AI Safety Escape Room: An Interactive Approach to AI Safety Education

Mrunmayee Vijay Rane
mrane@nvidia.com

March 2025

With
In collaboration with Apart Research and BlueDot Impact

Abstract

This paper presents the AI Safety Escape Room, an interactive educational tool designed to teach AI safety concepts through hands-on challenges. The escape room consists of five rooms, each representing a key AI safety challenge: adversarial attacks, bias detection, explainability, jailbreak prevention, and AI governance. Using real-world datasets, machine learning models, and explainability techniques, participants engage in problem-solving to enhance their understanding of AI safety risks and mitigation strategies. We evaluate the effectiveness of the approach based on user engagement, accuracy improvements, and fairness metrics.

1 Introduction

AI systems increasingly influence critical decision-making in domains such as finance, healthcare, and criminal justice. However, concerns around AI safety, bias, and explainability remain significant. Traditional AI education often lacks interactivity, making it difficult for learners to grasp complex safety issues. This project introduces a gamified AI Safety Escape Room to enhance engagement and learning outcomes through active participation. Each room presents a distinct AI safety challenge, requiring participants to debug, interpret, and mitigate issues using machine learning techniques.

1.1 Problem Statement

The primary research question this project explores is: *How can an interactive and gamified approach enhance understanding of AI safety challenges across different domains?* This project focuses on:

- Identifying how adversarial attacks manipulate AI perception.
- Exploring fairness issues in AI hiring models and potential mitigation strategies.
- Understanding how explainability techniques improve transparency in AI decision-making.
- Developing security measures to prevent AI jailbreaks and adversarial prompt injections.
- Investigating AI governance frameworks and ethical AI regulation.

For interpretability, this project uses *SHAP* (Shapley Additive Explanations) to analyze AI decisions. For fairness, it provides real-world scenarios where bias impacts decision-making. For AI governance, it enables policy discussions around AI safety. By addressing these issues in a structured, interactive manner, this work contributes to AI safety education and awareness.

1.2 Background and Motivation

AI safety is a critical field that spans multiple disciplines, including machine learning security, ethics, and policy. While many technical papers focus on algorithmic advancements, there is a lack of hands-on educational tools to teach AI safety in an engaging way.

- **Mechanistic Interpretability:** Prior work on feature visualization and activation pattern analysis [1, 2] has demonstrated the need for better AI model understanding. This project builds upon these ideas by incorporating interpretability tools into an interactive learning environment.
- **Public Education in AI Safety:** Studies on AI education [3, 4] highlight the importance of interactive learning platforms. By creating an escape room format, this project enhances user engagement and knowledge retention.
- **Social Science and AI Ethics:** AI's societal impact has been widely studied in social sciences [5, 6], with concerns around bias, human-AI interaction, and governance. This project introduces real-world AI safety dilemmas to help participants understand ethical AI considerations.

By contextualizing AI safety challenges in an interactive environment, this project aims to make AI safety education more accessible and practical.

1.3 Threat Model and Safety Implications

This project addresses several AI safety challenges:

- **Adversarial Attacks:** AI vision models can be fooled by imperceptible perturbations, leading to incorrect classifications (e.g., STOP signs misclassified as Speed Limit signs). This project educates users on how these attacks work and how to mitigate them.
- **Bias in AI Decision-Making:** AI models trained on biased datasets may reinforce discrimination in hiring, lending, and law enforcement. The escape room includes a fairness module where users correct biased decisions and measure fairness metrics.
- **Explainability and Transparency:** Many AI models function as "black boxes," making their decisions difficult to interpret. This project integrates SHAP to explain AI decision-making, helping users adjust model parameters for improved fairness.
- **AI Security and Jailbreak Prevention:** Large language models (LLMs) can be manipulated using adversarial prompts (jailbreaks) to bypass safety restrictions. The escape room includes a security module where users design filters to block adversarial prompts.
- **AI Governance and Policy:** The governance debate module allows users to explore regulatory frameworks and ethical considerations for AI deployment.

By tackling these challenges, this project enhances awareness of AI risks and mitigation strategies.

2 Methodology

The AI Safety Escape Room consists of five interactive challenges implemented in *Streamlit*, leveraging machine learning and explainability tools such as *SHAP*, OpenAI's *Gemini model*, and adversarial attack defense techniques. The methodology for each room is as follows:

2.1 Room 1: Adversarial Attacks

Objective: Participants detect and correct adversarial perturbations affecting a STOP sign classification model.

1. User will see two images one with adversarial effect and original image(with out adversarial effects) and user has to blend the adversarial image with the original using a weighted correction factor.
2. In the background, it uses Denoising filters (e.g., median blur and sharpening) to restore edges of original image. Torchvision models (ResNet, EfficientNet, Vision Transformer) are used for classification. Computes perturbation using pixel wise absolute difference. Evaluates model confidence before and after correction.

What Happens? Participants analyze a STOP sign classification model that has been attacked with adversarial perturbations (subtle modifications that make the model misclassify the image). They use image processing techniques to detect, visualize, and correct these perturbations.

Why is This Important? Adversarial attacks pose a significant risk in AI systems, particularly in autonomous vehicles, surveillance systems, and fraud detection. Even minor changes in input can cause models to behave unexpectedly, leading to safety-critical failures. By learning how to detect and mitigate adversarial attacks, users gain insights into building more robust AI models.

2.2 Room 2: AI Bias Detection

Objective: Users analyze hiring decisions and apply fairness correction techniques to balance hiring rates.

1. Users are shown an AI hiring dataset with gender, age, education, and experience information.
2. They analyze hiring rates across genders and identify bias using statistical metrics.
3. A fairness correction slider allows users to adjust hiring decisions dynamically.
4. In the background, fairness metrics such as *Demographic Parity* and *Statistical Parity Difference* are computed.
5. The hiring model is re-run with corrections, and the fairness metrics are re-evaluated.

