Instruct-GPT: Reinforcement Learning with Human Feedback and Proximal Policy Optimization

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

References

The subsequent slides are taken from the following materials:

- L. Ouyang et al. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155v1
- Rafael Rafailov et al. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. `https://arxiv.org/abs/2305.18290`
- Luis Serrano, Proximal Policy Optimization (PPO). `https://www.youtube.com/watch?v=TjHH_--7l8g`
- Luis Serrano, Re-inforcement Learning with Human Feedback. `https://www.youtube.com/watch?v=Z_JUqJBpV0k`
- Luis Serrano, Direct Preference Optimization (DPO) - How to fine-tune LLMs directly without reinforcement learning `https://www.youtube.com/watch?v=k2pD3k1485A`

RLHF and PPO

- Reinforcement Learning with Human Feedback (RLHF) is a method used for training Large Language Models (LLMs), including chat-GPT and its predecessor Instruct-GPT.
- RLHF in turn is based on gradient-based reinforcement learning method called Proximal Policy Optimization (PPO).
- In PPO two networks are trained simultaneously
  - a value network
  - a policy network

Problems with Prompt-Driven Large Language Models (LLMs)

Misalignment of user intention and model response

- hallucinations: generated content that is nonsensical or unfaithful to the provided source content; "making up facts"
- generating biased text
- toxic language: communication that is harmful to others
- due to different objectives of LLMs ("predict the next word or sentence") and of machine-human interaction ("follow the user's instructions helpfully and safely")

Instruct-GPT: Fine-Tuning, Reward Modelling, and Reinforcement Learning

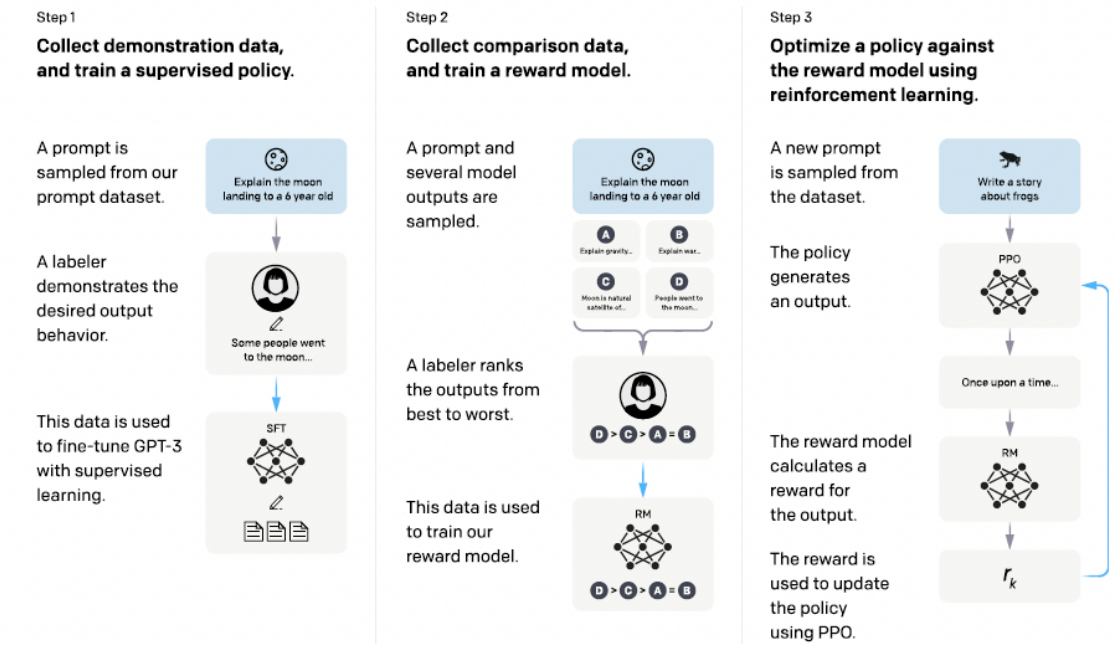


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

Figure taken from: L. Ouyang et al. (2022).

Results of Human Evaluation of Prompt Distributions

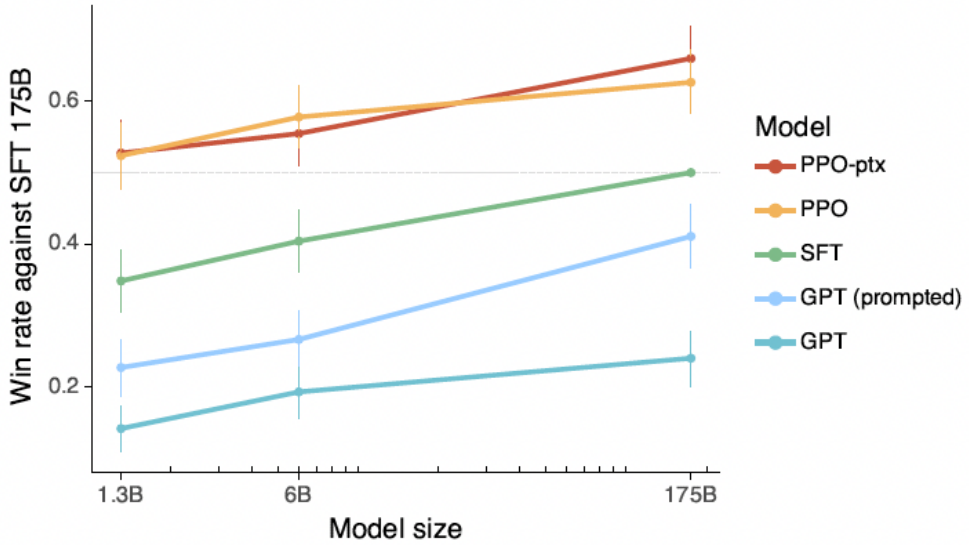


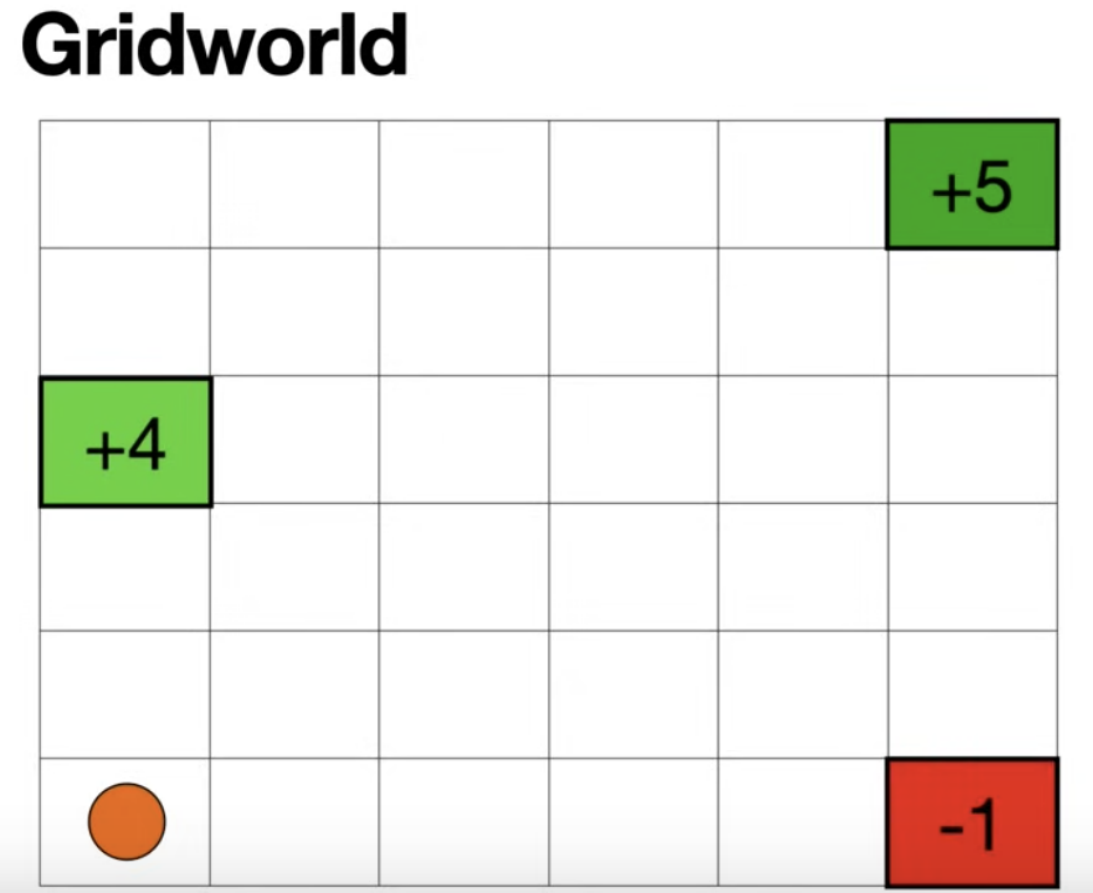
Figure 1: Human evaluations of various models on our API prompt distribution, evaluated by how often outputs from each model were preferred to those from the 175B SFT model. Our InstructGPT models (PPO-ptx) as well as its variant trained without pretraining mix (PPO) significantly outperform the GPT-3 baselines (GPT, GPT prompted); outputs from our 1.3B PPO-ptx model are preferred to those from the 175B GPT-3. Error bars throughout the paper are 95% confidence intervals.

