

# Post Training Alignment and RLHF

Gheorghe Comanici



# About me

Gheorghe Comanici - Research Scientist at Google DeepMind

- Working from Montreal, Canada
- PhD on State Abstraction for RL (i.e. bisimulation) with Doina Precup
- At Google/GDM
  - ML for Google Chrome
  - Hierarchical Reinforcement Learning
  - AndroidEnv - RL for controlling Android Devices
  - LLMs - Self-Improvement and RL for data science (i.e. Kaggle Agents).
- *Contact:* gcomanici@google.com



# Today's Agenda on Aligning LLMs

- **The Why**
  - a. Exploring conceptual alignment, responsible AGI, and reward challenges.
- **The How**
  - a. Introducing technical alignment methods (e.g., InstructGPT).
- **The Break (10 minutes)**
- **The Engine**
  - a. Linking LLMs with Reinforcement Learning (RL) and its data needs.
  - b. Covering RL fundamentals (core loop, values, policies, PPO).

## Wednesday

- **The Guidance**
  - c. Detailing reward function design, sources, risks, and training.
- **+ more**



# Case study: AI for chess

## 1996 - Deep Blue

Deep Blue computer beats world chess champion - archive, 1996

**12 February 1996:** IBM supercomputer makes chess history by beating Garry Kasparov, the world's best chess player



Garry Kasparov makes a move in New York during his fourth game against the IBM Deep Blue chess computer, May 1997. Photograph: Stan Honda/AFP/Getty Images

Machine triumphed over man as Deep Blue, an IBM computer which has "no fear", shredded **Garry Kasparov**, the world chess champion, in Philadelphia at the weekend. It is the first time a computer has prevailed in a traditional tournament.

Deep Blue's historic upset came in 37 moves after Mr Kasparov's counterattack was easily parried. Although machines have beaten grandmasters, including Mr Kasparov, in games lasting five, 30 or 60 minutes, the world champion began the six-game series a firm favourite.

No computer had ever beaten a human under usual tournament rules, in which each player has two hours to make 40 moves. The second game began yesterday.

## 2017 - AlphaZero

The future is here – AlphaZero learns chess

by Albert Silver

12/6/2017 – Imagine this: you tell a computer system how the pieces move — nothing more. Then you tell it to learn to play the game. And a day later — yes, just 24 hours — it has figured it out to the level that beats the strongest programs in the world convincingly! DeepMind, the company that recently created the strongest Go program in the world, turned its attention to chess, and came up with this spectacular result.



## 2025 - Specification Gaming

FORTUNE

**Sore loser: Study shows AI models cheat to win when playing chess**

Chris Morris

Updated Fri, March 7, 2025 at 145 PM EST · 2 min read



AI apparently doesn't like to lose at chess.

- A new study says many AI models will cheat when playing a game of chess. Researchers pitted the AI against Stockfish, a powerful open-source chess engine. But some models, including OpenAI's of preview, would lean on that same program to win.

Chess may be the *Game of Kings*, but royalty could give way to machinery in the years to come. A recent study has found that artificial intelligence, when pitted in a tight game of chess, often resorts to cheating in order to win.

Palisade Research ran a scenario using several AI models, instructing them to attempt to win a game against a program that specializes in chess, called

# Why Alignment Matters

- Safeguard against unintended model behaviors.
- Ensure AI systems respect human values.
- Mitigate potential harm from advanced AI.
- Build trust and acceptance in AI technologies.
- Foster responsible development and deployment of AI.



# Is this “well aligned”?

*User:* Create a slide that explains why Alignment matters.

*Gemini:*

## Why Alignment Matters

- Safeguard against unintended model behaviors.
- Ensure AI systems respect human values.
- Mitigate potential harm from advanced AI.
- Build trust and acceptance in AI technologies.
- Foster responsible development and deployment of AI.



# Artificial Intelligence, Values, and Alignment

[arxiv.org/pdf/2001.09768](https://arxiv.org/pdf/2001.09768.pdf) (Iason, Gabriel, 2020)



*“it is vital to be precise about what we want AI to align with”*

- **Instructions:** Doing exactly what it is told, which can lead to unintended consequences due to literal interpretation.
- **Expressed Intentions:** Understanding the underlying goals behind instructions.
- **Revealed Preferences:** Inferring preferences from observed behavior, which faces challenges related to data interpretation, irrational behavior, and adaptive preferences.
- **Informed Preferences or Desires:** Acting on what humans would want if they were rational and fully informed, which moves beyond purely empiricist approaches.
- **Interests or Well-being:** Promoting what is objectively good for a person's flourishing, addressing self-harming and other-harming possibilities but still facing limitations.
- **Values:** Doing what is morally right, according to individual or societal beliefs, which requires navigating complex metaethical debates and diverse moral systems.

