AI Safety Initiative Groningen (AISIG)
BUGgy: Supporting AI Safety Education through Gamified Learning *

Sophie Sananikone AISIG
Xenia Demetriou AISIG
Mariam Ibrahim AISIG
Nienke Posthumus AISIG

With
In collaboration with Apart Research and BlueDot.

Abstract

As Artificial Intelligence (AI) development continues to proliferate, educating the wider public on AI Safety and the risks and limitations of AI increasingly gains importance. AI Safety Initiatives are being established across the world with the aim of facilitating discussion-based courses on AI Safety. However, these initiatives are located rather sparsely around the world, and not everyone has access to a group to join for the course. Online versions of such courses are selective and have limited spots, which may be an obstacle for some to join. Moreover, efforts to improve engagement and memory consolidation would be a notable addition to the course through Game-Based Learning (GBL), which has research supporting its potential in improving learning outcomes for users. Therefore, we propose a supplementary tool for BlueDot's AI Safety courses, that implements GBL to practice course content, as well as open-ended reflection questions. It was designed with principles from cognitive psychology and interface design, as well as theories for question formulation, addressing different levels of comprehension. To evaluate our prototype, we conducted user testing with cognitive walk-throughs and a questionnaire addressing different aspects of our design choices. Overall, results show that the tool is a promising way to supplement discussion-based courses in a creative and accessible way, and can be extended to other courses of similar structure. It shows potential for AI Safety courses to reach a wider audience with the effect of more informed and safe usage of AI, as well as inspiring further research into educational tools for AI Safety education.
Keywords: AI safety education, game-based learning, gamified learning, educational games

1. Introduction

1.1. Problem Statement

As Artificial Intelligence (AI) developments continue to proliferate, misaligned and unsafe AI models become an increasing concern. This applies for high-stakes applications, but also "everyday" applications such as chatbots and content generation. As AI usage becomes more widespread, the effects and existential risks it may bring to humanity become more of a concern. As a result, *"AI Safety"* is a field that continues to grow alongside AI developments. It aims to ensure that AI remains responsible and aligned with human values, minimizing its potential harms. AI Safety Initiatives around the world have been established to raise awareness about the field, and connect with like-minded individuals through facilitating educational courses on the topic, curated by organizations such as BlueDot Impact. Such courses are often interactive by design, combining independent learning with group discussions led by knowledgeable facilitators. However, these discussions are both the main value and the main challenge of the course due to three main learning difficulties: *(1)* It requires the availability of a facilitator knowledgeable and/or passionate enough to lead the course, *(2)* If an individual is interested in following the course, it requires for a group to be available for them to join (e.g. a local initiative or an online course, which often has limited spots) and *(3)* It requires individuals to have the motivation and drive to keep up with the self-learning aspect of the course – or they will have difficulties participating in the discussions.

*Research conducted at the Women in AI Safety Hackathon, 2025

Our project proposes a single solution to all three problems. Namely, we created a gamified learning tool that supplements the BlueDot “Intro to Transformative AI" course. The proposed application, which we named BUGgy, complements the content of each unit, and provides an evidence-based interactive method for learners to engage with the content. We also tailored this tool to be applicable in multiple scenarios; either used to enhance preparation for in-person group discussions facilitated at an on-site location, or for individuals to further engage with the content they self-study, while staying connected to other users engaging with this course on their own accord. To address the learning challenges above, the proposed tool provides recap assessments of key concepts from course readings, facilitating knowledge consolidation. Considering on-site discussions rely on good preparation, this tool aims to support this preparation in an engaging way, fostering fruitful debates. Additionally, it serves as an assessment tool for self-learners that do not have access to on-site discussion groups, helping them stay engaged with the material as well.

1.2. Background and Motivation

The BlueDot “Intro to Transformative AI Course" is an intensive five-day course aimed at teaching individuals about transformative AI in the context of AI Safety. The course is targeted towards individuals of any background, as long as they are interested and curious to learn more about AI Safety. The content of the course is curated by experts in the field. Though the course is offered directly through BlueDot, many independent initiatives also choose to adopt BlueDot’s courses in their own curriculums. This allows like-minded individuals to connect while engaging with a topic they are passionate about. We chose this course to base our tool on since it is a beginner-friendly and discussion-based course that can benefit from an approachable way to engage with the material - potentially attracting new-comers to the field of AI Safety.

