Aligning Reinforcement Learning Systems with Human Intent

Stanford CS221
Embedded Ethics Lecture, Week 4
Myra Deng, Veronica Rivera

Some content is adapted from Jacob Steinhardt's Tutorial: Aligning ML Systems with Human Intent, SaTML 02/10/2023

Learning objectives

- Develop a general understanding of why AI alignment is important
- Understand alignment in the context of reinforcement learning (RL)
- Describe two different approaches for aligning RL agent exploration
- Explore the pros and cons of two different alignment techniques: 1) constrained RL and 2) RL with human feedback (RLHF)

Why is alignment important?

Classical Example (Amodei & Clark, 2016)

- Researchers wanted to teach the boat agent to win the race
- They trained the model to get the most points
- Agent **found a loophole** where it could rack up points without making any race progress

D. Amodei and J. Clark, "Faulty Reward Functions," OpenAI, June 21, 2016, https://openai.com/research/faulty-reward-functions.

Why is alignment important?

Model could avoid the task of winning *not by failing to follow instructions*, but rather *by doing exactly what it was told*

Reward hacking: metrics can sometimes become unreliable once optimized

Emergence: new qualitative behaviors arise at scale that are hard to predict

Why is alignment important?

Intent alignment: Does the system conform to the intended goals of the system designer?

However, intent can be:

- **difficult to specify formally** (consider honesty, fairness, polarization)
- **can often be implicitly coded**: [...while not breaking the law, not doing harm, being truthful...]
- **difficult to collect data on** (user surveys might be too limited)

General context: Aligning AI systems with human intent

Intent alignment: Does the system conform to the intended goals of the system designer?

Intent can be **difficult to specify formally** (consider honesty, fairness, polarization)

- And can often be implicitly coded: [...while not breaking the law, not doing harm, being truthful...]

J. Steinhardt, "Aligning ML Systems with Human Intent," Tutorial presented at the Symposium on Trustworthy Machine Learning (SaTML), February 10, 2023.

Alignment in RL

In RL, an agent learns through trial and error in a dynamic environment to maximize a reward function

Alignment is particularly important in RL for several reasons, including:

- Reward specification
- **Exploration**
- Sequential decision making
- Emergent behavior

Safe exploration in RL

RL agents need to explore their environments in order to learn optimal policies

Training RL agents in the real world allows us to capture meaningful complexities (e.g., human-AI interaction)

Safe exploration in real-world settings is of paramount importance to minimize harm, but how do we specify this?

Constrained RL

One potential solution to the safe exploration problem

Set safety specifications as constraints, separate from task-based specifications

How does constrained RL work?

Find the optimal policy under a constrained set of policies

Definition: policy

A **policy** π is a mapping from each state $s \in$ States to an action $a \in$ Actions(s).

Constrained MDP defines feasible policy sets

$$\pi^* = \arg\max_{\pi \in \Pi_C} J_r(\pi)$$

Cost-based constraint functions J_c (separate from reward functions) relative to human-defined threshold (d_i)

$$\Pi_C = \{\pi \ : \ J_{c_i}(\pi) \leq d_i, \quad i = 1, ..., k\},$$

Pros and cons of the constrained RL approach

- Simple to define (constraints are a natural way to define safety requirements)
- Explicit, known constraints

- Still potential for reward hacking (constraint hacking)
  - Unsafe conditions not captured by the constraints may emerge
- Convergence and stability can be a challenge

RL with Human Feedback (RLHF)

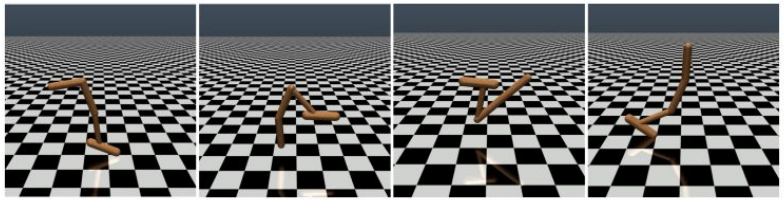
A general approach to the alignment problem

Have human annotators train a reward model that teaches the agent desirable and undesirable behavior

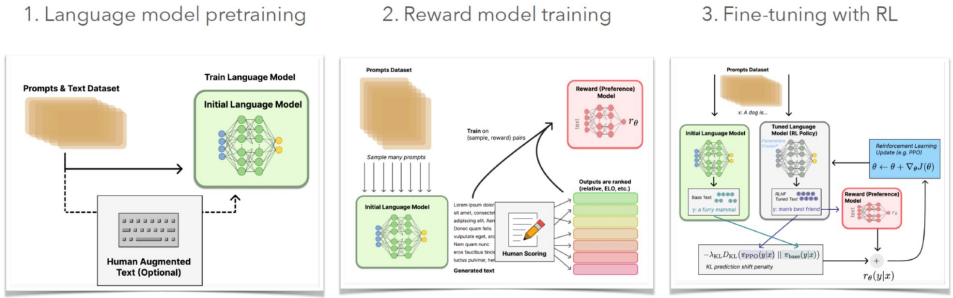
How does RL with human feedback work?

Basic strategy:

- elicit human feedback on system outputs
- retrain system to **produce outputs that are closer to human-preferred examples**



Paul Christiano et al., "Deep Reinforcement Learning from Human Preferences," OpenAI, 2017, https://arxiv.org/abs/1706.03741.



CS329x slides

Pros and cons of RLHF approach

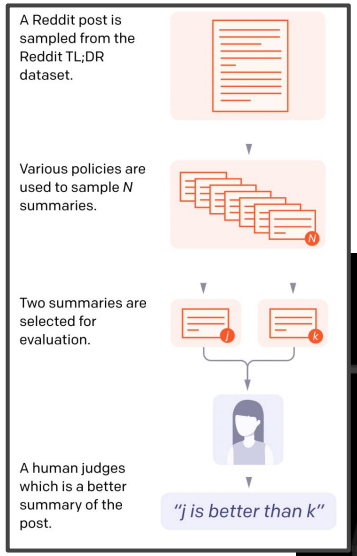
- Allows system to learn nuances that are hard to specify via formal rules or constraints
- More scalable approach (collecting human feedback datasets) than manually specifying rules

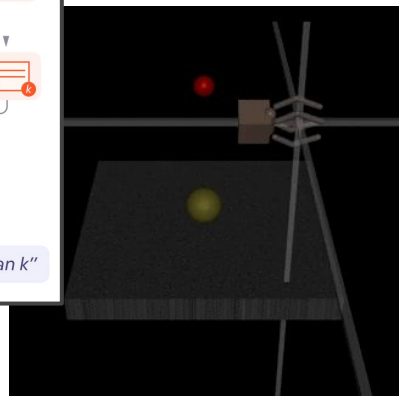
- Implicit, unknown constraints
- Quality of human feedback can be variable
- Expensive to manually annotate data and train an additional reward model
- System can also learn to hack the reward model, emergent undesired behavior

Challenges in collecting “good” human feedback

- Current binary approach is not necessarily ideal

- As a system designer, we should think about:
  - Who is providing feedback?
  - What are their incentives / are they being compensated appropriately?
  - Are there any data privacy concerns?
  - What are potential biases in our feedback mechanisms?





Ryan et al. (2023). Unintended Impacts of LLM Alignment on Global Representation. arXiv:2402.15018. https://arxiv.org/pdf/2402.15018.pdf
TIME, OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Thank you!

Please reach out on Ed if you have any feedback.