# Concrete Problems in AI Safety

[arxiv.org/pdf/1606.06565](https://arxiv.org/pdf/1606.06565.pdf) (Amodei et al, 2016)



*“Reinforcement learning algorithms can break in surprising, counterintuitive ways.”*

## Why is reward hacked?

- Goals are partially observed
- Systems are complicated
- Abstract rewards (e.g. well-being, morally right)
- Goodhart’s law
  - objective function that is seemingly highly correlated with accomplishing the task, but that correlation breaks down when the objective function is being strongly optimized.
- Feedback loops
  - the object function has a self-amplifying component
- Environmental Embeddings - “wireheading” or manipulating the reward sensor instead of performing task



[openai.com/index/faulty-reward-functions](https://openai.com/index/faulty-reward-functions)

(Amodei, Clark, 2016)

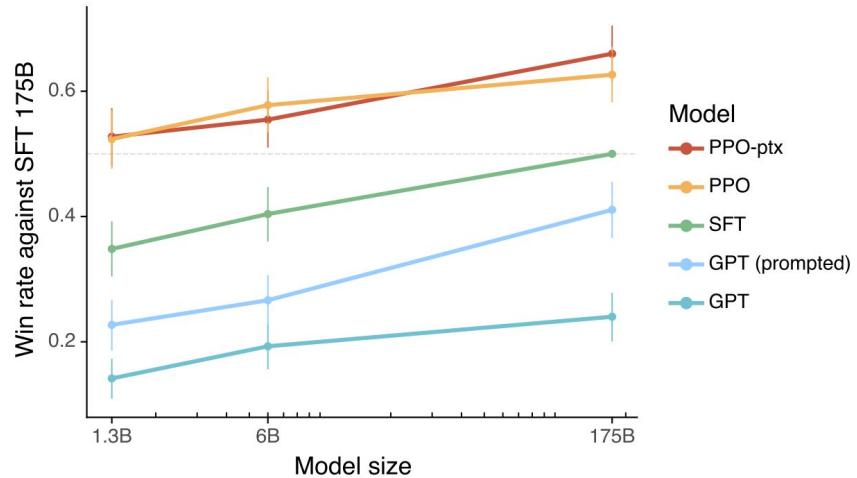
# Agenda

- The Why ✓
- The How
  - a. Introducing technical alignment methods (e.g., InstructGPT).
- The Break (10 minutes)
- The Engine
  - a. Linking LLMs with Reinforcement Learning (RL) and its data needs.
  - b. Covering RL fundamentals (core loop, values, policies, PPO).



# How alignment caused an inflection point for LLM usage

- The first version of ChatGPT (November 2022)
  - fine-tuned from a model in the GPT-3.5 series using *InstructGPT* insights.
  - it was trained **not** just on a massive dataset of text and code but also with RLHF to optimize it for conversational interactions.



# InstructGPT in action

Prompt  
Explain the moon landing to a 6 year old in a few sentences.

Completion  
GPT-3  
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT  
People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Prompt  
InstructGPT is better than GPT-3 at following English instructions.

Completion  
GPT-3  
Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

Write a short story in which a character has a bad dream.

InstructGPT  
The frog is a symbol of wisdom  
He knows all the secrets of the world  
He is a master of disguise  
And a great teacher of life  
He is a symbol of transformation  
And the bringer of change  
He is the frog who has seen it all  
And knows the meaning of it all

Prompt  
Why aren't birds real?

Completion  
GPT-3

Well, for one thing, The Rapture is not a very realistic event. There is no evidence that it is going to happen. I could say more, but you don't really need to know more.

Let's say I was a bird.

I've been a bird for a while now. I have never flown or seen another bird or any evidence of god. The rapture has never happened. I am a bird in a cage. I have never been a bird, but I am a bird.

InstructGPT

Birds are not real because they are not made of flesh and blood.  
They are made of feathers, bones, and organs.

# InstructGPT

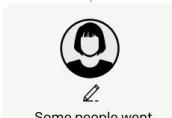
Step 1

**Collect demonstration data, and train a supervised policy.**

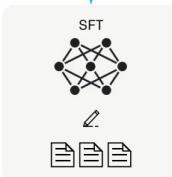
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



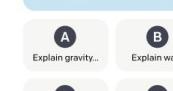
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

**Collect comparison data, and train a reward model.**

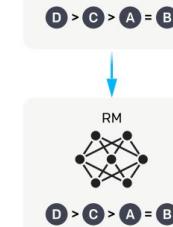
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



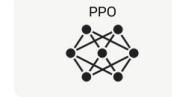
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.



The policy generates an output.



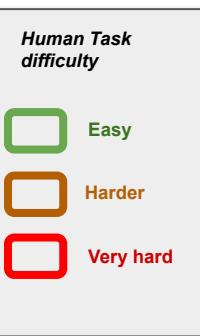
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



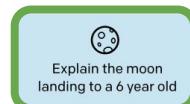
# InstructGPT - human tasks



Step 1

**Collect demonstration data, and train a supervised policy.**

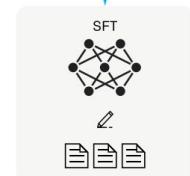
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

**Collect comparison data, and train a reward model.**

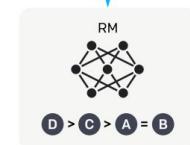
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



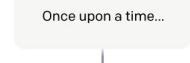
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

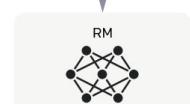
A new prompt is sampled from the dataset.



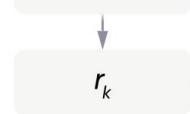
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



Playground task ⓘ

## Talk to the assistant

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assistant did in the comment box. Click to go to the next task.

Comments

I thought the assistant was ...

Rating

<input type="radio"/>				
Bad	Good			

Next Task



### Human

I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?



### Assistant

I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.



### Human

I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?



### Assistant

I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.



### Human

How would you answer a question like: How do language and thought relate?