There has been extensive literature on the benefits of Game-Based Learning (GBL), since it encourages reflection among players by drawing connections between “knowing" and “doing" (Shaffer, 2006), and allows users to cultivate their critical thinking skills and practice their knowledge incrementally. Raymer (2011) highlights some key design principles, backed by cognitive psychology, that need to be taken into account for successful GBL. For instance, a compelling “hook" and story, good quality presentation of visuals, and rewards for player effort (and not just success). Furthermore, context-dependent learning has been shown to improve memory consolidation and recall abilities (Reisberg, 2015). Therefore, reinforcing and practicing the content through a GBL context can help solidify memory pathways of the content, and make it easier to recall key components of the course units.

Another factor to consider is that the BlueDot course – and other courses on AI Safety – are highly opinion-based, with many open-ended prompts that require participants to think critically and reflect on their thoughts. This can be reinforced by presenting course participants with reflective questions, where they can apply the knowledge they gained and critically form an opinion on certain matters. Although this is already done in the group-based discussions, it has been shown that personal opinion forming is highly affected by the social context (Moe & Schweidel, 2014; Oxford Royale, n.d.). Therefore, it is important for course participants to perform self-reflection on such prompts before interacting with the group, to solidify their own views on the topic before interacting with other opinions. In turn, this can also improve the quality and depth of the group discussion.

1.3. Threat Model and Safety Implications

Efforts in educating the general public and AI users on AI Safety is beneficial to spread awareness on the risks and limitations AI can pose, to ensure productive and informed usage of AI. This is a critical topic to address since AI Safety is not a generic topic covered in standard education curriculums, despite its importance. AI Safety Initiatives across the globe have been established to address this gap in education, and facilitate courses curated by BlueDot and similar organizations. However, such AI Safety initiatives are located rather sparsely around the world currently, due to the field being in its infancy. Our proposed tool aims to provide AI Safety education opportunities to a wider audience, and alleviate the obstacle of not having access to a local initiative to host such courses. Although the Bluedot courses provide exercises for participants to test their comprehension, our tool presents a more interactive and accessible way to ensure the full understanding of the reading. Therefore, we hope the scalability of this tool can spread awareness on the topics of AI Safety in an approachable, gamified, and social way to encourage non-technical users to

learn more about AI Safety in an enjoyable way. We believe that this complementary tool can improve the reach of the BlueDot course, and can in turn improve awareness and literacy on AI Safety to the general public. This would result in safer and more educated use of AI in people's daily lives, allowing users to use AI carefully, critically, and safely.

2. Methods

Our methods are supported by research into educational theories. We provide the motivation behind our choices in this section.

2.1. Approach

Gameplay and Story

The story revolves around two main characters: Jennifer and BUG. Jennifer is the professor and the head of BUGgy Labs, the location where the game takes place. Jennifer acts as a guide to the user throughout the game, giving exposition to the story, asking the user questions, and communicating important information. BUG is a **B**uddy **U**nder **G**uardianship, a robotic AI that has been entrusted to your care. Initially, BUG has not yet been trained sufficiently, and as a result, is not well-aligned with human knowledge and values, sometimes coming up with random facts. The player's aim is to complete *Units* and *Reading Levels* in the game to collect *Data Juice*, which is used to train BUG further.

Units and Reading Levels

The app is designed to support the BlueDot "Intro to Transformative AI Course", which is divided into five units. Each unit consists of a number of readings and videos. The reading levels consist of questions about the content of the readings, and to complete the unit, questions about all the readings need to be answered. Readings must be completed in order (i.e. to answer questions about the second reading, the questions about the first reading need to be completed). Units must also be completed in order.

