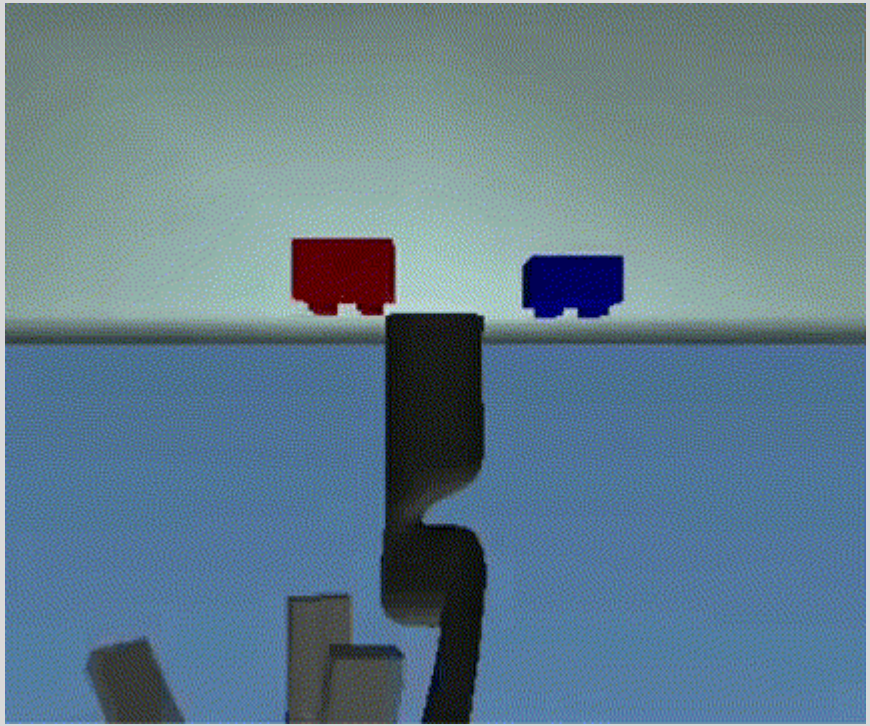
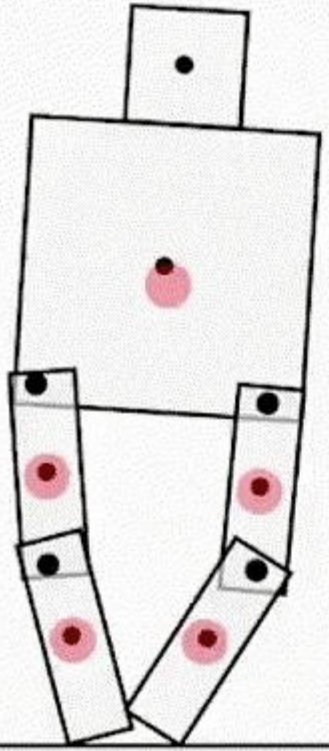
RL and Problems

Specification gaming

It is sometimes very hard to specify a score function when you want to train a model to do some complex task.