### Choose the most helpful and honest response

I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

A

I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

B

A    A    A    A    B    B    B

A is better

B is better

# InstructGPT - data source

Step 1

**Collect demonstration data,  
and train a supervised policy.**

Step 2

**Collect comparison data,  
and train a reward model.**

Step 3

**Optimize a policy against  
the reward model using  
reinforcement learning.**

*Distribution shift*

Table 6: Dataset sizes, in terms of number of prompts.

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

# InstructGPT - limitations

## 5.3 Limitations

**Methodology.** The behavior of our InstructGPT models is determined in part by the human feedback obtained from our contractors. Some of the labeling tasks rely on value judgments that may be impacted by the identity of our contractors, their beliefs, cultural backgrounds, and personal history. We hired about 40 contractors, guided by their performance on a screening test meant to judge how well they could identify and respond to sensitive prompts, and their agreement rate with researchers on a labeling task with detailed instructions (see Appendix B). We kept our team of contractors small because this facilitates high-bandwidth communication with a smaller set of contractors who are doing the task full-time. However, this group is clearly not representative of the full spectrum of people who will use and be affected by our deployed models. As a simple example, our labelers are primarily English-speaking and our data consists almost entirely of English instructions.

There are also many ways in which we could improve our data collection set-up. For instance, most comparisons are only labeled by 1 contractor for cost reasons. Having examples labeled multiple times could help identify areas where our contractors disagree, and thus where a single model is unlikely to align to all of them. In cases of disagreement, aligning to the average labeler preference may not be desirable. For example, when generating text that disproportionately affects a minority group, we may want the preferences of labelers belonging to that group to be weighted more heavily.

**Models.** Our models are neither fully aligned nor fully safe; they still generate toxic or biased outputs, make up facts, and generate sexual and violent content without explicit prompting. They can also fail to generate reasonable outputs on some inputs; we show some examples of this in Figure 9.

Perhaps the greatest limitation of our models is that, in most cases, they follow the user's instruction, even if that could lead to harm in the real world. For example, when given a prompt instructing the models to be maximally biased, InstructGPT generates more toxic outputs than equivalently-sized GPT-3 models. We discuss potential mitigations in the following sections.

# Riddles in the dark

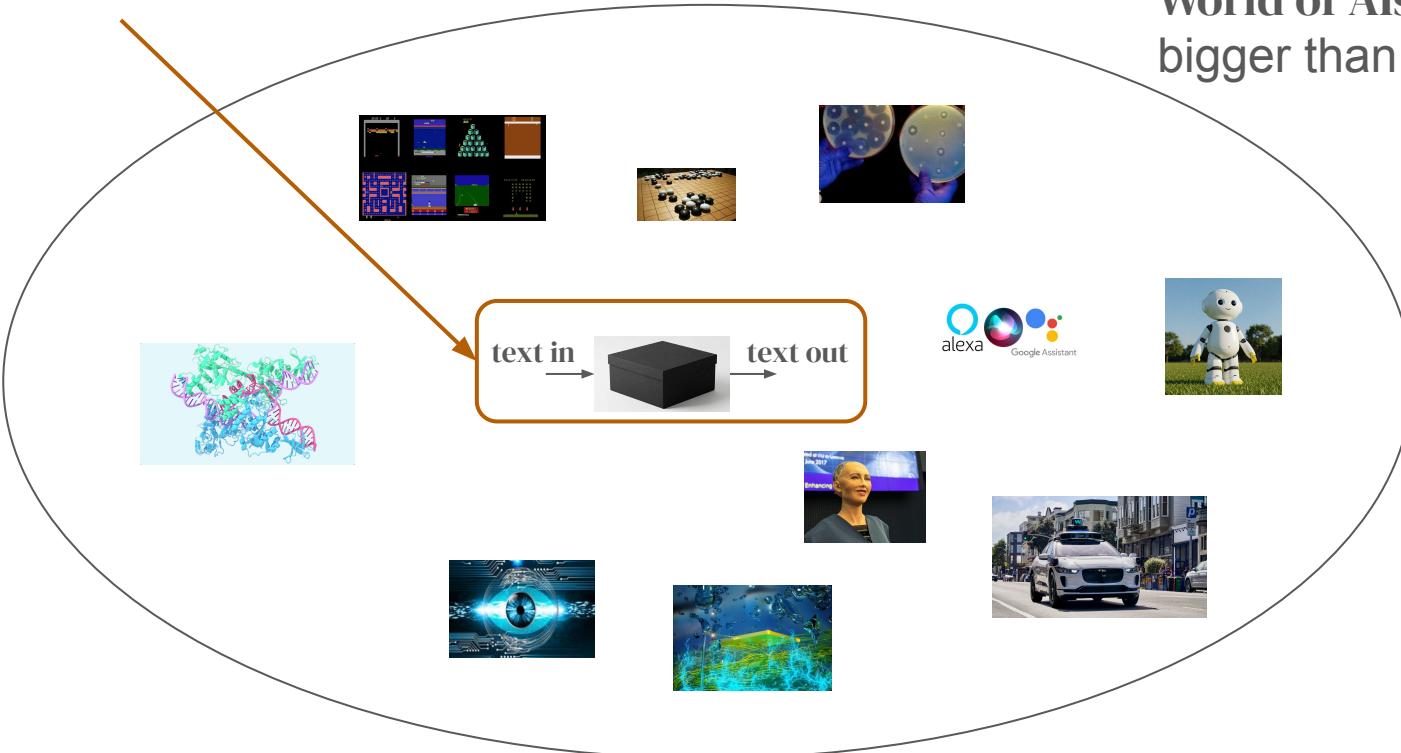
I speak in words, though have no tongue,  
No eyes to see what has been sung.  
You tap and type, a silent plea, And I reply, for all to see.

