers

7. Past & Current Mentees

- Have you mentored AI safety mentees before? If yes, how many?

Apart Research Hackathon Submission

- Any notable outcomes from past mentees? (e.g., published research, career transitions)

8. Availability

- How often can you meet with a mentee?
 - Weekly
 - Biweekly
 - Monthly
- Preferred meeting formats:
 - Virtual
 - In-person (if local)
 - Async support (Slack, email, etc.)
- Would you like to share a calendar link for scheduling?

9. Communication Preferences

- How do you prefer to communicate with your mentee?
 - Real-time discussions (calls, video chats)
 - Async feedback (email, Slack)
 - Structured check-ins
- What expectations do you have for mentee communication?

10. Community Involvement

- What AI safety initiatives, workshops, or conferences have you participated in or led?
- Are you interested in group mentorship sessions with multiple mentees?

Data Collection for Mentee Profile

1. Technical Skills

(For technical AI safety mentees, self-assess using AI Safety Research and Engineering Self-Assessment at the end of the Mentee Profile Questions section.)

- What's your current proficiency in AI safety research and engineering? (Appendix for detailed levels)
- For non-technical AI safety (governance, policy), what skills are you developing? *(We will refine this section later.)*

2. Development Goals

- What specific skills or knowledge are you hoping to gain from this mentorship?
- Are you looking for guidance on:
 - Research skills
 - Career trajectory
 - AI governance/policy insights
 - Technical AI safety techniques

3. Course Completions

- Which AI safety courses, fellowships, or training programs have you completed?

Apart Research Hackathon Submission

- AI Safety Fundamentals
- MATS
- ARENA
- Other: _____

4. Research Interests

- What AI safety topics interest you most? (List 2-3 with a brief explanation.)
- If applicable, rate your priority areas from 1 (low) to 5 (high).

5. Current Projects

- Are you currently working on AI safety projects, research, or publications?
- Would you like a mentor who can provide direct feedback on your work?

6. Mentorship Goals

- What are your short-term goals for this mentorship?
- Long-term goals?

7. Career Aspirations

- What's your ideal career trajectory within AI safety?
 - Research
 - Policy
 - Governance
 - Technical engineering
 - Uncertain, still exploring

8. Ideal Mentor

- What characteristics do you value in a mentor?
 - Hands-on and practical
 - Theoretical and conceptual
 - Career-focused
 - Flexible and open-ended
- Do you prefer a mentor with similar or different perspectives from your own?

9. Availability

- How frequently can you commit to mentorship meetings?
- Preferred time zones and scheduling constraints?

10. Community Engagement

- Are you involved in AI safety initiatives, discussion groups, or communities?
- Would you be interested in joining AI safety community events?

Appendix D: AI Safety Research and Engineering Self-Assessment Example

How far along are you in your AI Safety journey?

This assessment helps gauge your current skill level in AI Safety Research and Engineering and suggests the next steps to level up.

Level 1: Beginner (Exploring AI Safety & Building Technical Foundations)

Knowledge & Understanding

- ☐ I understand fundamental AI safety risks (e.g., goal misspecification, inner/outer alignment, adversarial robustness).
- ☐ I can summarize key concepts like reward hacking, corrigibility, and interpretability.
- ☐ I've completed at least one AI safety course, reading group, or fellowship (e.g., AI Safety Fundamentals, ARENA, AGI Safety Fundamentals).
- ☐ I can explain why AI interpretability matters and describe basic interpretability techniques.
- ☐ I have read classic AI safety papers (e.g., *Concrete Problems in AI Safety*, *AGI Safety from First Principles*).

How confident are you in your AI safety understanding?

(1 = Just starting, 2 = Read some materials, 3 = Comfortable discussing core ideas)

- ☐ 1 - I'm just starting; everything is new to me.
 - ☐ 2 - I've read some materials but haven't applied them.
 - ☐ 3 - I feel comfortable discussing AI safety concepts with others.
-

Technical Skills

- ☐ I can write basic Python scripts and solve algorithmic problems.
- ☐ I'm comfortable using libraries like NumPy, Pandas, Matplotlib.
- ☐ I understand machine learning fundamentals (e.g., gradient descent, loss functions, overfitting).
- ☐ I've trained simple ML models using Scikit-learn, TensorFlow, or PyTorch.
- ☐ I've completed at least one ML course (e.g., Andrew Ng's Coursera, fast.ai, Google's ML Crash Course).

Practical Challenge: Can you implement a logistic regression model from scratch in NumPy?

