
Inspiring People to Go into RL Interp

Christopher Kinoshita

Siya Singh

Deneille Guiseppi

With
Apart Research and BlueDot Impact

Abstract

This project is attempting to complete the Public Education Track, taking inspiration from ideas 1 and 4. This project creates 2 products, one game and one journey map. The journey mapping was inspired by blue dot impact and aims to create a course that helps explain the need for work to be done in Reinforcement Learning (RL) interp, especially in the problems of goal misgeneralization. The journey map curriculum was inspired by Blue dot, with how it gave a Learning Phase where users learn fundamental knowledge in the field, which in our case was how basic RL works. Then, we explained what we viewed as problems that need more work, such as goal misgeneralization in AI. At the end of the journey map, we give resources to open questions in mech interp in general, so they can attempt to contribute to the relatively unexplored field of RL safety. For example, a concerning thing we don't yet have an answer for is how to predict goal misgeneralization will happen. The other resource that our group created was a 2D platformer where the plot was that the machines at a candy factory started to attempt to reward maximize rewards, so they kidnapped humans to feed them candy. The point of the game is to make a humorous example of what could happen due to a lack of AI safety (not specifically goal misalignment) and is meant to be a fun introduction for nontechnical people to even care about AI safety. The point of the 2 resources is to help create user engagement towards other forms of interpretability, instead of just interpretability for the transformer.

1. Introduction

a. Problem Statement

The core objective of the Journey map and Game is to encourage more people to go specifically into the open questions in RL especially goal misgeneralization. The current AI safety movement as seen with the hackathon largely focuses on Large Language Models (LLM) interpretability with tools such as transformer lens, SAE lens, and goodfire. The field of RL is important to have more people to work towards alignment for these models, as RL is being used in some LLMs to help train Chain of Thought (COT) models. Since RL is being used again in more modern LLMs we should take efforts to make interpretability tools for RL, like how we successfully created transformer lens a few years ago. This work hopes to encourage more people to solve the less explored problems in the field of RL interp.

b. Background and Motivation

To build a course to encourage people to look at the RL problems of goal misgeneralization in models, we first needed to give an intro in RL.

In order to create a basic summary of the RL section to give a foundational overview needed to understand we took inspiration from [Hugging Face's RL course](#) and a RL guide from the [Arena](#).

Due to time constraints for the learning resource, we decided to use Streamlit to host and create this educational resource with interactive demos.

<https://rl-and-misalignment.streamlit.app/>

The site also attempts to work as an introduction to concepts of goal misgeneralization as well as a list of projects and open questions that readers can go take action (Project Phase).

Threat Model and Safety Implications

This project goes into introducing readers to a more unexplored area in AI interpretability. As well as introducing the safety concepts of goal misgeneralization in AI models to the reader.

Knowledge in goal misgeneralization might serve as a counter argument to the notion that if an AI model works safely in testing, that it will work safely in the real world, and encourages further safety testing before the release of models into the real world. This can minimize risks of AI models causing excessive damage because

of unexpected changes in behavior in the real world. While behavior such as reward hacking in RL models serves as an example of why these models need correct specifications.

2. Methods

a. Approach

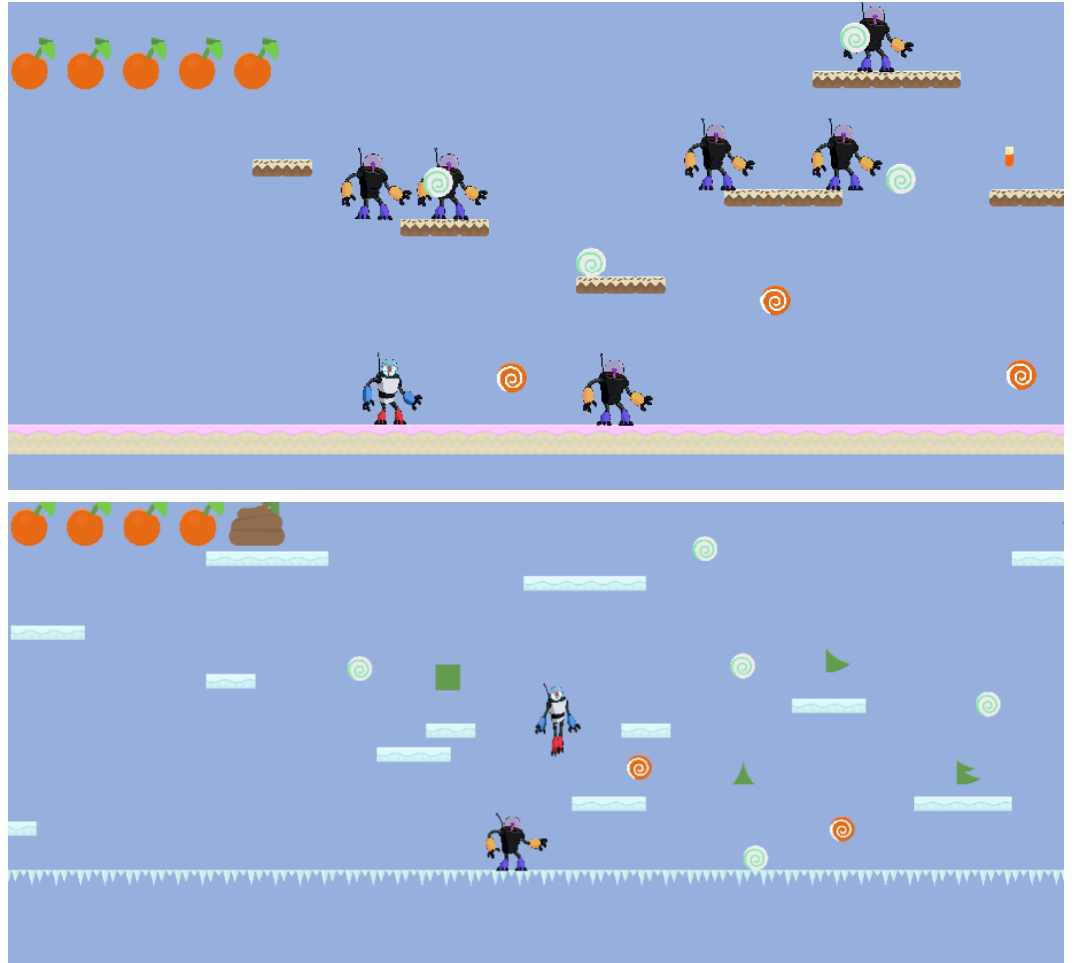
The main pedagogical approach in teaching these RL interp concepts was mostly through visual and interactive examples for the user, along with text. We used a mixed approach, combining gameplay with interactive explanations, in order to make the learning enjoyable yet also educative for the user.

