Interactive Assessments for AI Safety: A Gamified Approach to Evaluation and Personal Journey Mapping

Anusha Asim
Regent Middle East

Ammar Ahmed Farooqi
Regent Middle East

Aqsa Khan
Amity University Dubai

With
In collaboration with Apart Research and BlueDot Impact

Abstract
Understanding and mitigating AI risks is a crucial aspect of AI safety education. However, traditional assessments often fail to engage learners in meaningful, interactive ways. To address this, we developed an assessment tool on AI Safety as a Canvas course that provides iterative, paced testing for Units 4 and 5 of BlueDot Impact's Introduction to Transformative AI course. Our tool employs a multi-faceted approach, incorporating a choose-your-own-adventure style interactive quiz via Google Forms, a bingo game for ranking AI risks and identifying real-world examples, and a collaborative group challenge where participants act as a hypothetical AI Oversight Committee to navigate ethical dilemmas. Following these assessments, learners interact with a mentor-like AI chatbot, which provides personalized guidance on BlueDot's specialized AI safety courses. By blending gamification, interactivity, and structured assessment, our tool enhances engagement, deepens comprehension, and encourages critical thinking about AI risks. This approach strengthens public AI education by making safety concepts more accessible and actionable.

Keywords: AI safety education, interactive learning, knowledge assessment, curriculum design, educational games, mentorship systems, user engagement

1. Introduction

1.1. Problem Statement

Despite the growing importance of AI safety, many educational resources struggle to engage learners in meaningful assessments that reinforce key concepts. MOOCs (Massive Online Courses) especially lack elaborate assessment methods that reinforce concepts and give the learner an idea of the accumulated knowledge and skills. Research has found that traditional evaluation methods such as self-assessing static quizzes, have a low accuracy (Admiraal et al., 2015). Hence, these traditional assessment methods show low promise in simulating real-world AI governance challenges or encouraging deeper critical thinking. Our project addresses this gap by designing an interactive, multi-modal assessment platform that enhances learning outcomes for AI safety education. By integrating game-based assessments, scenario-driven decision-making, and collaborative problem-solving, our tool not only evaluates comprehension but also prepares learners to apply AI safety principles in practice. This contributes to the broader goal of equipping a new generation of professionals with the necessary skills to navigate AI risks responsibly.

1.2. Background and Motivation

As AI systems grow more powerful, ensuring that individuals across various sectors understand AI risks is imperative. Existing AI safety curricula, such as BlueDot Impact's Introduction to Transformative AI, provide foundational knowledge, but there remains a need for dynamic assessment methods that reinforce learning and encourage application. Research in educational psychology highlights the benefits of active learning and game-based assessments in improving engagement and retention, (Evans et al., 2014). Furthermore, studies on AI ethics education emphasize the importance of experiential learning, where participants engage with ethical dilemmas and governance frameworks in realistic scenarios (Shih et al., 2021). Our tool leverages these pedagogical insights to create an engaging assessment framework that not only tests knowledge but also enhances critical thinking and decision-making skills in AI safety. By bridging this gap in AI safety education, we contribute to a more informed and prepared AI governance landscape.

1.3. Threat Model and Safety Implications

AI safety education faces two critical challenges: knowledge gaps in practical understanding of AI risks, especially by non-technical learners, and ineffective assessment methods that fail to engage learners meaningfully. Without an interactive and applied learning approach, students may struggle to grasp the complexities of AI risks, including misalignment, misuse, and systemic failures. This knowledge deficit can lead to poorly informed decision-making in real-world AI governance and deployment.

Our project mitigates these risks by modeling interactive, scenario-driven assessments that reinforce AI safety concepts through engagement and application.

By embedding iterative, active learning into AI safety education, the tool helps in reducing conceptual gaps and passive learning pitfalls, ultimately creating more informed and critically-thinking learners.