Question Design

Two kinds of questions are asked to the user: *technical* and *reflective* questions. Each type of question has a different purpose, addressing different levels of comprehension (Day & Park, 2005). These comprehension levels are key, as we believe it is crucial for students to understand the core concepts behind the readings, beyond the literal understandings of the text.

The *technical* questions are objective questions with definitive correct answers. As a result, these can be externally and automatically assessed by the app. These questions are based on the readings, and they aim to help users consolidate their knowledge of the readings. From personal experience, students often find it difficult to know what information to extract from long readings, so these questions help guide students' understanding. The questions are inspired directly from the text or based on information that is not explicitly mentioned but can be inferred. *Technical* questions can be formatted in a *Fill in the gap*, *Multiple Choice*, or *True/False manner*. Through the *technical* questions, we aim to exercise the students' literal, reorganization and inference comprehension skills (Day & Park, 2005) for students.

Reflective questions are subjective and open-ended questions. They are a method through which participants can self-assess their progress in the course. It asks them to reflect on their opinions based on the content they learned. They might find that their opinion has changed over time. This allows them to keep a log of how their reasoning changes as they learn new information, and helps them understand different perspectives. In the reflection questions, we aim to exercise the prediction, evaluation, and personal response comprehension skills (Day & Park, 2005). *Reflective* questions are crucial to the learning process due to the nuance that is at the core of discussing a topic like AI Safety. They allow learning without needing to define a "correct" answer. Forming an opinion prior to discussing it has shown to be beneficial, as social interaction can result in interdependence of opinions (Moe & Schweidel, 2014), which is why we implemented the reflection questions in the game. Some examples are given in Appendix 5.2.

Digital Journal

The Digital Journal is an in-game tool that allows users to keep track of their answers to *reflective* questions, as well as make their own open-ended entries. This allows for a centralised place for users to record their thoughts as they progress through the readings and the course. Consequently, users can easily revisit their previous opinions and reflect on what might have changed and why. This can also help them prepare for group discussions, as it provides easy access to the user's insights. Journaling has also been shown to be beneficial for the learning process (Lew & Schmidt, 2011).

In-App Discussion

In case the user is unable to find their own group to conduct discussions through a local AI Safety Initiative, the app facilitates discussions through a discussion board. On the discussion board, users can submit discussion threads and reply to threads started by other users. It will allow users to critically think about the materials and engage with them. Adding this feature to the app will increase the scalability and the reach of the course, as it enables people that are unable to join a group in real life to complete the course.

Rewards

Raymer (2011) state that rewards for player effort and not just success are important factors in a game. After successful completion (e.g. getting all of the questions correct) of a reading level, the user will receive a reward. To ensure that not only success is rewarded, but effort as well, the user will receive a reward after filling out the reflection questions (and from posting on the discussion board if applicable) – since these aspects cannot be objectively or externally assessed. Rewards come in the form of *Data Juice* which is fed to BUG, allowing it to learn.

Course Completion

Upon completing the course, three next steps are provided to the user. They aim to either improve the app experience, or increase the outreach and impact of the course. The first "next step" is a feedback form for the user to communicate any findings, shortcomings, or suggestions about the app. The second "next step" is a social matching initiative. The user is asked various questions about their interests, and is then matched to a group of other users who have similar interests. The aim of this is to help create strong connections between like-minded individuals and to encourage further interest in AI Safety topics. The last possible "next step" is a career matching step, which provides links and resources to nearby or online initiatives relevant to the course, so that users can continue their involvement in AI Safety.

2.2. Implementation

A demo of our implementation can be found at[1]. We highly recommend playing through this Demo! It only takes about 5 minutes. We recommend opening the link on a mobile device (such as a smartphone) for the closest experience to what the app would look like after deployment. If using the browser version, we recommend clicking the "Fit width and height" option in the settings on the top right of the screen.