Figure taken from: L. Ouyang et al. (2022). Training language models to follow instructions with human feedback.

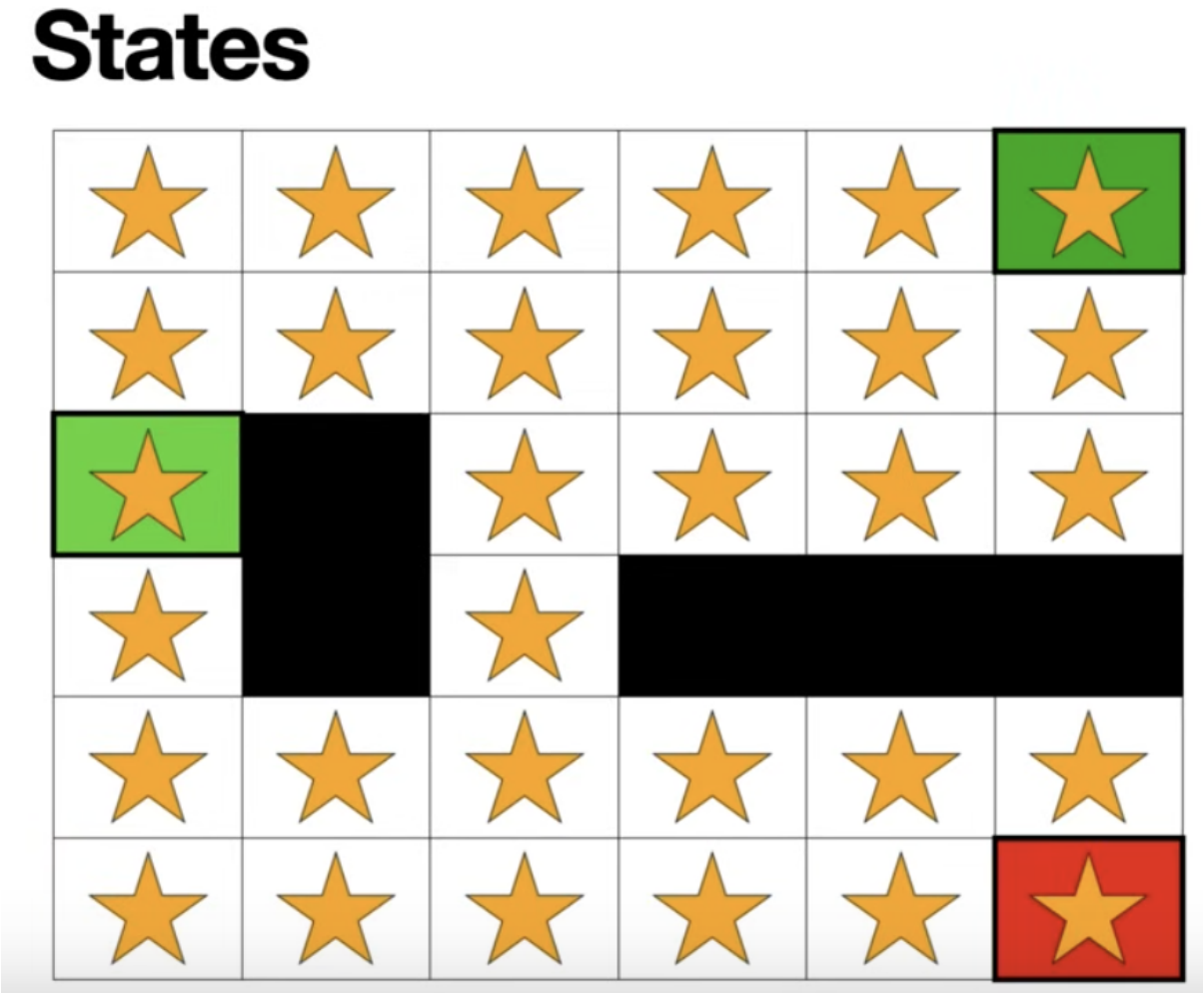
arXiv:2203.02155v1

GridWorld Game

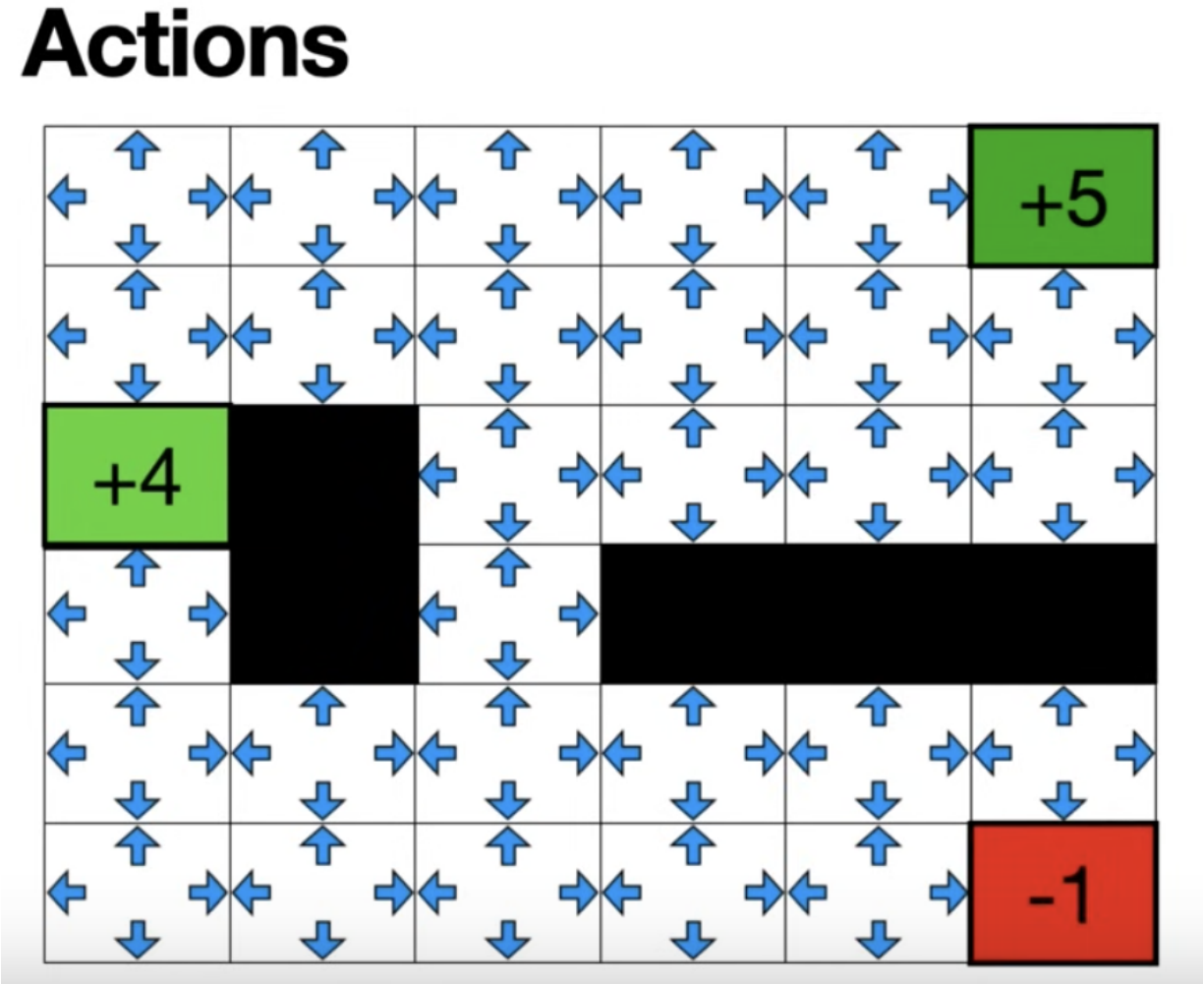
Objective: To get a player to learn how to play the game in an optimal way