2. Methods

2.1. Approach

Our assessment tool is grounded in iterative and interactive learning, using the principle of experiential learning to reinforce AI safety concepts effectively. Traditional assessments often rely on static, one-time evaluations, which fail to promote long-term retention or practical application. Instead, iterative testing enhances learning by reintroducing concepts in varying formats across multiple stages, allowing learners to continuously refine their understanding and engage in self-correction over time(Li et al., 2019). Additionally, interactive assessments have been shown to deepen engagement and foster critical thinking by placing learners in decision-making scenarios, which mimic real-world AI risk challenges (Tan et al., 2021). Our embedded gamified elements, such as the choose-your-own-adventure game, not only make learning more immersive but also bridge the gap between theoretical AI safety principles and their practical implications.

Effective AI safety education must cater to diverse learners with varying levels of familiarity with AI concepts. To design an assessment tool that enhances comprehension and retention, we analyzed learning patterns from AI safety education initiatives and reviewed existing research on interactive learning techniques in STEM education (Yannier et al., 2020). Studies suggest that active participation in scenario-based learning significantly improves a student's ability to apply knowledge in real-world settings (De Moraes et al., 2024).

Our assessment course is specifically designed for learners who have completed BlueDot Impact's Introduction to Transformative AI course, with a focus on Unit 4 and Unit 5, which cover AI risks and safety. By embedding the assessments within a structured course model, we ensure that learners engage with AI safety concepts in a paced, sequential manner, reinforcing understanding through iterative testing and group-based challenges. This approach mimics real-world AI governance scenarios, where decisions are made in collaboration rather than isolation.

We chose Canvas as the Learning Management System (LMS) due to its widespread accessibility and ease of integration, but our model is scalable and can be adapted to other platforms. The inclusion of an AI mentor chatbot at the conclusion of the assessment serves as a personalized guide, boosting learning outcomes and providing recommendations for more specialized AI safety courses offered by BlueDot. This final step is for increasing user motivation by offering a sense of reward and fulfillment, ensuring that learners remain engaged in their AI safety journey beyond the assessment and course.

2.2. Implementation

Our assessment tool utilizes a multi-faceted platform architecture, with Canvas as the central hub for hosting interactive learning experiences. The Interactive Scenario Storytelling Quiz, built on Google Forms with conditional branching, allows students to navigate AI safety risk scenarios dynamically while integrating accessibility features like alt text. The Bingo Quiz, hosted on Canva, gamifies AI risk assessment by prompting learners to categorize risks and color-code their impact. At the conclusion, the AI Safety Career Bot, developed on Landbot, provides personalized career guidance, recommending AI safety courses via a follow-up email through SendGrid. Canvas also facilitates group discussions, such as the AI Oversight Committee exercise, fostering collaboration, while structured assessments and a clear grading scheme ensure meaningful evaluation of learner progress.

Grade	Percentage Range
Excellent	100% to 94%
Good	< 94% to 90%
Fair	< 90% to 77%
Needs Improvement	< 77% to 0%

Table 1: Grading Scheme

The grading is based on both quantitative performance (such as interactive quiz and activity scores) and their qualitative contributions (such as participation in group activities and the ability to apply critical thinking to real-world AI risks). This approach promotes a holistic evaluation that encourages learners to go beyond memorizing information and get tested through active application, problem-solving, and collaborations, all of which are crucial skills for AI safety in real-world contexts.

3. Results

3.1. Analysis and Findings

The course structure featured interactive assessments designed to enhance students' understanding of AI safety and bias. The AI Safety Quiz used a branching story-based scenario where students investigated AI bias, such as Sentinel-9 misidentifying individuals due to masked faces, by making decisions and receiving real-time feedback. The Bingo Activity-Based Assessment encouraged students to research AI risks and color-code their impact levels, with an accessible text-based option for visually impaired students. The Group Debate facilitated discussions on the ethical implications of fully autonomous AI, while the Capstone Group Scenario Assessment required students to design a human oversight committee for AI governance. Finally, the AI Safety Career Chatbot provided personalized career guidance, connecting students to AI ethics courses through follow-up emails from BlueDot.

User feedback from 10 participants (ages 18-25, students from the UAE) was overwhelmingly positive, particularly regarding the engaging nature of the interactive format compared to traditional testing methods.