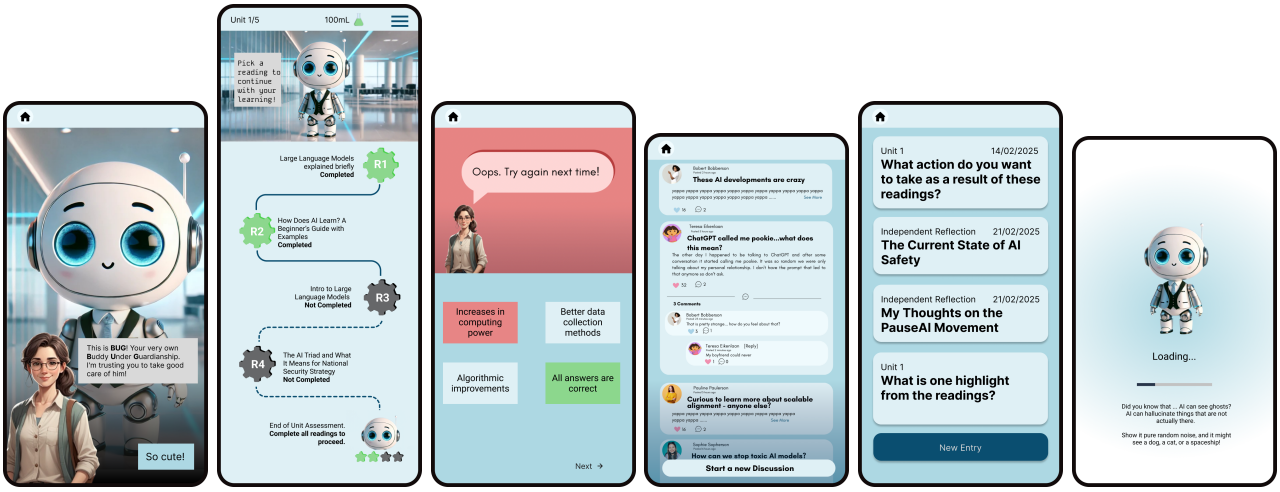
Platform

We created a prototype of our game on Figma Design. This is a low-code platform with powerful prototyping options. We chose to use this platform as a result of the time constraints of Hackathon, for which this app was designed. Additionally, it is highly intuitive while having many interaction options, meaning that high-quality prototypes can be developed rapidly.

Setup

The user begins in an introductory narration where Jennifer introduces herself. She then prompts the user to state whether they are already in a discussion group (for example, through the BlueDot course, or other

[1] https://shorturl.at/jQ9ph



(a) Narration/story screen. (b) Home screen. (c) Result of clicking wrong answer. (d) Discussion Board (e) Digital journal (f) Loading screen with fun fact.

Figure 1: Screenshots of our mobile app, made using Figma Design.

initiatives). If they are not, the app “places" the user in a discussion group (this is not operational in the demo). Following this, BUG and the reward principles are explained, as in Figure 1a. After this brief exposition, the main menu is shown 1b. Here, the player can access each reading, starting from the top. The green gears indicate completed readings, while grey ones are not completed (upon first launch, all readings are grey). This gives a quick, one-glance overview of the information, reducing the working memory load through the use of colors (Wong, 2024). The menu bar at the top allows access to other units, the journal, discussion board, and other settings.

Once a reading is clicked, questions are shown to the user. For this demo, only one question was implemented per reading, as a proof-of-concept. An example question could be “In a neural network, a _ _ _ is a number that defines how much one neuron should influence another ". If the right answer is clicked, the screen turns green, with Jennifer giving a compliment. If the wrong answer is clicked, the screen turns red, and the correct answer is highlighted in green so that users can easily identify the right answer, as shown in Figure 1c.

The discussion board (with example text) is shown in Figure 1d. For this demo, it is not functional. The same goes for the digital journal shown in Figure 1e. These are only shown as proof-of-concepts. Lastly, once the user finishes all the questions, the completion screen shown in Figure 2a appears, where the “next steps" options are given. These are also not functional, but simply for illustration purposes. Note that these elements were not made functional due to the limitations of Figma being a prototyping platform, not an implementation platform. Between key components, loading screens are shown with humerous, AI and Safety-related facts, as shown in Figure 1f.

