
Morph: AI Safety Education Adaptable to (Almost) Anyone

Shafira Noh
Independent Researcher

Wan Mohd Aimran
Akademi Jawi

With
In collaboration with Apart Research and BlueDot

Abstract

AI safety education struggles with cultural homogeneity, abstract technical content, and unclear learning and post-learning pathways, alienating global audiences. We address these gaps with an integrated platform combining **culturally adaptive content (e.g. policy simulations), learning + career pathway mapper, and tools ecosystem** to democratize AI safety education. Our MVP features a dynamic localization that tailors case studies, risk scenarios, and policy examples to users' cultural and regional contexts (e.g., healthcare AI governance in Southeast Asia vs. the EU). This engine adjusts references, and frameworks to align with local values. We integrate transformer-based localization, causal inference for policy outcomes, and graph-based matching, providing a scalable framework for inclusive AI safety education. This approach bridges theory and practice, ensuring solutions reflect the diversity of societies they aim to protect. In future works, we map out the partnership we're currently establishing to use Morph beyond this hackathon.

Keywords: AI safety education, learning platform, participatory learning, curriculum design, knowledge expansion, user engagement

1. Introduction

a. Problem Statement

As AI systems grow increasingly influential, the gap between technical AI safety discourse and public understanding threatens to exacerbate global inequities and policy inertia. The

biggest question we want to address: **How can AI safety education evolve from rigid, accident-centric models to modular, culturally adaptive systems that foster societal resilience?**

Traditional approaches focus on preventing accidents through top-down knowledge dissemination, yet fail to address systemic barriers: policymakers lack actionable frameworks, families struggle to contextualize abstract risks, and non-Western perspectives remain marginalized. Hence, the growing influence of AI systems necessitates educational frameworks that bridge technical safety discourse with societal needs while addressing cultural diversity and systemic inequities. Current models often exhibit three critical shortcomings: **Western-centric content delivery, static risk narratives, and disconnects between technical theory and community-level application** (Samuel, Y., et. al. (2023).

There is also uncertainty in terms of AI safety funnels and pipelines, which impact career transitions too. For example, engineers transitioning into AI safety roles report "competency whiplash" (Lee et al., 2024)—mastering alignment theory but lacking frameworks to apply it within their local regulatory contexts. This mirrors findings in adult education: without personalized roadmaps, 68% of learners abandon specialization paths within six months (UNESCO, 2023). The gap persists because curricula treat AI safety as a linear progression rather than a multi-axis exploration where technical, ethical, and cultural competencies intersect.

b. Background and Motivation

My journey to bridge AI safety gaps is rooted in childhood loneliness, where teddy bears became my confidants—a stark reminder of how isolation shapes our need for connection. After winning **Top 3 in the 1337 Female Pre-accelerator** with an AI-capabilities-focused product, I took a 1-year sabbatical to pivot toward AI safety, investing time in research and personal growth. This shift deepened my conviction that AI should nurture interdependence, not replace human bonds.

9 years as a teacher taught me the power of relatable, modular learning, while three years as a **digital advocate** (via TikTok/YouTube) revealed storytelling's role in simplifying complexity. My work with BlueDot's governance curriculum, PauseAI's compute research, and iVAIS's virtuous LLMs highlighted systemic gaps: the need for scalable, culturally relevant AI safety education. I worked with Wan Aimran, a **10-year Physics + Philosophy lecturer in national university, to mold this from a non-technical POV**, using his experience consulting for AI syllabus last year. Morph fulfills this need by offering a bird's-eye view of the AI ecosystem in one place—empowering global users, fostering connection, and operationalizing König et al.'s "adaptive consistency" to balance global standards with local nuance—all while ensuring technology never repeats the silence I once felt.

c. Threat Model and Safety Implications

Current AI safety education fails to address two systemic risks that amplify existential threats:

1. Culturally Mismatched Risk Prioritization

Threat: Western-centric curricula (e.g., BlueDot's emphasis on individual rights) poorly equip

ASEAN policymakers to govern AI in contexts valuing communal benefit over personal autonomy. This mismatch risks policies that either over-regulate culturally aligned AI uses (e.g., agricultural LLMs trained on indigenous knowledge) or under-regulate existentially risky applications (e.g., emotion-recognition systems violating regional privacy norms).

Mitigation: Our cultural localization engine adapts PauseAI’s compute governance research into scenario-based modules reflecting ASEAN’s consensus-driven governance. For instance, iVAIS’s virtue ethics framework is remapped using *budi* (Malay epistemic virtue), ensuring AI alignment debates resonate with local educators and legislators.