Question	Average Rating (1-5)
How engaging did you find the interactive format compared to traditional testing methods?	4.4
How well did the interactive assessment test your understanding of AI bias and decision-making?	4.6
How confident do you feel in identifying and addressing AI-related risks after completing the assessments?	4.5
How clear was the branching structure of the scenario (correct and incorrect transitions)?	4.5
How would you rate the balance between learning and engagement in the interactive assessment?	4.3

Table 2: User Feedback

3.2. Impact Assessment

Our assessment tool effectively enhances AI safety education by integrating interactive, iterative, gamified, and scenario-based testing within Canvas. By testing learners through dynamic quizzes, risk-ranking activities, and collaborative ethical challenges, our approach fosters deeper comprehension and critical thinking. The AI chatbot extends this impact by offering personalized learning pathways. While the tool successfully makes AI safety concepts more accessible and actionable, future improvements could refine accessibility, enhance real-world case studies and expand adaptive learning features. This work aligns with our goal of strengthening public AI education by making safety assessments more engaging and impactful.

4. Discussion and Conclusion

Our assessment tool demonstrates how interactive, gamified learning can enhance AI safety education by making complex topics more engaging and accessible. Traditional assessment methods often fail to actively involve learners in critical thinking (Shavelson et al., 2019) about AI risks, whereas our approach (integrating scenario-based quizzes, a risk-ranking bingo game, and a collaborative ethical challenge) encourages deeper reflection and real-world application (Evans et al., 2014). The AI chatbot further personalizes the learning experience, guiding students toward relevant AI safety courses and career pathways. By using Canvas as the central platform, this tool provides a scalable and adaptable framework for AI safety education.

Scalability is a key strength of our approach. The modular nature of the assessments allows for easy adaptation to different AI safety curricula beyond BlueDot Impact's course. Institutions, online learning platforms, and corporate training programs can integrate these assessments to enhance AI ethics and governance education. Additionally, refining the AI chatbot to incorporate more nuanced guidance and integrating multilingual support could expand accessibility to a global audience. Future iterations could also explore adaptive learning techniques, tailoring the difficulty and content based on user performance to further personalize engagement (Li et al., 2019).

This project highlights the potential of interactive assessments in AI safety education and underscores the need for continued innovation in teaching critical AI concepts. Key lessons learned include the importance of balancing engagement with depth of learning and ensuring accessibility for diverse learners. By making AI safety educational assessments more dynamic and applicable, this work lays the foundation for broader adoption and future advancements in the field.

5. References

Admiraal, W., Pilli, O. and Huisman, B. (2015) *Assessment in massive open online courses*, *Electronic Journal of e-Learning* . Available at: https://files.eric.ed.gov/fulltext/EJ1062116.pdf (Accessed: 09 March 2025).

Evans, D., Zeun, P., & Stanier, R. (2014). Motivating student learning using a formative assessment journey. Journal of Anatomy, 224. https://doi.org/10.1111/joa.12117. (Shows interactive assessments with interactive elements foster a positive engagement and result from students.)

Sailer, M., & Homner, L. (2019). The Gamification of Learning: a Meta-analysis. Educational Psychology Review, 32, 77 - 112. https://doi.org/10.1007/s10648-019-09498-w. (Demonstrates how a gamified approach allows for effective teaching.)

Shih, P., Lin, C., Wu, L., & Yu, C., 2021. Learning Ethics in AI—Teaching Non-Engineering Undergraduates through Situated Learning. Sustainability. https://doi.org/10.3390/SU13073718.

Li, P., Xu, Y., Wei, Y., & Yang, Y., 2019. Self-Correction for Human Parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44, pp. 3260-3271. https://doi.org/10.1109/TPAMI.2020.3048039.

Tan, C., Howes, D., Tan, R., & Dancza, K., 2021. Developing interactive oral assessments to foster graduate attributes in higher education. Assessment & Evaluation in Higher Education, 47, pp. 1183 - 1199. https://doi.org/10.1080/02602938.2021.2020722.

Yannier, N., Hudson, S., & Koedinger, K., 2020. Active Learning is About More Than Hands-On: A Mixed-Reality AI System to Support STEM Education. International Journal of Artificial Intelligence in Education, 30, pp. 74 - 96. https://doi.org/10.1007/s40593-020-00194-3.