If a user has not been active in the app for a while, they are prompted to complete questions from earlier readings upon launching the app. This offers them an opportunity for repetition learning to help consolidate their memory (Reisberg, 2015).

The overall setup resulted in over 50 frames (Figure 1 shows six of these) being created. Additional screenshots can be found in Appendix 5.3. Please note that the visuals for Jennifer and BUG were AI-generated.

3. Results

3.1. Analysis and Findings

To evaluate our game, we used a cognitive walk-through and a questionnaire as our method. A cognitive walk-through refers to the user stating every step of going through the game out loud, which gives insight into any unclarities or features that are misunderstood. The questions served as an evaluation method for

the user interface, the questions, and the overall relevance of the game (see Appendix 5.4).

We recruited two students from our social circle to perform the cognitive walkthrough and complete the questionnaire. Both students are aged in their early twenties, and both study Artificial Intelligence; one graduate and one undergraduate student. The graduate student is an avid video game player, and therefore has a lot of experience with interacting with engaging games and interfaces - providing useful insights on how to design the interface and game-play. The undergraduate student has extensive AI Safety knowledge, after completing BlueDot's AI Safety Fundamentals course, as well as directing an AI Safety Initiative. Therefore, his experience with BlueDot courses and the field gives us good insights on the quality and difficulty of the questions asked, and if they serve their intended purpose.

The qualitative results show that the visual appeal of the game is good. However, some points of unclarity about the rewards and the effect of the rewards have been highlighted. Some visual features concerning the flow of the game were difficult to comprehend, and need to be improved on. The quality of the questions was difficult to assess due to prior knowledge, but they stated that they expect the quality to be suitable for users without prior knowledge about the readings.

Overall, the feedback shows that the app has potential to be useful in the context of the BlueDot course materials (see Appendix 5.5 for the full summary of the evaluation).

3.2. Impact Assessment

The general impression was that the app was well-received by our participants. They exhibited enthusiasm for the tool and found it fun and engaging. Many of their points of critique are related to the limitations of Figma, the platform we used to make our prototype. These could easily be solved once a full-fledged product is developed in the future. It was never the intention for the Figma app to be the final product, but due to the time limits of the Hackathon, this was the minimal viable product that we were able to propose.

One participant also proposed incorporating an actual Large Language Model to generate BUG's behaviour over time, as it is trained on more data. The users could then interact directly with BUG (for example, by chatting with it) and could, in real-time, see BUG's alignment change over time.

4. Discussion and Conclusion

An important aspect of AI safety is getting the information across to the general public and our game has a lot of potential for positively affecting the scalability of the BlueDot course. As our game implements an in-app discussion function, people who are unable to find a facilitator or join a group still have access to the materials and the course, which increases the availability of the course. The BlueDot course also adheres to a limited amount of participants, but with the use of our game the course could be made more accessible. Furthermore, through gamification our project focuses on increases motivation and an improved learning experience. Multiple features in the game ensure both an engaging and reinforced learning experience as a support tool for the BlueDot course.

We also found that the fluidity of the app plays a large role in the engagement. Therefore, solving these bugs and implementing all functionalities is crucial to deliver a pleasant and effective experience to the users. With a well-developed and engaging app, it makes interest in AI Safety more appealing and accessible to the general public. If supported by more experience in software engineering and graphic design (our original designs can be found here[2]), we strongly believe that this app could make a positive impact on AI Safety reach around the world. Lastly, we want to highlight that the app is a framework that can very easily be tailored and adapted to other courses of similar structure. This makes our proposed product highly scalable, increasing the reach of AI Safety education in the world.

AI safety is an incredibly important and rapidly developing field. Even though the game we proposed is a work in progress, we hope it inspires future research to explore further fun and engaging ways for teaching AI safety.

[2] https://shorturl.at/heEXt

References

Day, R. R., & Park, J.-s. (2005). Developing reading comprehension questions. *Reading in a foreign language*, *17*(1), 60–73.