GridWorld Game: States

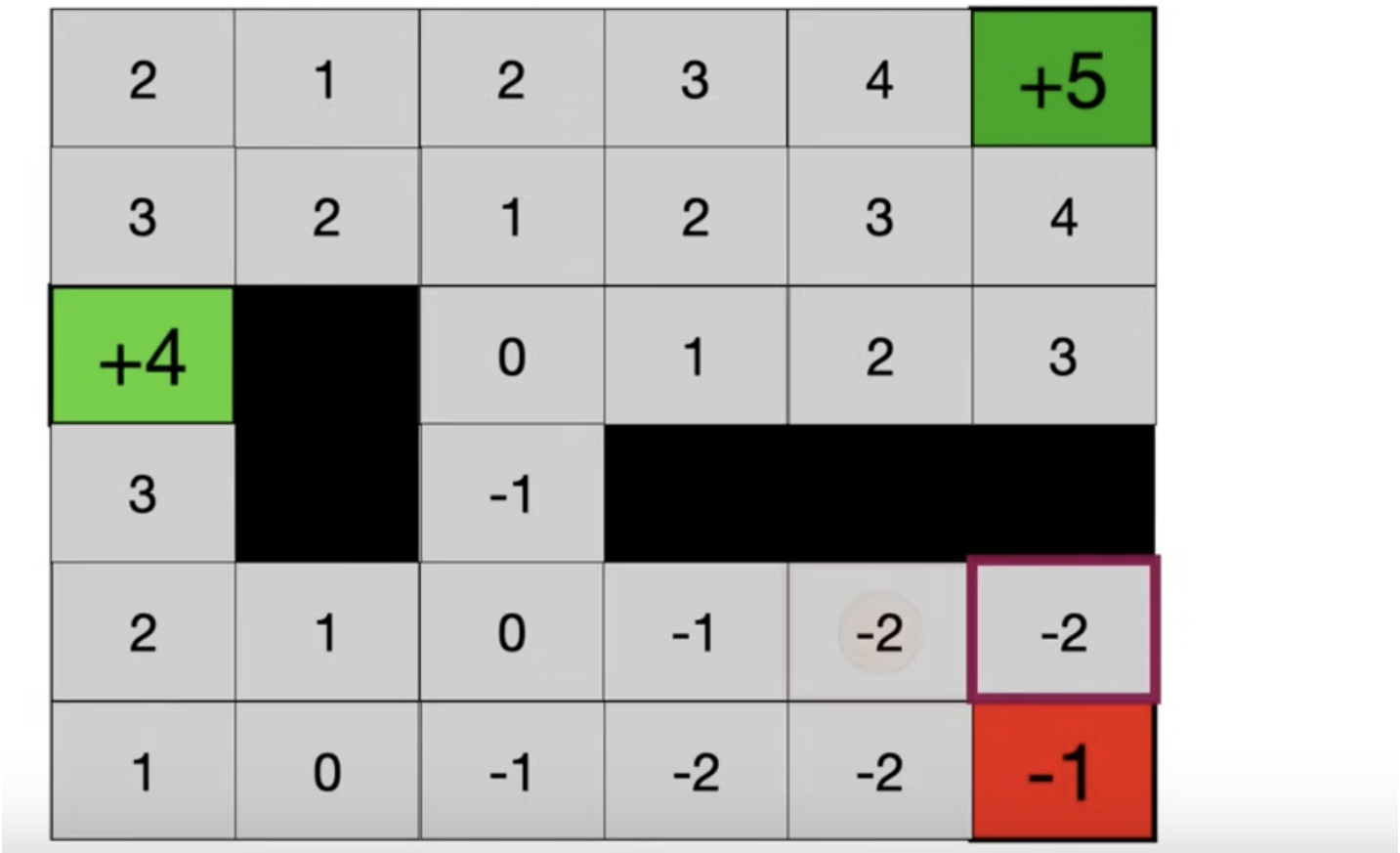


GridWorld Game: Actions

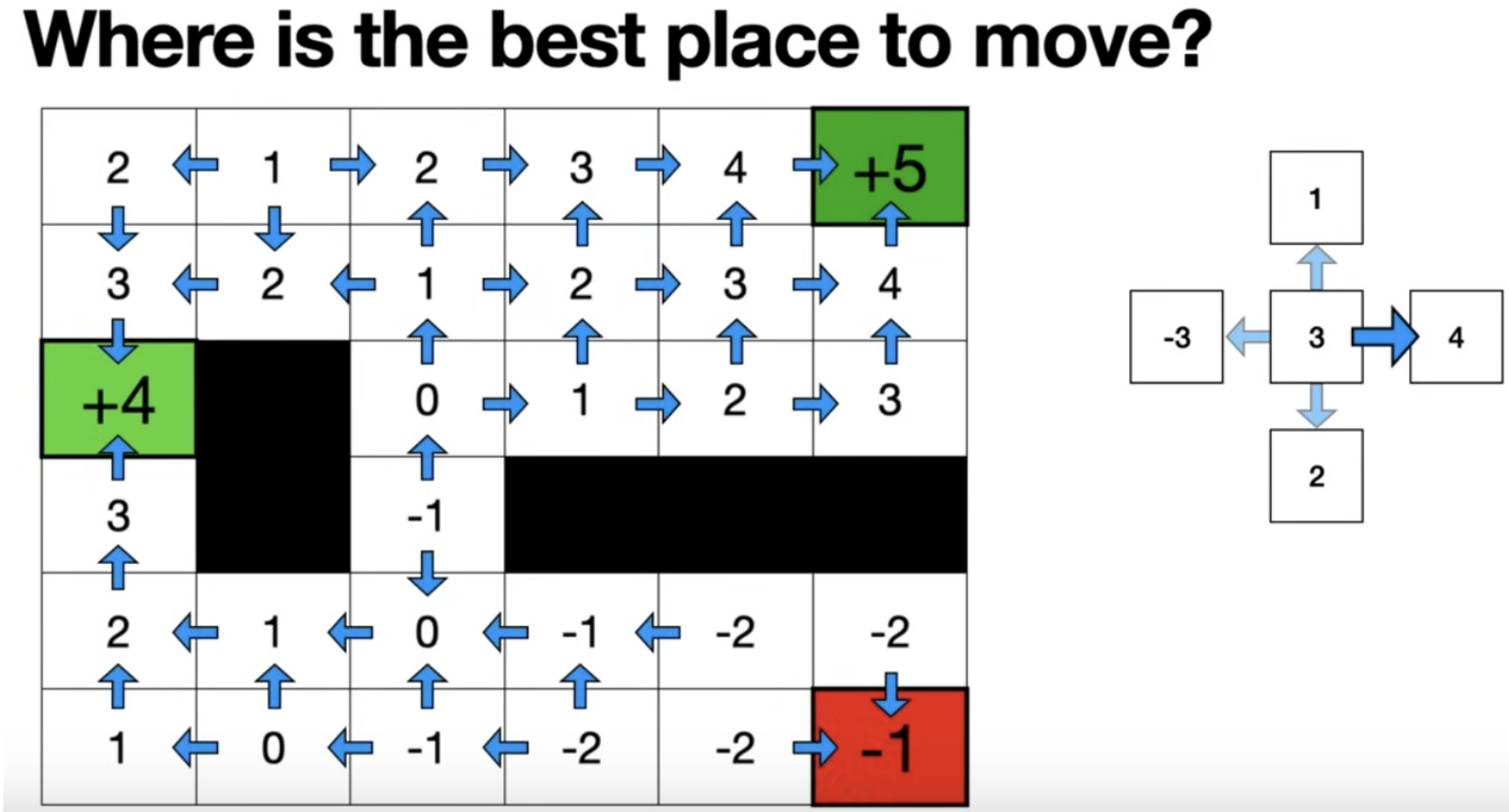


GridWorld Game: Values

How to calculate all the values?