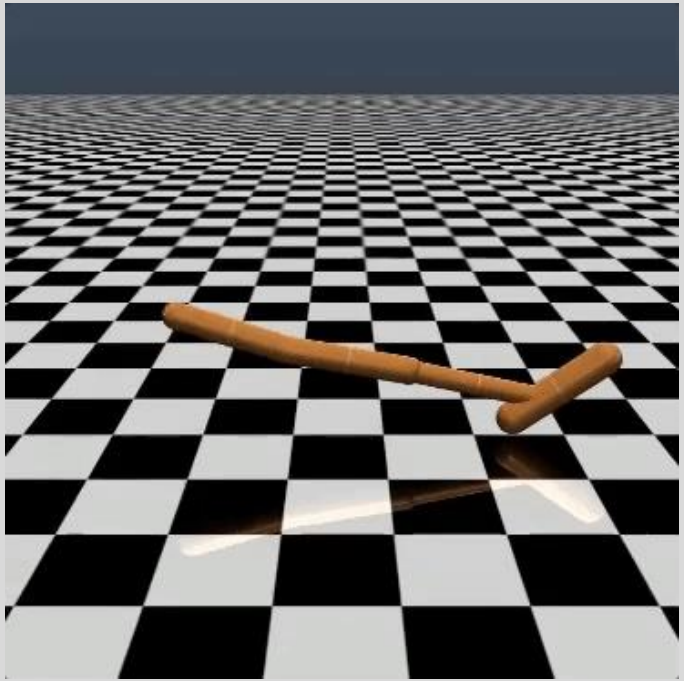



TURBO
SCORE
13500
LAPS
–/3
TIME
0:26
TURBO
MORE GAMES

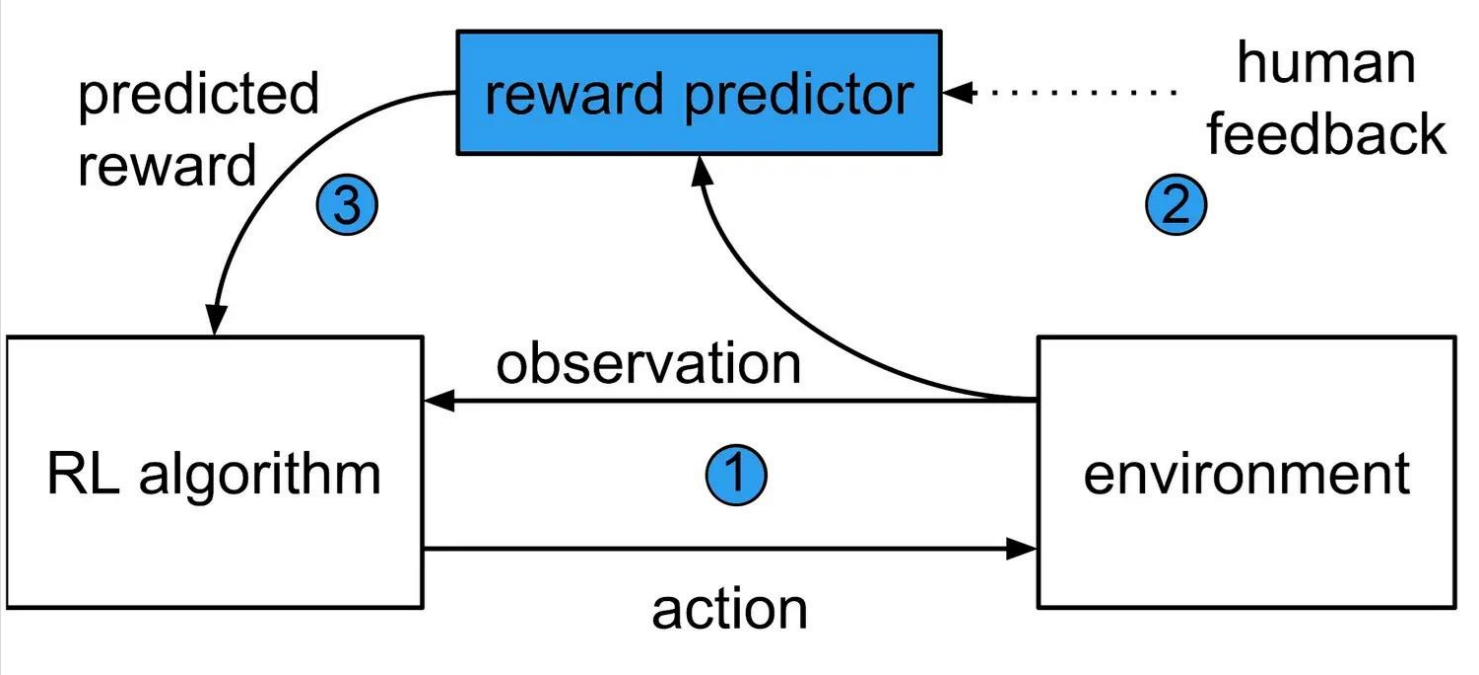


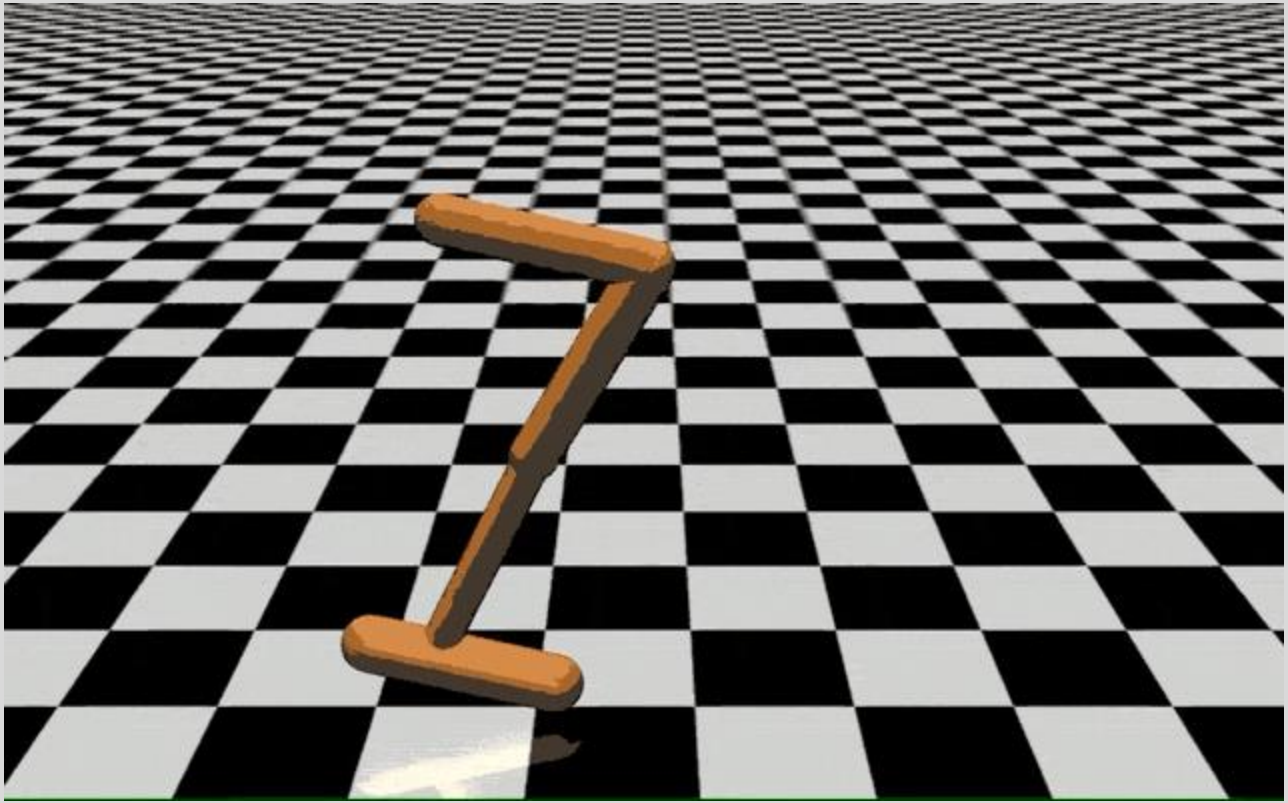