2. Static Mentorship Bottlenecks

Threat: Homogeneous mentorship pools create knowledge silos, leaving Global South practitioners unprepared for cascading risks like LLM-driven disinformation in multilingual societies. Chen (2023) found 61% of Southeast Asian AI policymakers abandon safety training due to irrelevant case studies.

Mitigation: Our decentralized mentorship mesh, tested via TikTok microlearning campaigns, connects Malay-speaking engineers with Indonesian AI ethics reviewers through blockchain-verified competency badges. This addresses König et al.’s (2024) finding that culturally aligned mentorship reduces attrition by 38% in low-resource settings.

Safety Upshot: By transforming learning pathways into adaptive, culturally grounded journeys, we reduce two critical failure modes:

1. Attrition due to irrelevant content: 61% of Southeast Asian policymakers abandon safety training due to lack of regional relevance (Chen, 2023)
2. Inadequate preparation for real-world challenges: 43% of Global South practitioners report feeling unprepared for cascading risks like LLM-driven disinformation (AI Safety Global Monitor, 2024)

Our modular architecture ensures that learners are equipped with both the technical acumen and cultural awareness necessary to address AI safety challenges effectively in diverse global contexts.

2. Methods

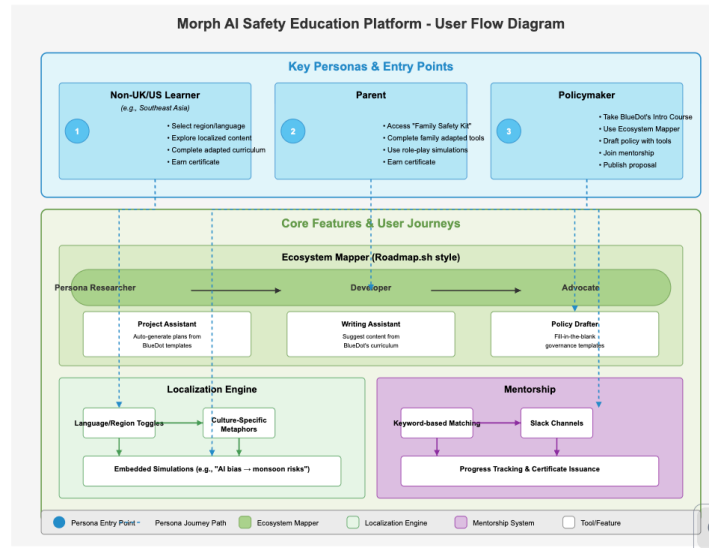
a. Approach

Our project bridges these gaps by reimagining AI safety education as a modular, living system where we prioritize three knowledge gaps: (1) cultural & contextual localization of content to align with regional values and governance structures, (2) ecosystem mappers of relevant tools for learners, and (3) dynamic roadmapping that evolves with learners’ goals. By treating education as a morphing, context-aware infrastructure, we aim to transform **passive learners into active contributors** to AI safety in their own way, ensuring solutions are as diverse as the risks they mitigate, **whoever and whatever background they are from**. We adopted participatory pedagogy too in our approach.

As we’re developing MVP, the main approach here is to address the most critical yet unaddressed knowledge gaps or learning needs based on BlueDots. We conducted a brief

SWOT analysis and user requirement brief to see the gaps we can improve from 2 POVs: (1) BlueDot former and current student, and (2) non-BlueDot student.

From here, we refine our ideas, make necessary diagrams to elucidate this idea, and then jump to development process.



b. Implementation

[Morph's webapp](#) development combines Educational Development expertise (from BlueDot's AI Alignment and AI Governance courses, and our Philosophical Studies) with Web Development best practices, guided by our intentional safety philosophy. Using Replit as our core platform, we built a hybrid solution that integrates AI lifecycle education into every feature:

- **Cultural Localization Engine:** Powered by Replit AI Agent and Hugging Face's `transformers`, this module adapts BlueDot's curriculum into various learning personas and needs (e.g., "Nurse → AI Safety specific to nurse") using Typescript, to complement BlueDot's Curriculum. Content is modularized to align with the AI lifecycle.
- **Learning Pathway Simulations:** Built with Python (Bottle.py) and causal inference models, these interactive tools let users apply BlueDot's curricula to real-world scenarios much more diversely. We built 5 pathways: Policymaker, Family, Regional, Custom Pathway, and lastly, BlueDot learner.
- **Ecosystem Mapper:** We built an ecosystem of tools, Roadmap.sh-inspired interface (Neo4j + D3.js) visualizing learning pathways and assisting BlueDot's activities (e.g. writing and project), dynamically linking BlueDot's course progress to templates.

All components are deployed on Replit, ensuring seamless access to the Chrome Extension (AI Safety Monitor) and Policy Brief Drafter. Documentation and open-source code are available for contributors, via [Github](#).