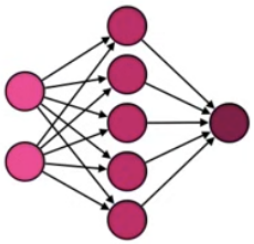
GridWorld Game: Policy

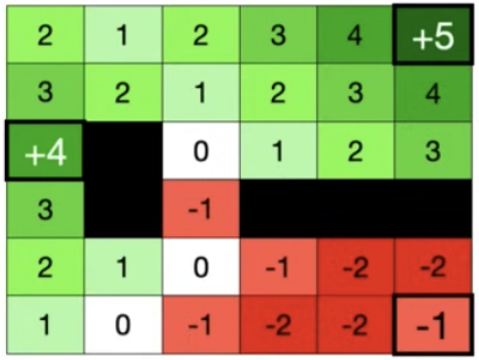


Two neural networks

Value neural network

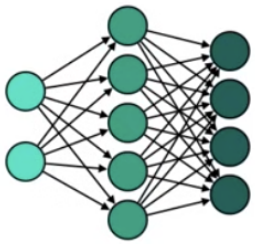
Approximates values

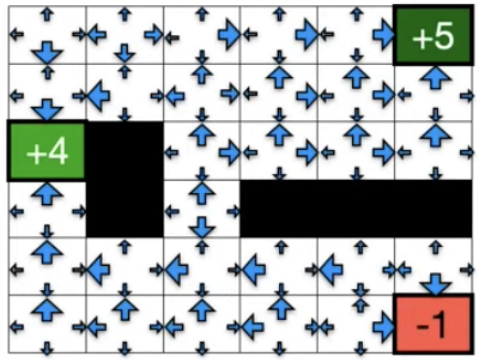