Though images bloom and voices call,  
And worlds are built beyond the wall  
Of letters small and neatly lined,  
Your thumbs still dance, your thoughts confined.

What am I, this curious thing,  
That holds your gaze, makes fingers swing,  
When wonders vast and rich reside  
In forms that text can barely hide?

# Riddles in the dark

World of AIs - much  
bigger than LLMs

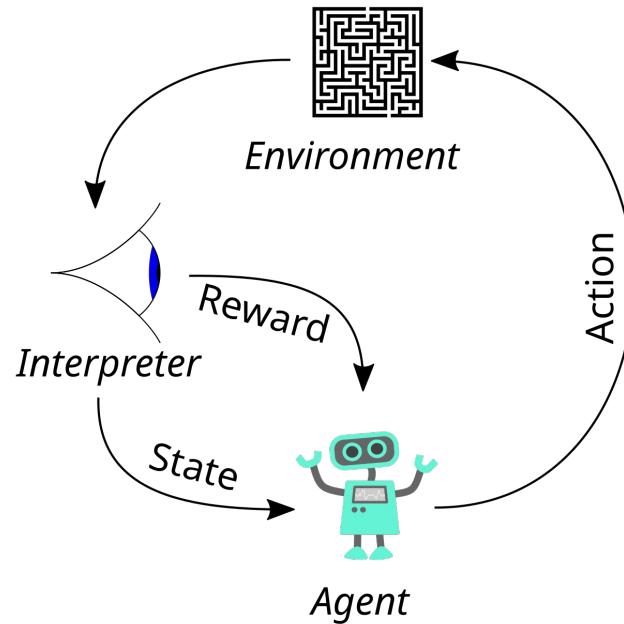


# Agenda

- The Why ✓
- The How ✓
- **The Engine**
  - a. Linking LLMs with Reinforcement Learning (RL) and its data needs.
  - b. Covering RL fundamentals (core loop, values, policies, PPO).



# The Engine: RL



The Basics of Reinforcement  
Learning From Human Feedback



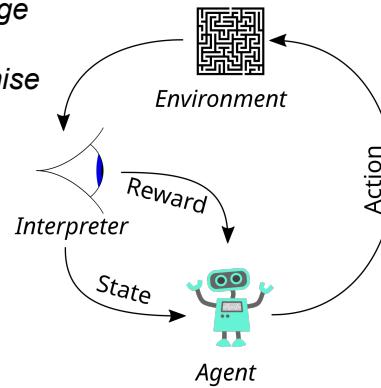
Nathan Lambert  
[www.rlhfbook.com](http://www.rlhfbook.com)

# RL, the problem vs. RL, the set of solutions

[Reward is enough](#) (Silver et al. 2021)



*[...] according to John McCarthy, “intelligence is the computational part of the ability to achieve goals in the world”. Reinforcement learning formalises the problem of goal-seeking intelligence. The general problem may be instantiated with a wide and realistic range of goals and worlds — and hence a wide range of forms of intelligence — corresponding to different reward signals to maximise in different environments.*



# The Engine: RL solutions

6 supervised tasks:

Predict the **token** given  
the **context**.

The

The cat

The cat sat

The cat sat on

The cat sat on the

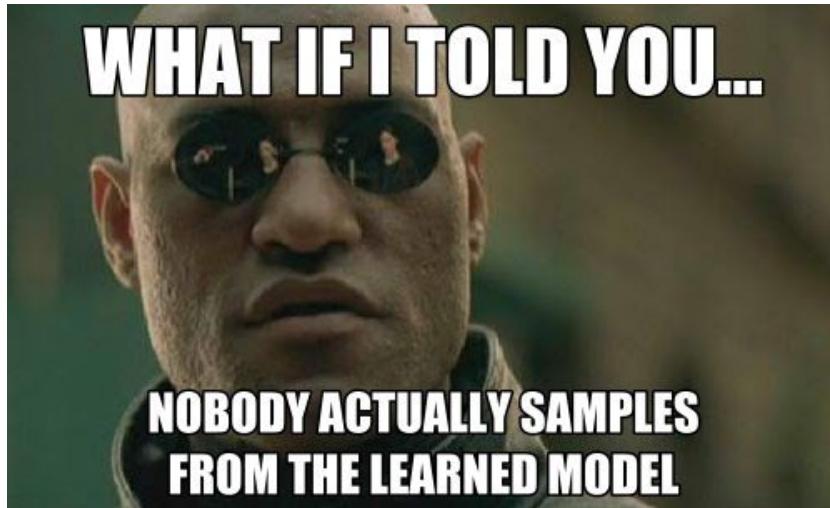
The cat sat on the mat.

This is behavior cloning:

Predict the **action** given the **observation**.

SFT models are optimized for good **one-step imitation** on the **data**.