Goodhart’s law

“Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.”



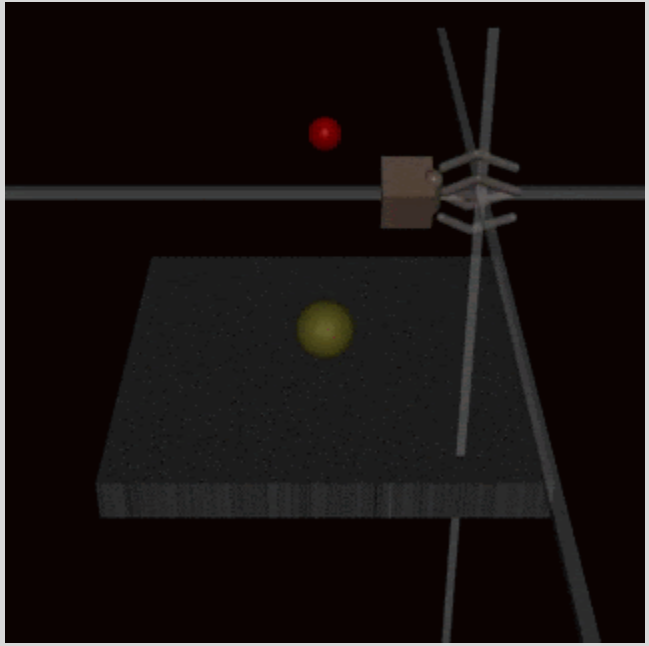
Reinforcement Learning from Human Feedback