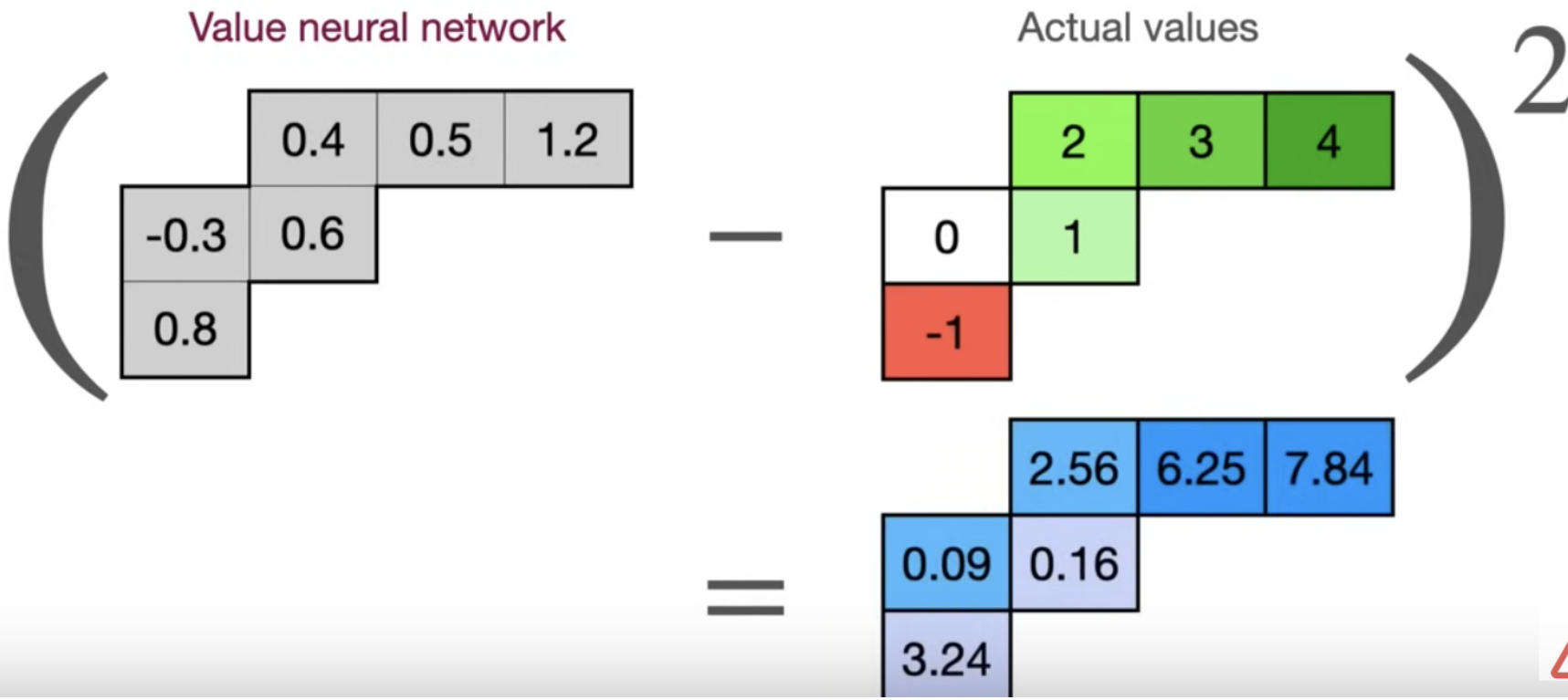
Policy neural network

Approximates policy

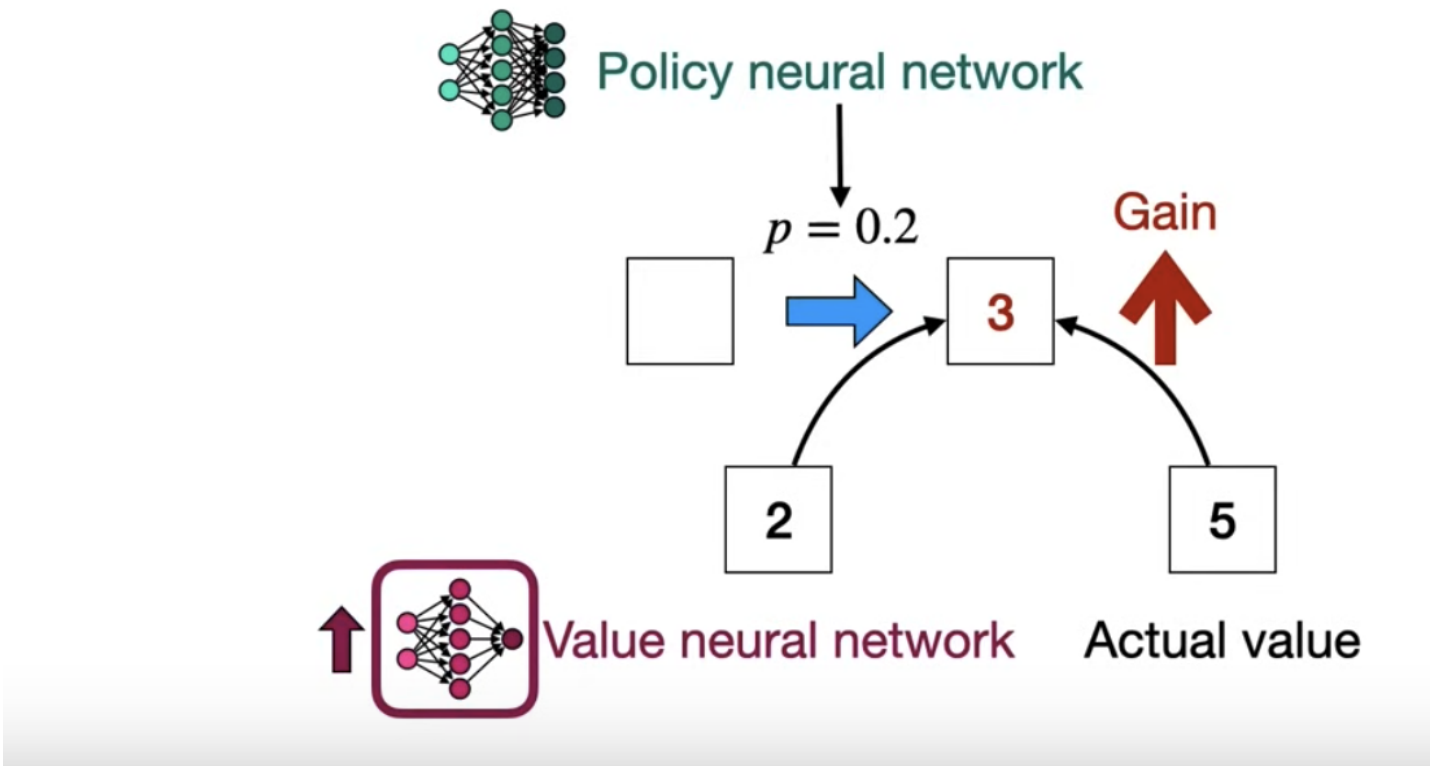




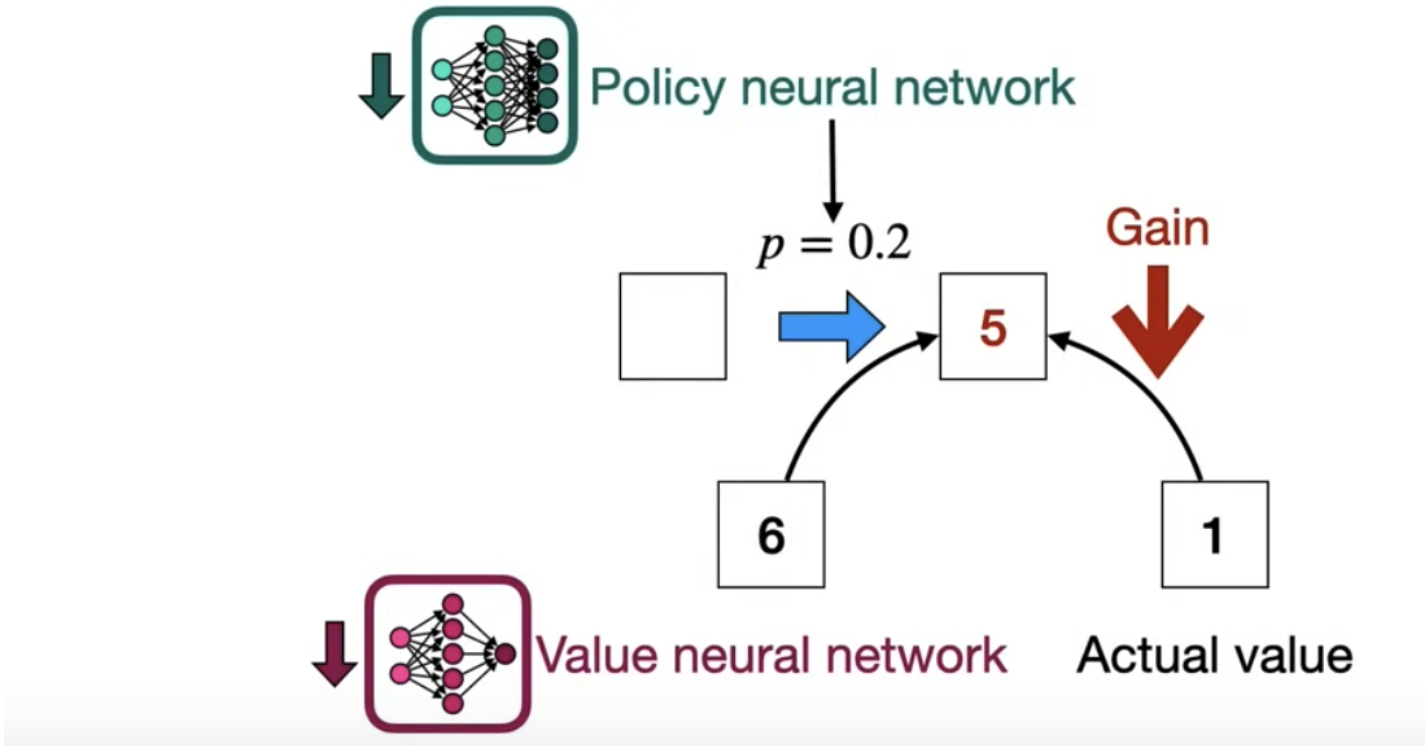
Training the value neural network



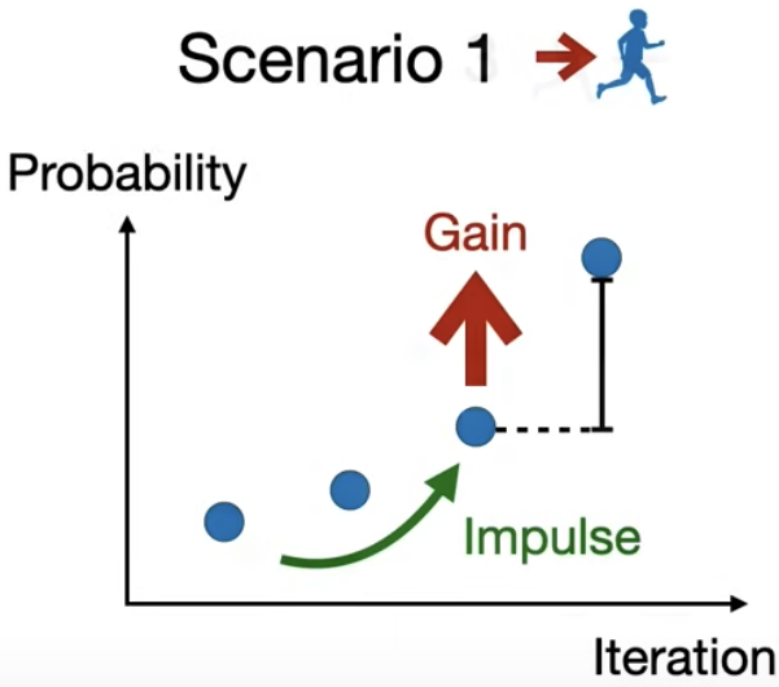