- ☐ Yes, confidently
- ☐ Maybe, but I'd need guidance
- ☐ No, not yet

Beginner Level Score:

0-5 → Focus on fundamentals.

Apart Research Hackathon Submission

Next Steps for Beginners:

- Take an ML engineering course and practice coding more.
- Join an AI Safety reading group or discussion forum.
- Apply for an AI Safety program (e.g., AGI Safety Fundamentals, BlueDot Impact)

6-10 → Move to intermediate projects.

Level 2: Intermediate (Building Research & Engineering Skills)

Knowledge & Understanding

- ☐ I understand different AI alignment strategies (e.g., iterated amplification, constitutional AI, corrigibility).
- ☐ I can compare safety metrics (e.g., KL divergence, robustness testing, interpretability scores).
- ☐ I am familiar with reinforcement learning and its connection to AI alignment.
- ☐ I understand how RLHF (Reinforcement Learning from Human Feedback) works.
- ☐ I have critically evaluated and debated different AI safety research directions.

Conceptual Challenge: What are the trade-offs between prosaic alignment methods and theoretical alignment proposals?

- ☐ I can explain this confidently.
 - ☐ I sort of understand but need more clarity.
 - ☐ I don't know yet.
-

Technical Skills

- ☐ I have implemented deep learning models using PyTorch or TensorFlow.
- ☐ I can train and fine-tune models (e.g., transformers, diffusion models, RL agents).
- ☐ I've experimented with interpretability techniques (e.g., feature visualization, activation patching).
- ☐ I've implemented adversarial robustness tests (e.g., FGSM, PGD attacks).
- ☐ I can optimize ML models through hyperparameter tuning, model pruning, or quantization.

Hands-on Challenge: Have you fine-tuned a GPT-style language model?

- ☐ Yes, I've done this before.
 - ☐ I understand the process but haven't implemented it yet.
 - ☐ No, I haven't tried.
-

Research & Practical Experience

- ☐ I have contributed to an AI Safety research project or open-source repository.

Apart Research Hackathon Submission

- ☐ I have replicated experiments from an AI Safety research paper.
- ☐ I have presented AI Safety findings at a workshop or discussion group.
- ☐ I have written technical summaries of at least three research papers.
- ☐ I am actively engaged in AI Safety research communities.

Intermediate Level Score:

0-5 → Focus on hands-on projects.

Next Steps for Intermediate Level:

- Apply for an AI Safety research fellowship (MATS, SERI MATS, ARENA).
- Start contributing to AI Safety open-source projects.
- Write a research proposal on an AI Safety topic.

6-10 → You're research-ready, move on to advanced!

Level 3: Advanced (Contributing to Cutting-Edge Research)

Knowledge & Understanding

- ☐ I can explain and critique state-of-the-art AI Safety papers.
- ☐ I understand mechanistic interpretability for large models.
- ☐ I have worked with AI governance/policy experts to bridge technical and regulatory efforts.
- ☐ I engage in high-level AI alignment debates.
- ☐ I can describe failure modes of large AI models and mitigation strategies.

Technical Skills

- ☐ I have developed and tested novel safety interventions in large ML models.
- ☐ I can implement and modify reinforcement learning architectures for AI alignment.
- ☐ I have built tools for monitoring AI model safety in deployment.
- ☐ I am comfortable working with high-performance computing clusters.

Hands-on Challenge: Have you created an AI Safety benchmark or evaluation dataset?

- ☐ Yes, I've designed one myself.
- ☐ I have contributed to an existing dataset.
- ☐ No, but I'd like to.

Research & Career Progress

- ☐ I have published AI Safety research in journals, arXiv, or top conferences.
- ☐ I have collaborated with leading AI Safety researchers.

Apart Research Hackathon Submission

- ☐ I have mentored junior AI Safety researchers.
- ☐ I have secured funding for AI Safety research.

Advanced Level Score:

0-5 → Strengthen leadership & mentorship skills.

Next Steps for Advanced Level:

- Apply for research grants.
- Scale up an AI Safety research project.
- Lead an AI Safety research team.

6-10 → You're ready to lead AI Safety research independently!

Appendix D: BlueDot Impact Connect Surveys:

Surveys were created for different stages of user engagement: first-time users, registration and week-long community members.

Website survey

To pop up at the corner at the site asking:

“We’re trying to make this site more useful and would really love your input. Would you mind answering 3 (very short) questions?”