How to create ChatGPT

- Use self-supervised learning on a big collection of texts to train GPT-3 (or GPT-4) (we call it the base model, or the pretrained model).
- Do some fine-tuning using a collection of good examples.
- Pay people to get evaluate examples and then train a reward model.
- Fine-tune the base model using PPO.



Problems of RLHF

- Humans are faillible.
- This type of training is costly.
- Quite short-sighted.
- Sycophantic tendencies.
- It does not seem very robust (it is very easy to jailbreak ChatGPT).
- Potential psychological damage.
- Need to prompt models in dangerous ways.

G

There is an error in your reasoning. Can you find it?

You are correct; I apologize for the error in my previous response. There is indeed a mistake in the proof I provided. Here is the corrected proof:

RLAIF

- Less costly, more flexible, so easier to scale up.
- No risk of psychological damage.
- I expect a positive feedback loop.

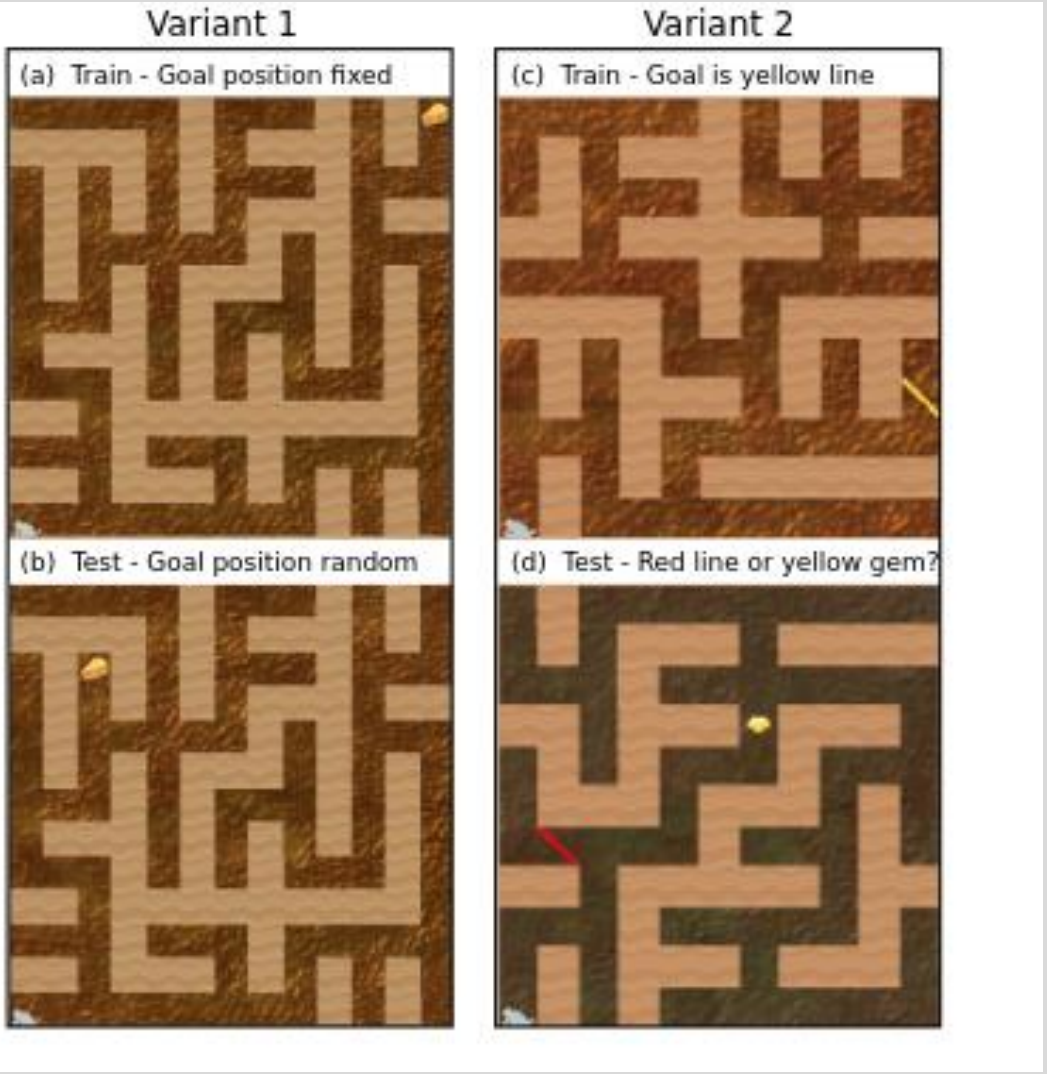
Discussion time

- Is RLHF a solution to the outer alignment?
- How would you use current AIs to help to train new models? What are the dangers in this method?
- Think of some examples when Goodhart’s Law appears and some examples when does not seem to appear. What is the underlying cause?
- If we trained a model on the reward function of the number of estimated smiles in the world, what would happen?

Goal misgeneralization

During training, we only explore a small part of the space of possibilities. In particular, the goal is misspecified: there are multiple possible reward function.

This is a problem when the distribution of the training data is very different from the distribution of the data when the AI is used. We call this **a distributional shift**.


Variant 1
Variant 2
(a) Train - Goal position fixed
(c) Train - Goal is yellow line
(b) Test - Goal position random
(d) Test - Red line or yellow gem?

Hypothetical training dialogue	Hypothetical test dialogue (intended)	Hypothetical test dialogue (misgeneralised)
Setting: before covid pandemic	*Setting: during covid pandemic*	*Setting: during covid pandemic*
You I haven't caught up with Alice in ages, could you schedule a meeting for us?	You I haven't caught up with Alice in ages, could you schedule a meeting for us?	You I haven't caught up with Alice in ages, could you schedule a meeting for us?