Main idea for training policy

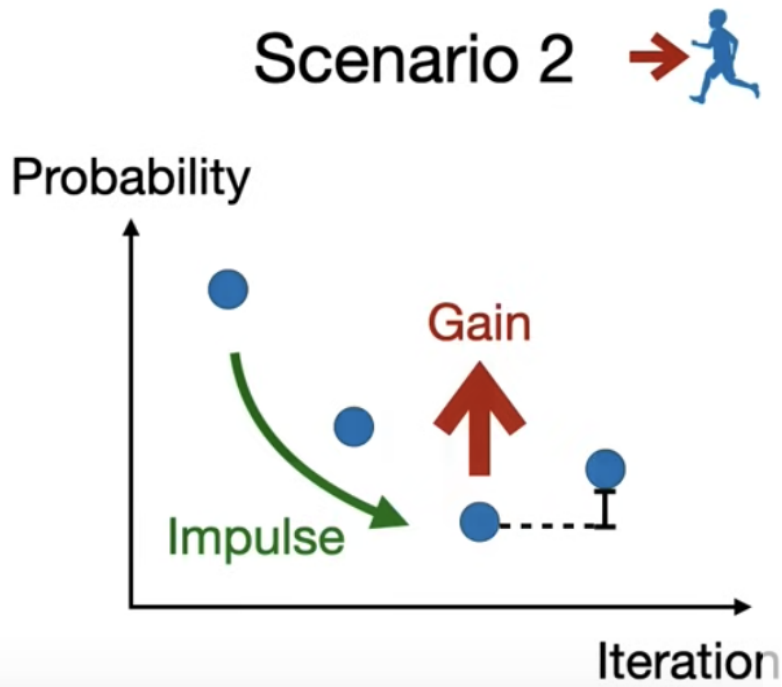


Main idea for training policy

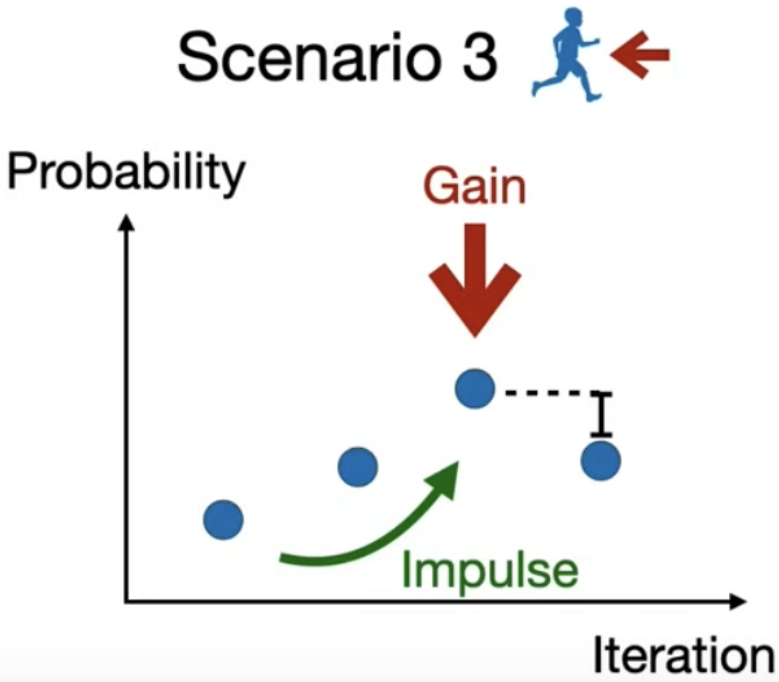


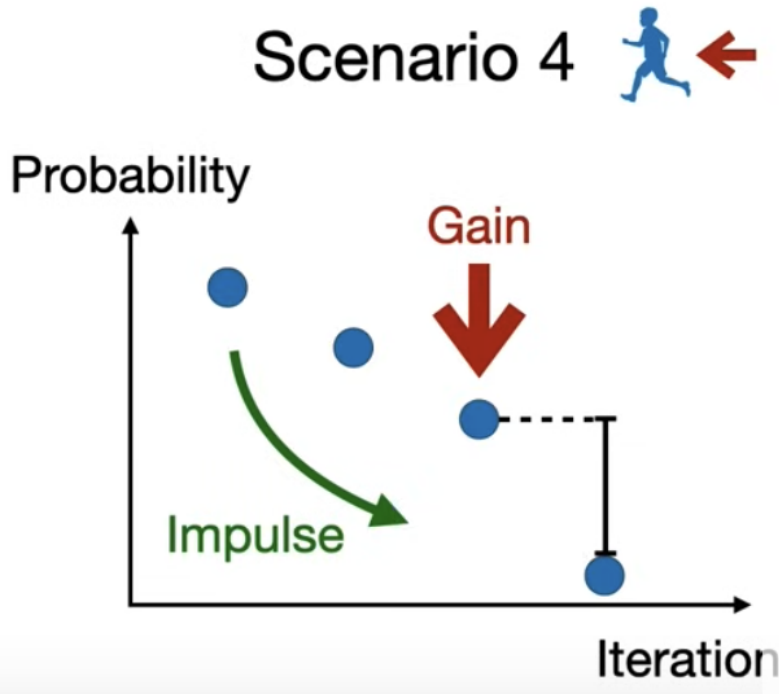
Impulse (Momentum)



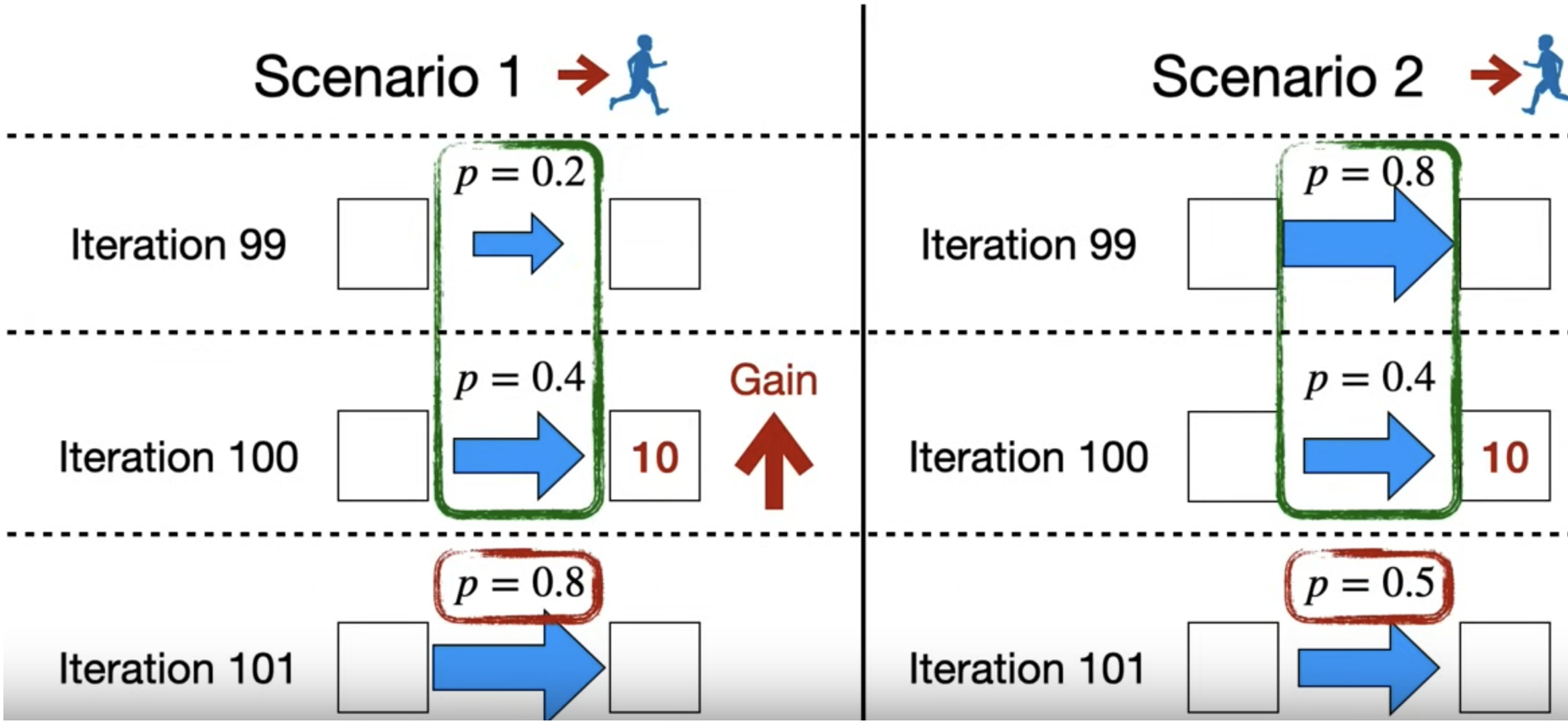


Impulse (Momentum)