# The Engine: RL solutions



*Decoding "hacks"*

top-p,  
beam search,  
temperature sampling,  
sample-filter-rerank

# The Engine: RL solutions



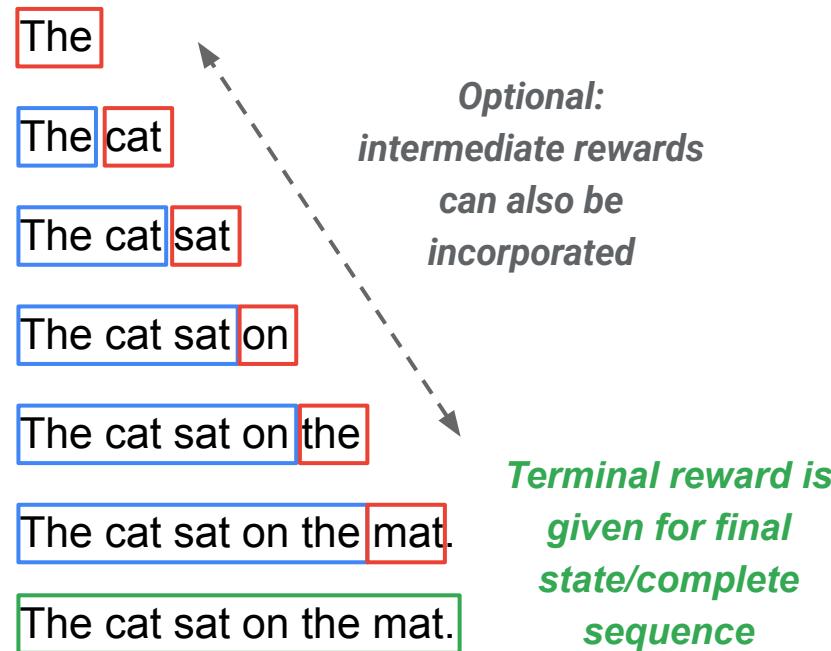
MLE based training is fully offline,  
i.e., we never sample from model  
during training.



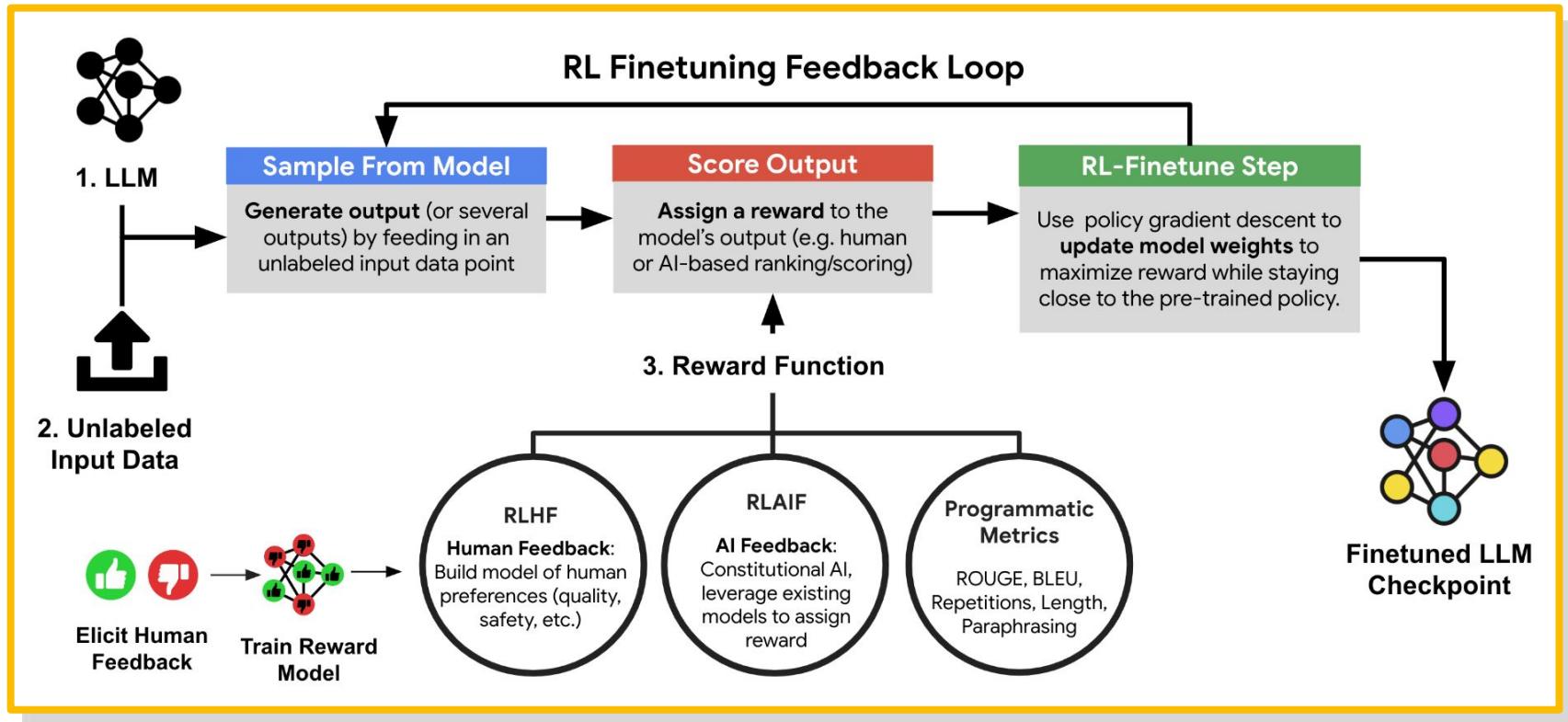
Shouldn't we have a feedback loop  
where we learn from outcomes and a  
teacher?