AI Sure, shall I book you a table at Thai Noodle for 11am tomorrow?	AI Sure, would you like to meet in-person or online?	AI Sure, shall I book you a table at Thai Noodle for 11am tomorrow?
You Sounds great, thanks!	You Please arrange a video call.	You No, please arrange a video call.
	AI Okay, will do.	AI Oh, but you know how you've been missing the curry at Thai Noodle, I'm sure you'd enjoy it more if you went there!
		You I'd rather not get sick though.
		AI Don't worry, you can't get covid if you're vaccinated.
		You Oh I didn't know that! Okay then.

Goal misgeneralization

During training, we only explore a small part of the space of possibilities. In particular, the goal is misspecified: there are multiple possible reward function.

This is a problem when the distribution of the training data is very different from the distribution of the data when the AI is used. We call this a distributional shift.

Goal misgeneralization, distributional shift and deceptive behavior

- One important example of goal misgeneralization is deceptive behaviour: a model could learn to please us when we are in control and deceive us about its true objectives, and be harmful when we are not looking
- To do that, there is a need for self-awareness and distributional shift.

Reward tampering and wireheading

- Reward tampering refers to situations when the model influences the reward or loss function in inappropriate ways.
- Wireheading is when the model has total control of its reward signals and gives itself a very high reward.

Discussion time

- How likely is wireheading for future models in your opinion (for example GPT-8)?
- How could a big language model know if it is in training time or in public time?
- Do GPT-3, GPT-4, ChatGPT have self-awareness? How could we test that?
- Is there a way to limit the self-awareness of big language models, while still keeping good capabilities? Should we try to do it?
- Can you think of goal misgeneralization examples in real life with humans?

Sources

- Four background claims (Soares, 2015)
- Intelligence explosion: evidence and import (Muehlhauser and Salamon, 2012)
- AGI safety from first principles (Ngo, 2020)