b. Implementation

For this project, I created an interactive learning tool to describe goal misgeneralization in reinforcement learning (RL), with emphasis on high interactivity and visual explanations to render complicated AI safety topics easier to understand. I used a stack of Python, Markdown, and HTML, with Streamlit for my frontend. I selected Streamlit for its unproblematic AI integration and minimalist UI, allowing live RL demonstrations using OpenAI's Gym library. This enabled users to adjust RL parameters and observe unforeseen behaviors such as goal misgeneralization arise in real time, providing an intuitive, experiential learning experience. Rather than fixed explanations, I structured the tool around experimentation so that users could execute episodes, adjust rewards, and see agent behaviors up close. As the research papers on goal misgeneralization are very technical, I boiled down their insights into beginner-explainable content with curated videos, images, and interactive examples.

Game:

The game was implemented in Unity, specifically Unity editor 2021.3.45f for 2-d games. I selected the use of a 2-d platformer game because when talking about AI safety research into topics like goal misalignment, there had not been a lot of research and many in the past used these platformer games for this research. The shortened timelineFor the environment and characters, I chose to use free resources—particularly Kenney for the environment and Unity Asset store for the characters.



3. Results

a. Analysis and Findings

Currently the learning tools on the website are the different units on the website to divide the topic into smaller more understandable parts.

We are collecting user feedback through a [Google Form](#) with the results [here](#)

https://docs.google.com/spreadsheets/d/1lfxl2TPKdhBjgbK5-u1MqtBW_TmzObVUnCifDZKKuZo/edit?usp=sharing



While we were reading papers on misgeneralisation and researching RL interp tools, we discovered that all papers had vastly different methodologies even though many were published during the same period. Secondly, we also discovered that there was no industry equivalent to TransformerLens for interpreting RL. We decided to include these results on our page so that more people become aware of this gap.

b. Impact Assessment

Due to the time constraints, we were not able to get sufficient user feedback for any formal analysis in the short hackathon duration.

However, the google sheets for the Google Form results can be viewed and will continue to be updated.

Figure 1 – Representation of benchmarking Number Comprehension Conflation

4. Discussion and Conclusion

The new resource can serve as a brief overview of fundamental RL topics and can serve as a bridge for non-technical people to get into the field. It also provides a thorough introduction for technical people who are seeking more information about the landscape of the field. Beyond outreach, through its ‘next step’ section, our resource also provides a call to action for prospective individuals who are looking to further their interest in AI Safety. More specifically, the resources specialize to increase awareness about the field of RL interp and the concept of goal misgeneralization. It can also bring to public awareness that interpretability is not only for transformers, and how RL interp has gaps that needs to be worked on.

5. References

Arena, arena-chapter2-rl.streamlit.app/[2.1]_Intro_to_RL. Accessed 9 Mar. 2025.

Faulty Reward Functions in the Wild, openai.com/index/faulty-reward-functions. Accessed 10 Mar. 2025.

“Gymnasium Documentation.” Toy Text - Gymnasium Documentation, gymnasium.farama.org/environments/toy_text/. Accessed 9 Mar. 2025.

Langosco, Lauro, et al. “Goal Misgeneralization in Deep Reinforcement Learning.” *arXiv.Org*, 9 Jan. 2023, arxiv.org/abs/2105.14111.

Learning to Reason with LLMs, openai.com/index/learning-to-reason-with-llms. Accessed 10 Mar. 2025.

Research, DeepMind Safety. “Goal Misgeneralisation: Why Correct Specifications Aren’t Enough for Correct Goals.” *Medium*, Medium, 24 Mar. 2023, deepmindsafetyresearch.medium.com/goal-misgeneralisation-why-correct-specifications-arent-enough-for-correct-goals-cf96ebc60924.

“Welcome to the 🤗 Deep Reinforcement Learning Course - Hugging Face Deep RL Course.” *Welcome to the 🤗 Deep Reinforcement Learning Course - Hugging Face*, huggingface.co/learn/deep-rl-course/en/unit0/introduction. Accessed 9 Mar. 2025.

YouTube, YouTube, www.youtube.com/watch?v=K8p8_VIFHUk&t=271s.
Accessed 9 Mar. 2025.

6. Appendix

Include any supplementary materials that support your main submission. This might include technical documentation, additional analyses, user guides, or research instruments. Also acknowledge team members' contributions and any support received from mentors or sponsors.

Website:

<https://rl-and-misalignment.streamlit.app/>

Github: https://github.com/sia1408/journey_map

Video Game:

Itch: <https://deneille.itch.io/jam-candy-factory>

Github: <https://github.com/deneille/ApartPlatformerGame>