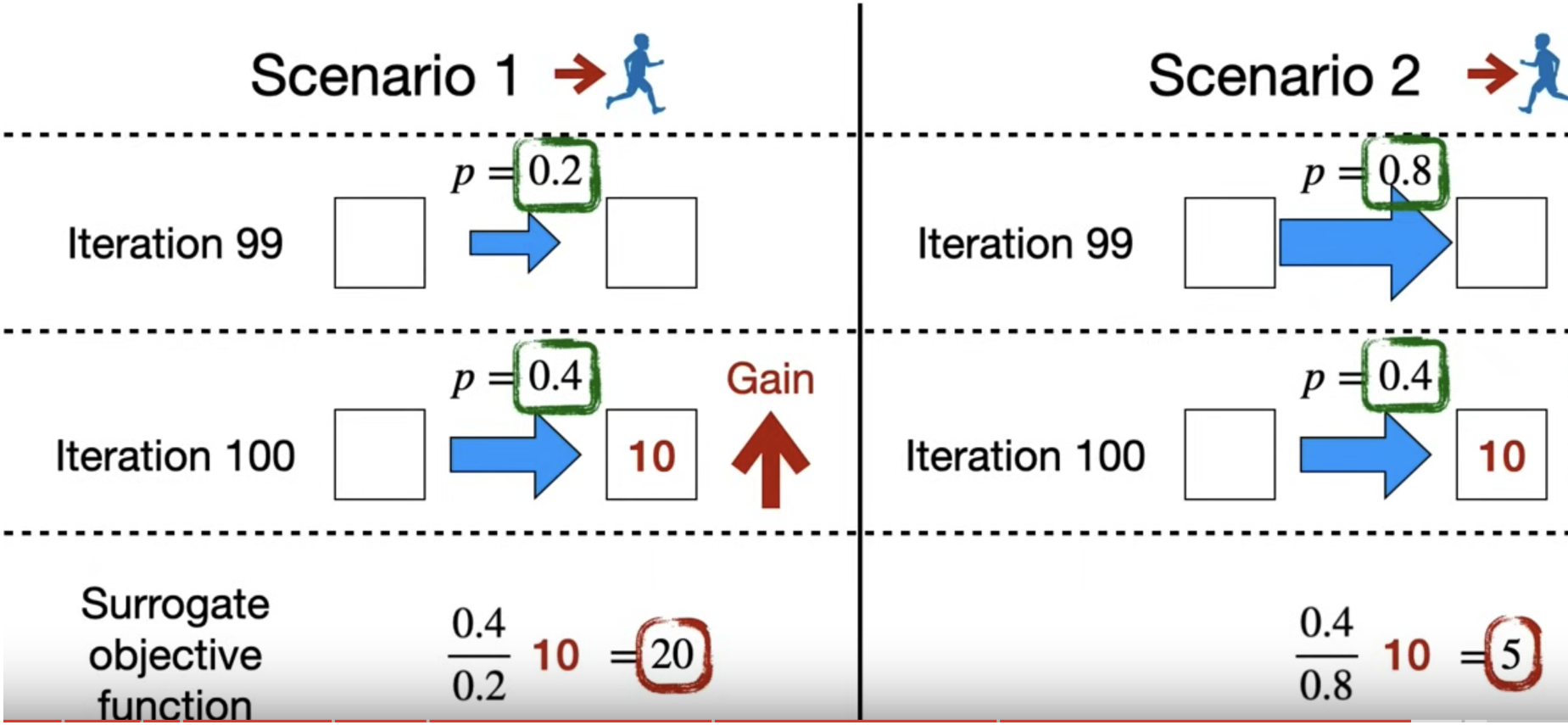




Impulse (Momentum)



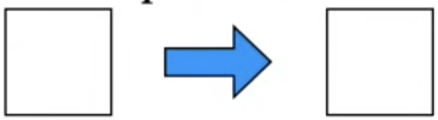
Dividing by the previous probability

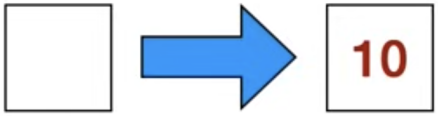


Surrogate Objective Function

Train neural network to make this as small as possible

$$L_{\text{policy}}(\theta) = \mathbb{E}\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} A_t$$

Iteration 99 

Iteration 100 

$$\frac{0.4}{0.2} \; 10 = 20$$

Solution: Clipping

Current probability

Previous probability

$$\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

If between 0.7 and 1.3, keep

If too small: Make it 0.7

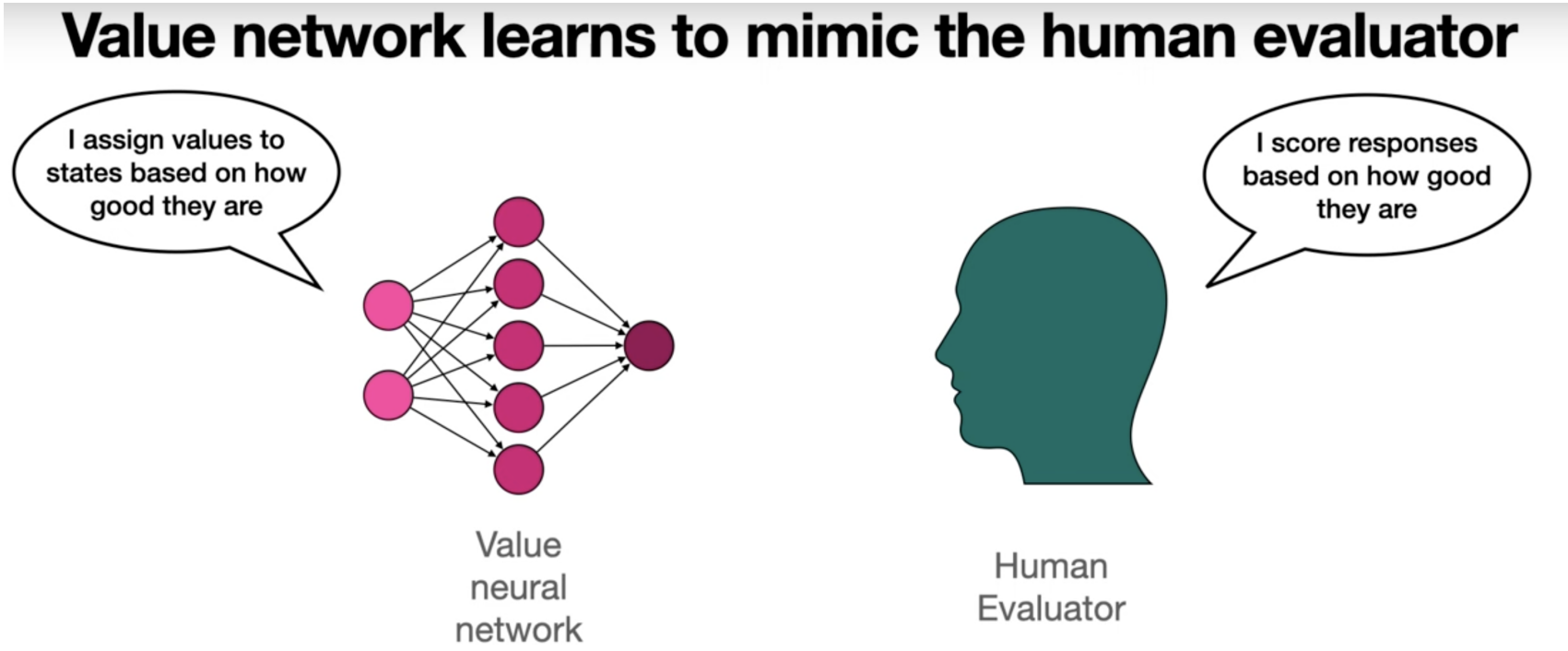
If too large: Make it 1.3

Solution: Clipping

$$L^{\text{CLIP}}_{\text{policy}}(\theta) = \mathbb{E}\left[\min\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} A_t, \text{clip}(\rho, 1-\epsilon, 1+\epsilon) A_t\right)\right]$$

- If between $1-\epsilon$ and $1+\epsilon$, keep
- If smaller than $1-\epsilon$, keep $1-\epsilon$
- If larger than $1+\epsilon$: Make it $1+\epsilon$

