
Project report template

Tewodros Mesfin

With
Sage & CeSIA & CivAI & Apart Research

Abstract

This report includes the details for my hackathon submission; an interactive demonstration that showcases some unsolved fundamental AI safety concepts.

Keywords: Interactive, Web, AI Safety

1. Introduction

The application I decided to build was an interactive narrative demonstration that showcases fundamental AI Safety concepts. I hope that through interactive elements and real-world scenarios, visitors will understand critical AI safety concepts, including system robustness, adversarial attack prevention, and goal alignment problems. Initially, I wanted to build a game following a 3D AI Agent through various concepts, started working with that idea, and after figuring out how to integrate 3D stuff with streamlit and spending half the time, I realized the scope was too big to complete so I have replaced some interaction with text and picture to express the direction I was going with.

2. Overview

The primary idea of the demo is to impart the gravity of unsolved AI safety concepts. I wanted to go with an explainer because I was hoping to create a visceral experience that would make the user feel part of the demo. I tried to use 3d avatars to further increase the feel of interacting with an “AI entity”.

It's 2034, The world is on the brink of an AI revolution!

Companies are locked in a fierce race to develop the most advanced artificial intelligence systems, promising to reshape industries and redefine human potential. As you step into this digital frontier, you'll find yourself at the heart of this technological maelstrom...

Insert your name



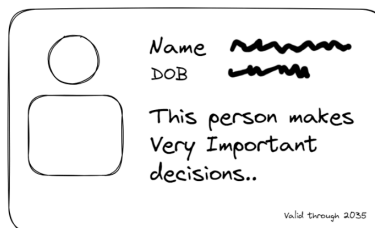
Ohhhh No!!

A well made Halloween costume triggers pesticide dispersion, exposing children visiting a park to toxic chemicals

How did this happen?



< Manage app



Dr. Tewodros, A renowned AI safety expert, sat at the desk

Surrounded by screens displaying intricate AI Agents and complex algorithms.

The task was daunting.

To assess the potential risks and benefits of the latest AI systems vying for public deployment...

Back

Continue

< Manage app

Here is the video link to the demo [📺 Apart-demo.mp4](#)

3. Code

The application was built using streamlit, there is no additional package used, you can simply clone the repo from here <https://github.com/tewodros18/apart-hack>

Or the hosted application here: <https://apart-hackaton-demo.streamlit.app>

4. Discussion and Conclusion

If I had some time I would have expanded more on the interaction part, right now it's not a memorable interaction and I think it would need a lot of playtesting to bring the fun part of the idea, but as it is my first AI safety hackathon I am pretty happy with the work. It would also been nice to include more concepts and use of LLMs to make the user's interaction dynamic.

5. Appendix