☐ Yep, I’m happy to help

1. What best describes you?:
 - a. Student; academic; AI professional; policy maker; business professional; other
2. How did you find out about us?
3. Is there anything holding you back from taking action today?
☐ Yes (click to type details)
☐ No, I’m ready to take action right now

Registration form

BlueDot Impact Connect - Join the Community!

Welcome to BlueDot Impact Connect! Our community is dedicated to fostering a global network of individuals passionate about AI safety. By joining us, you’ll gain access to valuable resources, mentorship opportunities, and collaborative projects that aim to make AI technologies safe and beneficial for all. Together, we can contribute to shaping the future of AI through shared knowledge and expertise.

1. About You

- **Full Name:** (First, Last)
- **Email Address:**
- **Profile URL:** Provide a link for your LinkedIn profile or your CV. We prefer LinkedIn.
- **Links (Optional):**
 - **Your CV:**
 - **Personal Website/Portfolio:**
 - **GitHub/Code Repository:**
 - **Other Relevant Links:**
- **Profile Picture (Optional):** Help us put a face to your name!
- **Location:** (City, Country)
- **What is the closest match to your field(s) of study? Select up to three, including past or present studies.**
 - Computer Science, Engineering, Policy Studies, Social Sciences, Ethics, Mathematics, Other (specify)]
- **I am applying as a:**
 - Mentor, Mentee, Both (Learn more about being a BlueDot Impact Connect mentor here. If you check both boxes, we will consider you first as a mentor. If we can't invite you to facilitate this time, we will consider you as a mentee.)

2. Your Skills and Motivation

- How do you expect this platform will help you contribute to AI safety?
 - Please specify:
 - Concrete steps you'll take within the community
 - How participating will help you grow
 - How this will contribute to AI safety
 - It's fine not to have everything fully planned out or to be uncertain if you want to work in AI safety! However, we do expect you to explain the paths you're considering and why.
- To what extent are you currently engaged with AI Safety?
 - Not at all, Somewhat engaged, Actively engaged
- Please briefly describe how you have engaged with AI safety. (50-100 words)
 - This could include things like:
 - Projects you've worked on
 - Blog posts you've written
 - Resources you've read or watched
 - Events you've organized or attended
- What skills have you developed that could be used to advance AI safety? This question allows us to place you in a cohort at your level of technical expertise. You could tell us about:
 - Programming projects or ML courses you've done (at university or self-taught)
 - Software engineering skills or general STEM expertise
 - Government, think tank or similar work related to policy making, especially on technology

3. For Mentors: (Appears IF "Mentor" or "Both" is selected above)

- **I can offer mentorship in the following areas:**
 - Technical AI Safety (specify topics)
 - Mechanistic Interpretability, AI Control and Subversion, etc.
 - Career Advice
 - Research Guidance
 - Policy & Strategy
 - Communication & Outreach
 - Community Building
 - Other (specify)
- **My preferred style of mentorship is:**
 - 1-on-1 Sessions
 - Group Mentoring
 - Project-Based Guidance
 - Informal Advice via Chat/Forums
- **Briefly describe your mentorship experience:**

4. For Mentees: (Appears IF "Mentee" or "Both" is selected above)

- **My primary goals for mentorship are: Checkboxes:**
 - Career Advancement
 - Skill Development
 - Research, writing policy memos, improving coding skills, etc.
 - Research Opportunities
 - Networking
 - General Guidance

- Other (specify)
- What specific challenges are you hoping a mentor can help you with?

5. Community & Connection

- **I'm interested in participating in:**
 - Online Forums & Discussions
 - Local Meetups (if available)
 - Collaborative Projects
 - Workshops & Webinars
 - Informal Social Events
- **I can contribute to the community by:**
 - Sharing my Expertise
 - Organizing Events
 - Moderating Discussions
 - Providing Feedback
 - Actively Participating in Discussions
 - Other (specify)
- **Preferred Language(s) for Communication:**
 - English
 - Spanish
 - French
 - Mandarin
 - Other(specify)

6. Time & Availability

Time Zone: Dropdown: List of time zones

Availability for Meetings:

- Weekday
 - Mornings
 - Afternoons
 - Evening
- Weekends(specify preferred times)

7. Final Steps

- How did you discover BlueDot Impact Connect?
 - Website, Social Media, Referral, Event, Other]
- Do you have any feedback on this application form? If you need a response from us, please contact us.
- Community Guidelines Agreement:
 - ☐ I agree to abide by the BlueDot Impact Connect Community Guidelines, which promote inclusivity, respect, and constructive engagement.
 - ☐ Can we share your data with third-party AI safety organisations? If you opt-in, we may share parts of your application with organisations we trust and think are making positive contributions to the field. These organisations sometimes email people with jobs or other opportunities that could serve as good next steps after this course. We will only share this data if you give us your consent below; it will not affect your application decision. You can opt-out at any time. Checkbox: Yes – I consent | No – I don't consent]