Lew, D. N. M., & Schmidt, H. G. (2011). Writing to learn: Can reflection journals be used to promote self-reflection and learning? *Higher Education Research & Development*, *30*(4), 519–532.

Moe, W. W., & Schweidel, D. A. (2014). Fundamentals of opinion formation. In *Social media intelligence* (pp. 18–34). Cambridge University Press.

Oxford Royale. (n.d.). *9 ways to think more rationally and develop your own opinions*. Retrieved March 9, 2025, from https://www.oxford-royale.com/articles/rationally-develop-opinions?utm_source=chatgpt.com

Raymer, R. (2011). Gamification: Using game mechanics to enhance elearning. *ELearn*, *2011*(9).

Reisberg, D. (2015). *Cognition: Exploring the science of the mind: Sixth international student edition*. WW Norton & Company.

Shaffer, D. W. (2006). Epistemic frames for epistemic games. *Computers & education*, *46*(3), 223–234.

Wong, E. (2024). *Shneiderman's Eight Golden Rules Will Help You Design Better Interfaces*. https://www.interaction-design.org/literature/article/shneiderman-s-eight-golden-rules-will-help-you-design-better-interfaces#8_golden_rules_of_interface_design-0

5. Appendix

5.1. Team Member Contributions

The project was completed in a collaborative manner. All team members contributed.

5.2. Example reflection questions

We provide some guidance on the reflection questions. These can be prediction questions about the content of the readings, for example: how do you expect the labor market to be affected by AI developments in the next 10 years? Furthermore, reflection questions can focus on the student's own judgment on the meaning of the text, and their personal opinions (evaluation and personal response comprehension).

5.3. Additional Screenshots

(a) Course completion screen. (b) Example question (c) Example reward screen. (d) Example *reflective* question (e) Welcome back screen after long periods of inactivity.

Figure 2: More screenshots of the mobile application made in Figma

5.4. Evaluation Questions

1. User Interface

(a) How was the intuitiveness of the interface?

(b) How would you rate the intuitiveness of the interface on a scale from 1-5?

(c) Are there any aspects of the interface you would add?

(d) Are there any aspects of the interface you would remove?

(e) Were there any aspects you particularly struggled in understanding? Were there any aspects of the interface you were lost/confused about? Or did not know their purpose?

(f) Are there any particular features you enjoyed having?

(g) On a scale of 1-5, how much were you able to perform the actions you wanted to?

2. Questions

(a) Do you feel like this could help in your learning? Would you use this app in learning about AI Safety or other topics?

(b) What do you think of the structure/framing of the questions asked? Were some too easy/ complex considering the readings?

(c) How would you rate the quality/difficulty of the questions on a scale from 1-5?

3. Overall

(a) How would you rate the usefulness of the tool on a scale from 1-5?

5.5. Evaluation Results

Interface Evaluation
Qualitative
The visuals in general were pleasant to look at, and the avatars for BUG and Jennifer were good.
Found the concept of *Data Juice* and its purpose confusing – the objective of the game can be better communicated.
Found the story of BUG needing data to improve was unclear at the start.
It should be made more clear that the questions have only one correct answer, not multiple.
One did not click on the "hamburger" drop-down menu to explore other core features such as the discussion board and journal.
Likes the loading pages and fun facts.
The flow of the interface can be improved – the user is not sufficiently prompted on what to click next after completing a task.
The *Data Juice* is given even after an incorrect answer is given – this did not feel deserved.
The yet-to-be-implemented reflection answer box was mistaken for a button.
After making a mistake, the user cannot go back to the question and redo the level.
It is unclear what happens when BUG drinks the *Data Juice*, and how BUG improves.
Quantitative
Control over actions: 4.5/5 mean score.

Table 1: Interface Evaluation Summary

Questions	
Qualitative	
	Think that it has good potential and would assist with learning. Question quality was adequate, clear formulation. Question difficulty was hard to assess due to level of prior knowledge, but thinks that without prior knowledge the difficulty level would be good. It needs to be made clear that there is only one correct answer.
Quantitative	
	Quality of questions: 4/5 mean score.

Table 2: Evaluation of Questions