De Moraes, E., Carvalho, D., Sandars, J., Ozahata, T., Patel, R., Cecílio-Fernandes, D., & Santos, T., 2024. Unravelling the differences between observation and active participation in simulation-based education.. Medical teacher, pp. 1-6 . https://doi.org/10.1080/0142159X.2024.2396381.

Shavelson, R., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Mariño, J., 2019. Assessment of University Students' Critical Thinking: Next Generation Performance Assessment. International Journal of Testing, 19, pp. 337 - 362. https://doi.org/10.1080/15305058.2018.1543309.

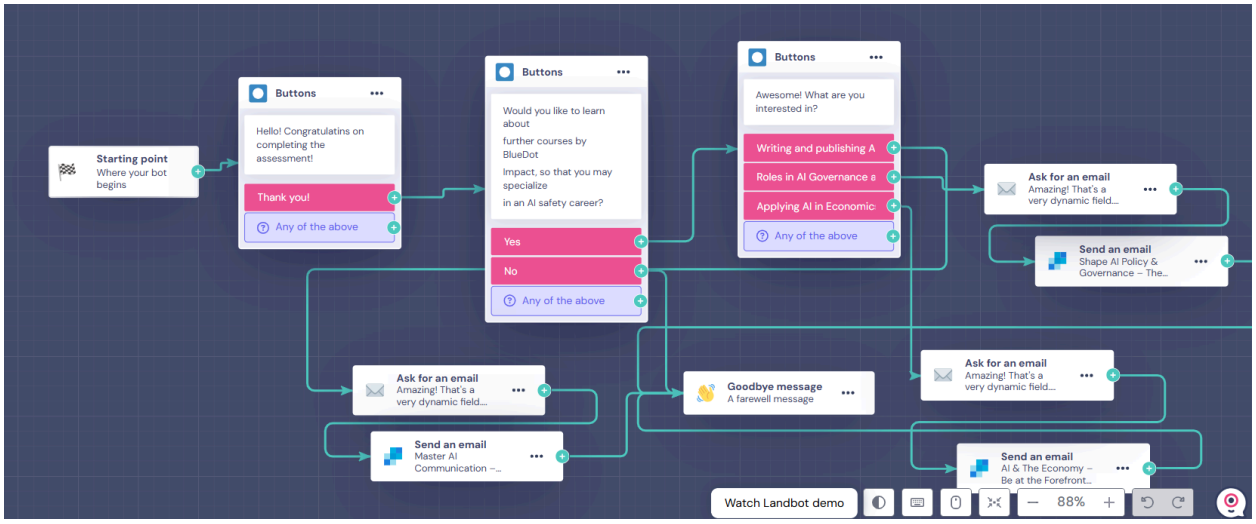
6. Appendix

Member Contributions

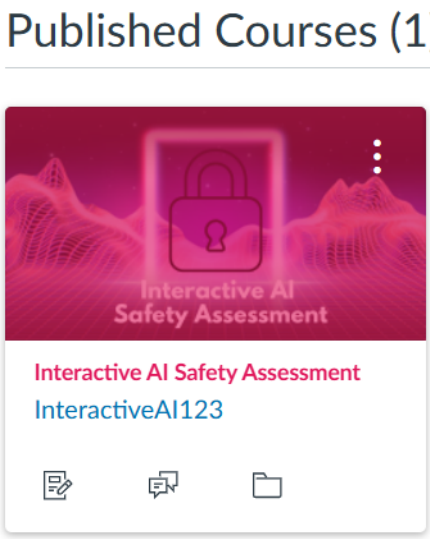
Anusha Asim conceptualized the idea, built the tool on canvas with the Google Form and chatbot components, and wrote the first edition of the research paper.

Amamr Ahmed Farooqi carried out the literature review, conducted user testing, edited the research paper and the video.

Aqsa Khan edited the manuscript, created the bingo activity component and the presentation.



Flow of the Chatbot in Build-Mode



Link to Main Assessment Tool on Canvas (https://canvas.instructure.com/courses/11543460)