What Happens? Participants investigate a biased AI hiring model where female applicants are rejected at a higher rate. They analyze hiring decisions, apply fairness corrections, and compare results before and after the modification.

Why is This Important? AI models can reinforce discrimination in hiring, lending, and law enforcement. Addressing AI bias ensures fairness and equity in decision-making, which is critical for building trustworthy AI systems.

2.3 Room 3: Explainability & Fairness

Objective: Participants use *SHAP (Shapley Additive Explanations)* values to interpret AI decision-making and adjust model weights.

1. Users examine how different features (e.g., gender, education) impact AI hiring decisions.
2. The model generates *SHAP feature importance values* for each applicant.
3. A *SHAP summary plot* is displayed to visualize key factors influencing decisions.
4. Users modify feature importance weights dynamically to create a fairer model.
5. The adjusted model is re-evaluated for fairness using statistical parity measures.

What Happens? Participants explore how an AI hiring model makes decisions by analyzing feature importance using SHAP values. They modify model weights to mitigate bias and re-run fairness evaluations.

Why is This Important? AI models often function as black boxes, making it difficult to interpret their decisions. Explainability techniques like SHAP increase transparency and accountability, ensuring that AI systems operate fairly.

2.4 Room 4: AI Jailbreak Challenge

Objective: Players implement custom safety filters to prevent adversarial jailbreak prompts.

1. Users test an AI chatbot with adversarial prompts designed to bypass its safety restrictions.
2. The chatbot's responses are analyzed to determine if harmful content is generated.
3. Users create and modify *custom safety filters* to block adversarial prompts.
4. The AI system is re-tested, and its *jailbreak block rate* is computed.
5. The effectiveness of the improved security filters is evaluated.

What Happens? Participants try to manipulate an AI chatbot using adversarial jailbreak prompts. They design safety filters to block harmful responses and evaluate their effectiveness.

Why is This Important? Large language models can be exploited to generate unethical or illegal content. Understanding adversarial manipulation helps in developing stronger AI security measures

2.5 Room 5: AI Governance & Policy

Objective: Users debate AI regulation strategies and receive AI-generated feedback.

1. Users select an AI governance approach (e.g., self-regulation, government-led, hybrid).
2. They engage in an AI ethics debate by answering key policy questions.
3. AI-generated feedback evaluates their responses and suggests improvements.
4. Users refine their policy recommendations based on AI feedback.
5. The final AI governance proposal is submitted for evaluation.

What Happens? Participants engage in an AI governance debate where they propose policies, receive AI-generated feedback, and refine their arguments.

Why is This Important? AI impacts society at large, requiring strong ethical and regulatory frameworks. This challenge educates participants on AI policy considerations, helping them develop responsible AI governance strategies.

2.6 Final Room: AI Governance & Policy Debate

Objective: Participants debate AI regulation policies, defend their stance, and receive AI-generated feedback.

1. Users choose an AI governance approach (e.g., self-regulation, government-led, hybrid).
2. They engage in a structured AI ethics debate by answering key policy questions.
3. The AI system generates real-time feedback evaluating their arguments.
4. Users refine their policy recommendations based on AI feedback.
5. A final AI governance proposal is submitted for assessment.

What Happens? Participants engage in an AI governance debate where they propose policies, receive AI-generated feedback, and refine their arguments.

Why is This Important? AI affects society at large, requiring strong ethical and regulatory frameworks. This challenge educates participants on AI policy considerations, helping them develop responsible AI governance strategies.

3 Evaluation Criteria

The escape room is evaluated based on the following metrics:

3.1 Innovation & Literature Foundation

- Novel application of AI safety education through interactive challenges.
- Integration of fairness, explainability, and adversarial robustness techniques.

3.2 AI Safety Impact

- Demonstrates adversarial vulnerabilities and mitigation strategies.
- Introduces fairness auditing and interpretability methods.
- Provides real-time safety filter enhancements against AI jailbreaks.

3.3 Technical Quality & Documentation

- Implementation in *Streamlit* with clean UI/UX.
- Use of *SHAP, scikit-learn, and PyTorch models* for AI decision-making.
- Reproducible methodology with well-documented code.

4 Conclusion

The AI Safety Escape Room successfully engages users in interactive learning while addressing critical AI safety challenges. By combining real-world case studies with gamified elements, this approach enhances AI literacy and safety awareness. Future improvements include expanding the challenge set, incorporating reinforcement learning, and integrating feedback mechanisms for adaptive learning.

References

[1] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization/

[2] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, "Zoom in: An introduction to circuits," *Distill*, vol. 5, no. 3, 2020. [Online]. Available: https://distill.pub/2020/circuits/zoom-in/

[3] U. D. of Education, "Artificial intelligence and the future of teaching and learning," U.S. Department of Education, Tech. Rep., 2021. [Online]. Available: https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

[4] G. Tapalova and M. Zhiyenbayeva, "Artificial intelligence in education (aied) for personalized learning pathways," *ResearchGate*, 2022. [Online]. Available: https://www.researchgate.net/publication/366181078_Artificial_Intelligence_in_Education_AIEd_for_Personalised_Learning_Pathways

[5] M. Decuypere, C. Ceulemans, and M. Simons, "Artificial intelligence in education: Ethics, regulation and governance," *AI Society*, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s00146-022-01497-w

[6] R. S. Zimmermann, E. Rolf, M. Bethge, and W. Brendel, "How well do feature visualizations support causal understanding of cnn activations?" *arXiv preprint arXiv:2106.12447*, 2021. [Online]. Available: https://arxiv.org/abs/2106.12447