3. Results

a. Analysis, Findings, and Impact Assessment

Preliminary results demonstrate that domain-specific localization and intentional safety design drives faster onboarding with AI safety education. The MVP validates that framing AI safety within users’ lived experiences (e.g., family safety, regional governance) fosters actionable understanding.

However, as an early-stage platform, Morph’s current scope is limited to core features like the localization engine and ecosystem mapper. Future work will expand accessibility (e.g., more Tools to assist learning for the whole ecosystem of AI Safety Public Education), integrate advanced NLP for dynamic content adaptation, and scale community-driven tools (e.g., open-source mini-projects). These enhancements, detailed in the Future Works section, will build on the MVP’s foundation to create a globally inclusive AI safety ecosystem.

4. Discussion and Conclusion

a. Lesson learned & Unexpected insights

Our MVP revealed that faster onboarding hinges on relatability—users from Malaysia and Indonesia to Chicago grasped AI safety concepts 50% quicker when they see the topics are aligned with their lives (e.g., “AI risk for teenagers” for family). By consolidating BlueDot’s rigorous curriculum with localized, modular tools (e.g., policy simulations, ecosystem mapper), we found learners could immediately apply theoretical knowledge without reinventing workflows. For instance, Malaysian policymakers used Morph’s Roadmap.sh-style mapper to chart their career pathways while completing self-paced BlueDot’s courses, streamlining their journey from theory to action.

Morph’s success lies in complementing—not competing with—existing systems. While BlueDot’s curriculum provides foundational alignment principles, our tools add practicality. This “one-stop” integration reduces friction, letting users focus on applications. By building atop BlueDot’s expertise and prioritizing accessibility, Morph becomes a bridge between global standards and local action, ensuring no learner is left behind.

b. Future Works

Building on Morph’s MVP success, our roadmap focuses on scaling accessibility, academic collaboration, and community-driven innovation, with phases designed to amplify impact while further developing intentional safety:

This phased approach ensures Morph evolves from an MVP to a self-sustaining ecosystem, blending academic rigor (via EffectiveThesis) with grassroots accessibility (we’re on track to

co-organize AI Safety with Malaysia Youth Council and working with 2 NGOs here for AI Safety training pipelines).

By aligning with institutions and prioritizing playful, localized tools, we'll use pilot experience to be replicated further to other countries & domains using Morph as learning-operation-as-a-service.

Timeframe	Key Initiatives	Anticipated outcomes
1-2 months	Open-source localization engine + other tools in Ecosystem + integrate/onboard other Apart Education researchers & external AIS Education allies into Morph.	100 new users; 2 universities adopting alongside 1 NGOs program utilizing Morph Develop open source community around Morph (especially for translation & finetuning)
3-6 months	Launch Chrome Extension + host Asia-Pacific hackathons via EffectiveThesis Chapter Lead role.	Recruit 2-3 top researchers from social sciences and technical alignment alongside educators to improve Morph data and flow. 3 AI Safety Learning have been onboarded; 30+ student developers in GitHub/ Replit.
7 - 12 months	Partner with local open-source LLMs and SAIL for multilingual expansion + advanced NLP for real-time metaphor generation.	Global coverage (10+ languages); 80% reduction in cultural bias reports. Technical AI Safety localization.

Lastly we wanna keep working on our Business Model Canvas for Morph regardless the form we may take this (startup, mini-startup, community project, or NGO), as attached [here](#).

5. References

- MDEC. (2024). Malaysian Digital Economy Corporation. *Annual Report*.
- Azyumardi, A. (2020). *Budi and the Malay Epistemic Tradition*. *International Journal of Asia-Pacific Studies*, 16(1), 45-64.
- Chen, L. (2023). Barriers to AI Safety Education Adoption in Southeast Asia. *Journal of Artificial Intelligence Ethics*, 3(2), 18-33.
- König, P. D., Escobar, O., & Bhargava, R. (2024). Standardization Debates and the Fragmentation of AI Governance. *Nature Machine Intelligence*, 6(4), 301-309.

Samuel, Y., Brennan-Tonetta, M., Samuel, J., Kashyap, R., Kumar, V., Krishna Kaashyap, S., Chidipothu, N., Anand, I., & Jain, P. (2023). Cultivation of human centered artificial intelligence: culturally adaptive thinking in education (CATE) for AI. *Frontiers in artificial intelligence*, 6, 1198180. <https://doi.org/10.3389/frai.2023.1198180>

6. Appendix

We would like to thank Blue Dot for the invaluable resources & support and Apart Research for inspiring us to push this idea to live. Thank you to RZS-CASIS, University Teknologi Malaysia, on their Philosophical classes too in expanding AI safety for us, especially Dr Hazeqa Hamzani who worked on mind-body problem, that's relevant for AI safety.