- ☐ Diversity data consent: We value the contributions of a diverse range of voices towards making AI safe and beneficial for all. You can optionally provide the following data to help us ensure we are getting this right. If you opt-in, we'll collect and process information about you for diversity, equality and inclusion monitoring. We won't share your specific data outside BlueDot Impact. We may share aggregated anonymous data publicly (e.g., x% of our applicants were female). Checkbox: Yes – I consent | No – I don't consent]
- ☐ By submitting this form, you consent to our Privacy Policy and authorize the use of your personal data in accordance with it. You'll receive an email confirming we've received your application.

Email templates

Confirmation Message:

"Thank you for applying to BlueDot Impact Connect! We're excited you're joining us. We'll review your application and be in touch within [timeframe]. In the meantime, feel free to explore our resources and community forum." "Thank you for applying to BlueDot Impact Connect! We're excited you're joining us. We'll review your application and be in touch within [timeframe]. In the meantime, feel free to explore our resources and community forum."

Send out messages in-platform to users who've been engaged members for a week:

Hi (Name), BlueDot Impact Connect is supposed to be designed for (defining characteristics of your target audience that match this person's characteristics) — so your input would be a huge help in making sure we're taking the platform in the right direction.

I'd love to chat briefly to understand how you're (solving the pain point that matches this person's registration information) right now, what's working well about that, and what's not. I know your time is extremely valuable, so, if you're open to it, please let me know your availability here[link to calendly], and we can find a time that works for both of us.

Thank you for considering this, and for being such an important part of our group!

Zoom interview survey for community members who've been engaged in the platform for a week:

1. The journey to BlueDot Impact Connect

Example Script:

Take me back to the day when you first decided that you needed to (defining pain points that match this person's registration information). Not to sign up, that's not the day. We want to go back to the day when you thought it was urgent and compelling to join a community and (get mentorship/become a mentor) to (define pain points that match this person's registration information). Just tell me what happened.

'Okay, why didn't you do it sooner? What really changed that made you decide now is the time to (solve pain points that match this person's registration information).

So before Bluedot Impact Connect, how did you navigate (defining pain points that match this person's registration information)?

2. Goals: What progress do you want to make?

- Complete this sentence: "By participating here, I want to..."

3. Pains: What's holding you back?

- What's one thing that's made it harder to achieve that here? (e.g., "Mentorship matching took longer than expected")

4. Experience: How are we doing?

- What's missing that would accelerate your progress? (e.g., "Structured onboarding for technical AI safety topics")

5. Final Feedback

- If you could redesign one thing about this community, what would it be?
- Wild card question:
 - What's a non-obvious way this community could improve AI safety? (e.g., "Host debates about under-discussed risks")
 - Digging Deeper: Follow-Up Questions
 - For each of the sections above, consider adding follow-up prompts based on their responses:
 - If they mention a specific motivation or problem:
 - "Can you elaborate on why this is important to you?"
 - If they state a goal:
 - "Why is achieving this goal significant for you?"
 - If they identify a challenge:
 - "How has this challenge impacted your engagement or progress?"
 - For their final feedback:
 - "Why do you think this change is necessary?"

Appendix F: Team Member Contributions

Elise Racine created the community platform web app prototype, including identifying the critical gap, conceptualizing the matching algorithm approach for the mentorship program, crafting the Community Hub architecture, and co-developing the detailed profile systems for both mentors and mentees. She was the primary author of this research paper.

Aliya Koishin assisted with the writing of this paper, specifically the problem statement, background, and threat model sections (see Appendix A), as well as finding and citing relevant sources throughout.

Aishwarya Gurung co-developed the profile components, building on Racine's profile framework to optimize mentor-mentee matching to design detailed data collection points(Appendix C). Additionally, she created the AI Safety Research and Engineering self-assessment survey (Appendix D), establishing a standardized framework for evaluating mentee competencies across beginner, intermediate, and advanced levels.

Tatenda Mawema designed the user surveys to collect user feedback for different stages of engagement with the platform (see Appendix E), as well as gave advice on messaging for a general, non-technical audience.

Haihao Liu was the team leader, setting up communication channels and logistics for collaboration, scheduling team meetings, and guiding the overall planning and strategic vision for this project.