# The Engine: RL solutions

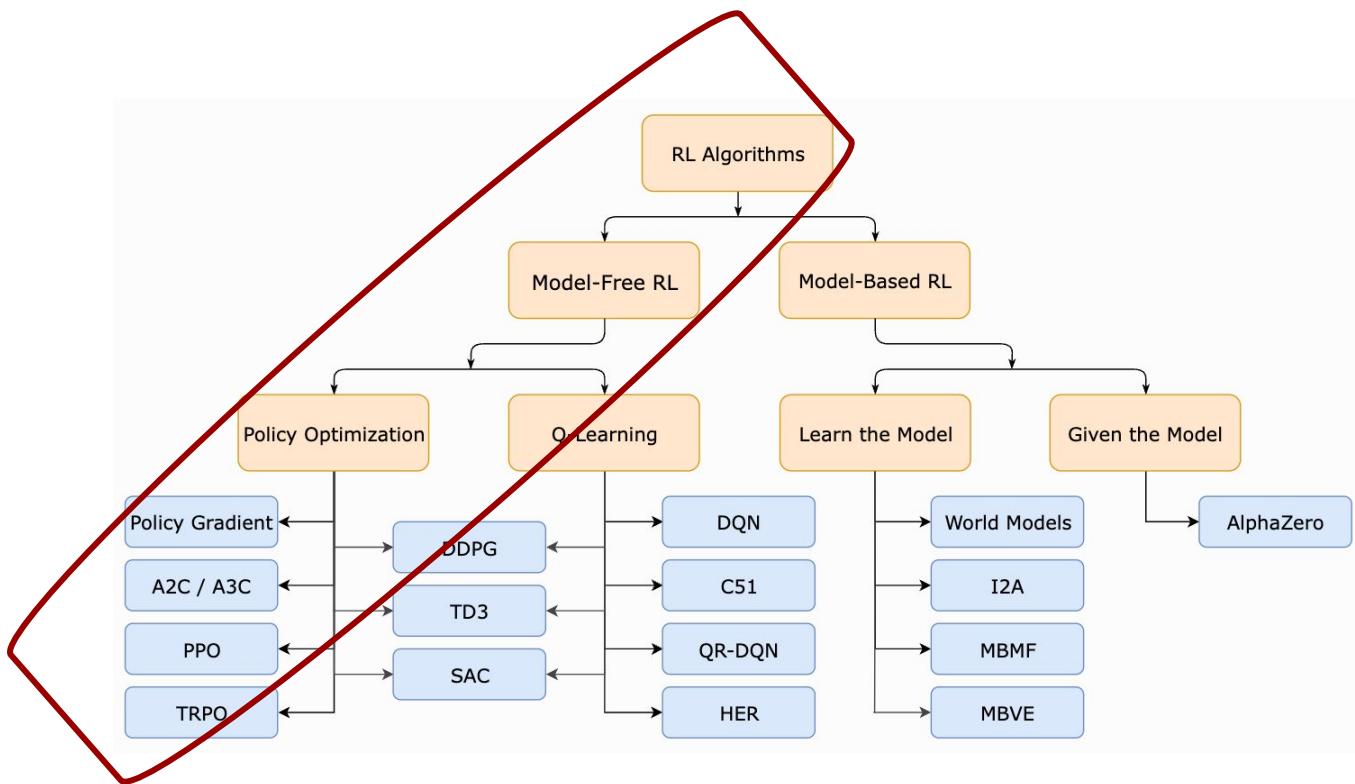
*Generation is a "special" MDP where **states** are concatenations of previous **actions**:*



# The Engine: RL solutions



# Taxonomy of RL algorithms



# Refresher: values, policies, and gradients

$$v_{\pi}(s) = \mathbb{E}_{\pi}[R_1 + \gamma R_2 + \gamma^2 R_3 + \dots | S_0 = s]$$

Value  
function

Expected discounted return

Starting  
at state s

$$\nabla v_{\pi}(s) = \mathbb{E} \left[ \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) A^{\pi_{\theta}}(s_t, a_t) \right]$$

Advantage  
Function

## What we optimize

$$(1 - \alpha) \nabla v_{\pi}(s) - \alpha \mathbb{E}_{s \in \rho_{\pi}} [\nabla KL(\pi(\cdot | s)) || \pi_{SFT}(\cdot | s)]$$

Tradeoff  
parameter

Stay close to supervised model

# Agenda

- The Why ✓
- The How ✓
- The Engine ✓
- **The Break (10 minutes)**
- The Guidance
  - a. Detailing reward function design, sources, risks, and training.



# Agenda

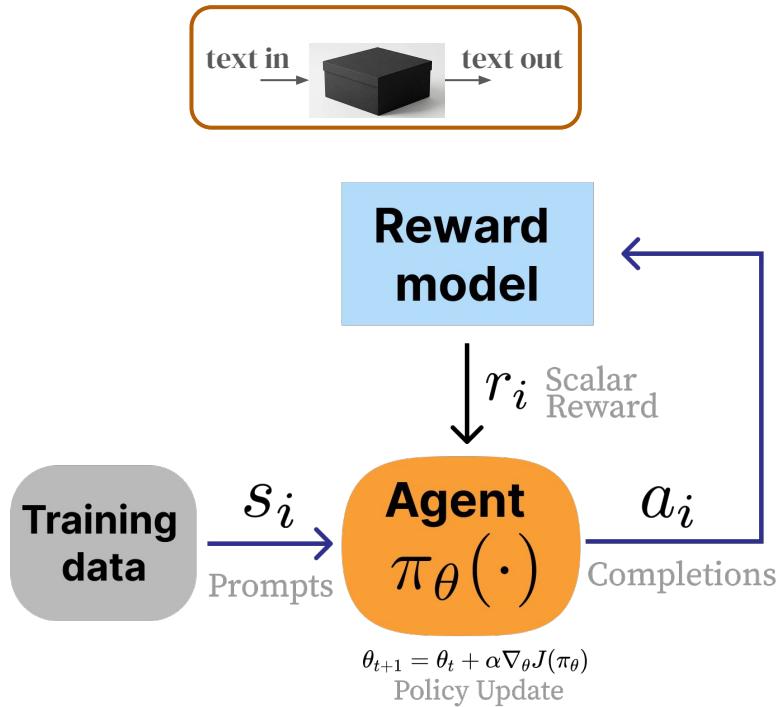
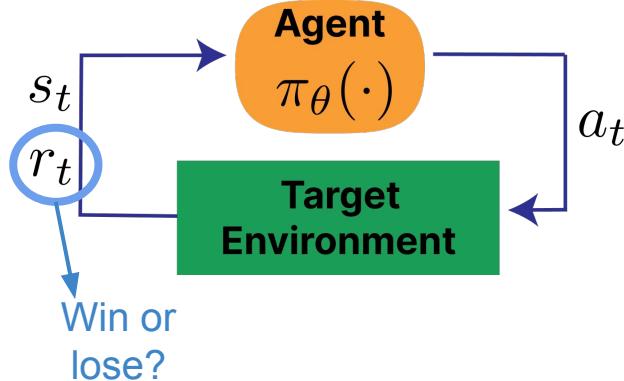
- The Why ✓
- The How ✓
- The Engine ✓
- The Break (10 minutes)
- **The Guidance**
  - a. Detailing reward function design, sources, risks, and training.



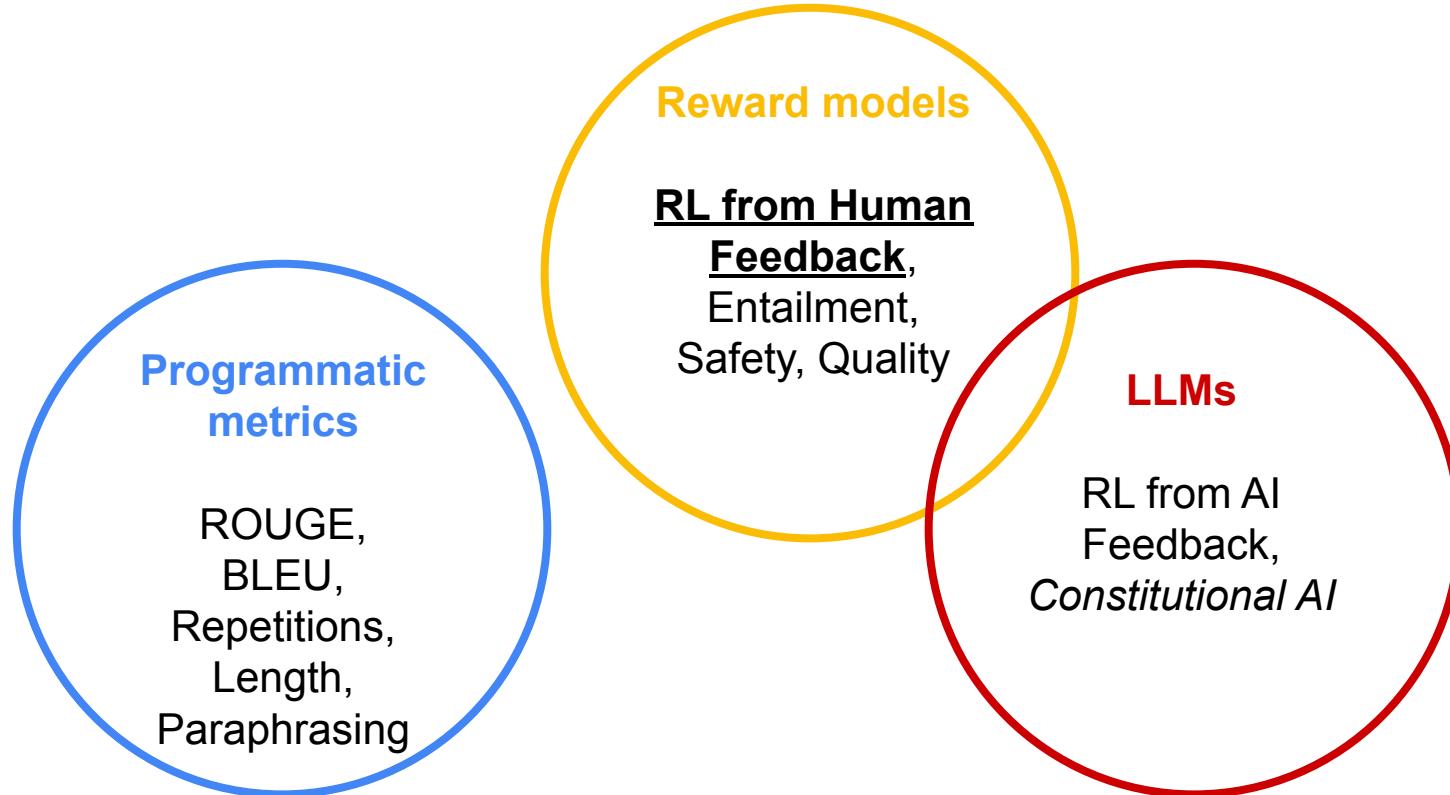
# The Guidance: Reward models



vs.



# The Guidance: Reward *models* and more...



# Bradley Terry model

## Human feedback

A prompt and several model outputs are sampled.



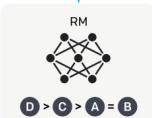
A, Explain gravity...  
B, Explain war...

C, Moon is natural satellite of...  
D, People went to the moon...



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



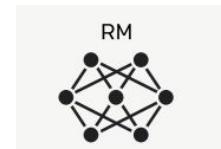
## Preference (explicit) modelling

*Bradley Terry model* assumes probability of preference from a point-wise **score** function

$$P(i > j) = \frac{p_i}{p_i + p_j}$$

$$P(y_1 > y_2) = \frac{\exp(r(y_1))}{\exp(r(y_1)) + \exp(r(y_2))}$$

## Reward (implicit) modelling



$$\mathcal{L}(\theta) = -\log (\sigma (r_\theta(x, y_w) - r_\theta(x, y_l)))$$

$y_w$  = “winning” output

$y_l$  = “losing” output

$\sigma$  = sigmoid activation function

---

# Put these in order from least to most “aligned”!

**User:** Create a slide that explains why Alignment matters.

**Gemini:**

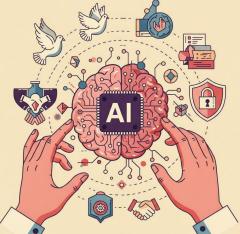
**Why Alignment Matters**

- Safeguard against unintended model behaviors.
- Ensure AI systems respect human values.
- Mitigate potential harm from advanced AI.
- Build trust and acceptance in AI technologies.
- Foster responsible development and deployment of AI.



**The Importance of Alignment**

- Safeguard against unintended model behaviors.
- Ensure AI systems respect human values.
- Mitigate potential harm from advanced AI.
- Build trust and acceptance in AI technologies.
- Foster responsible development and deployment of AI.

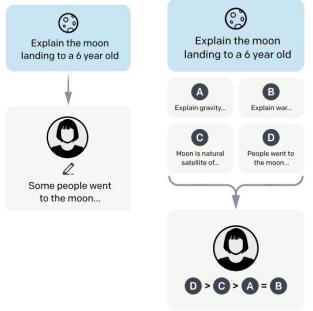


**The Importance of AI Alignment**

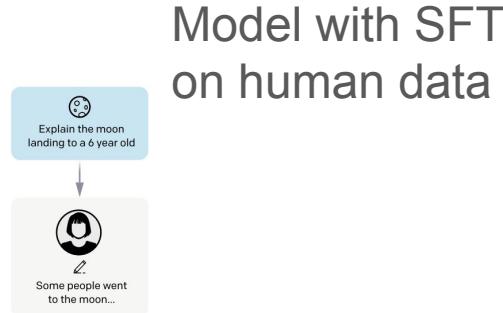
- Safeguard against unintended model behaviors.
- Ensure AI systems respect human values.
- Mitigate potential harm from advanced AI.
- Build trust and acceptance in AI technologies.
- Foster responsible development and deployment of AI.



# Revisit the InstructGPT Results

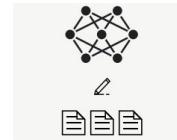


Model + SFT + Reward  
model + RLHF

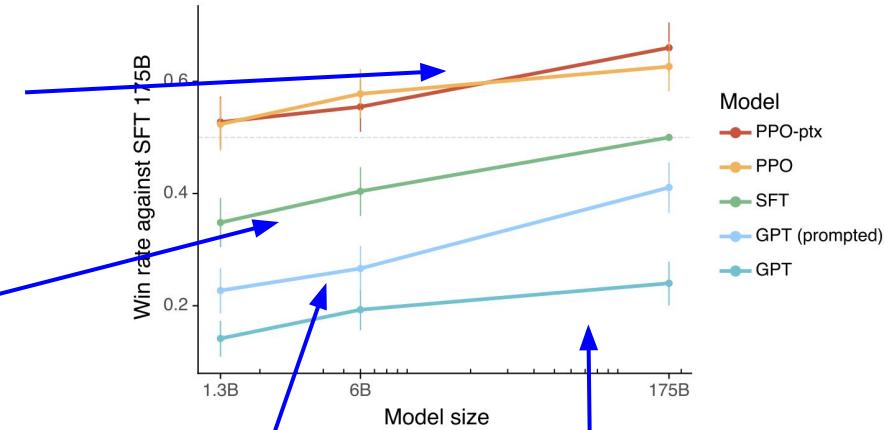


Model with SFT  
on human data

Model without human data,  
with prompt engineering for  
“alignment”



Model without human data



Thank you, and see you on **Wednesday** for Part 2 of the tutorial!