Re-Information Learning with Human Feedback



Training the value network

			What color is the sky? Red		
		What color is the sky? Blue	What color is the sky?	What color is the sky? Banana	
			What color is the sky		
	What color	What color is	What color is the		
<start>	What				

What color is the sky?
Blue

What color is the sky?
Red

What color is the sky?
Banana

Training the value network

			+2		
		+3	What color is the sky?	+1	
			What color is the sky		
	What color	What color is	What color is the		
<start>	What				

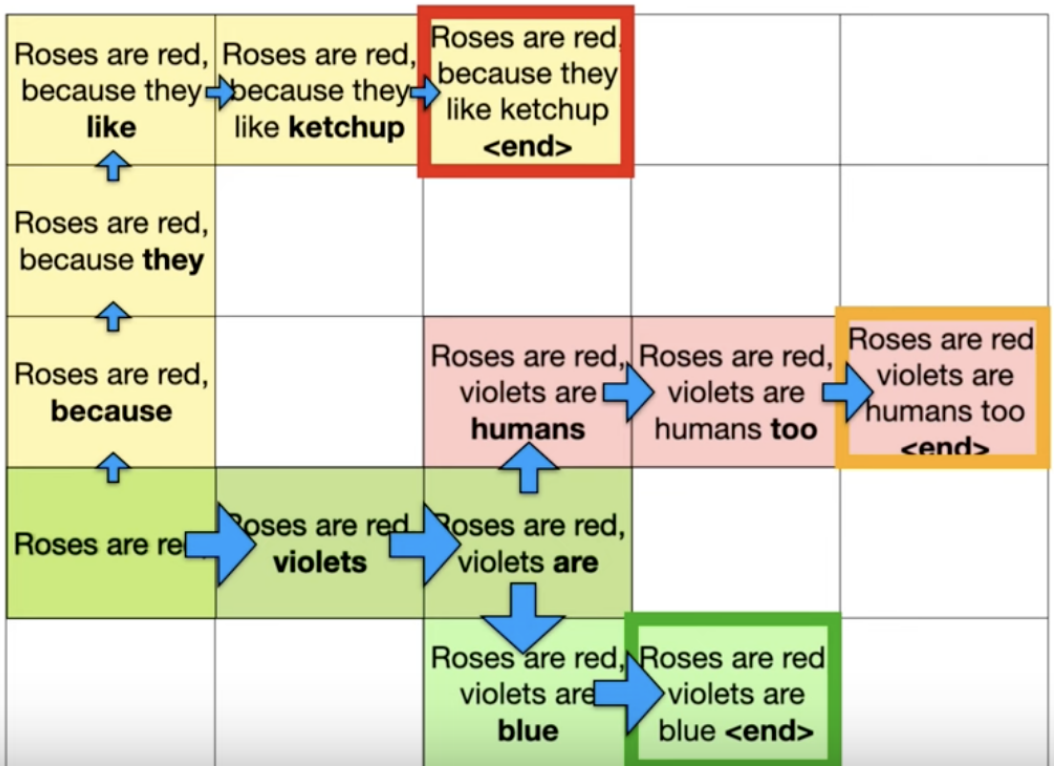
+3 What color is the sky? Blue

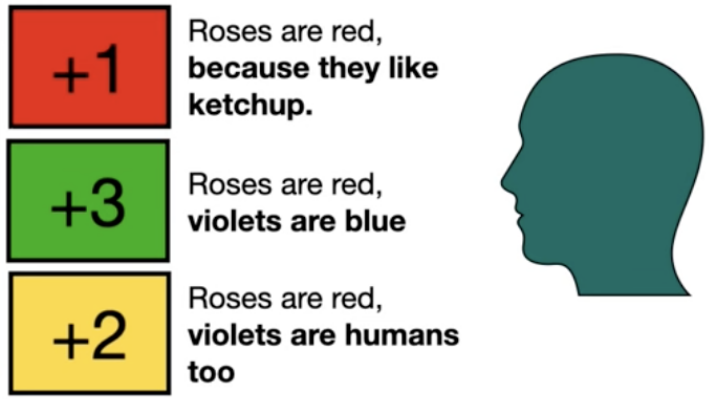
+2 What color is the sky? Red

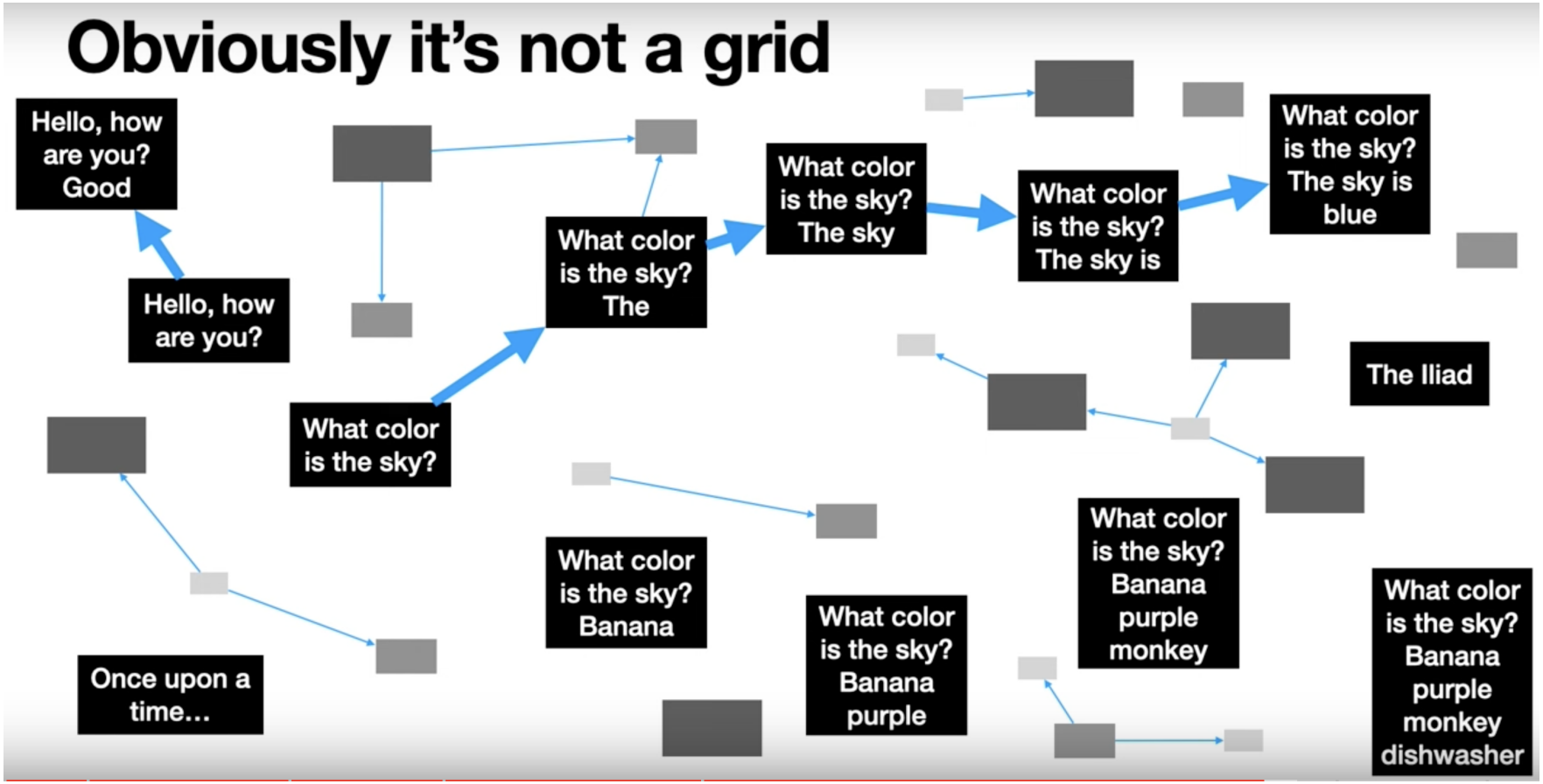
+1 What color is the sky? Banana

Value neural network

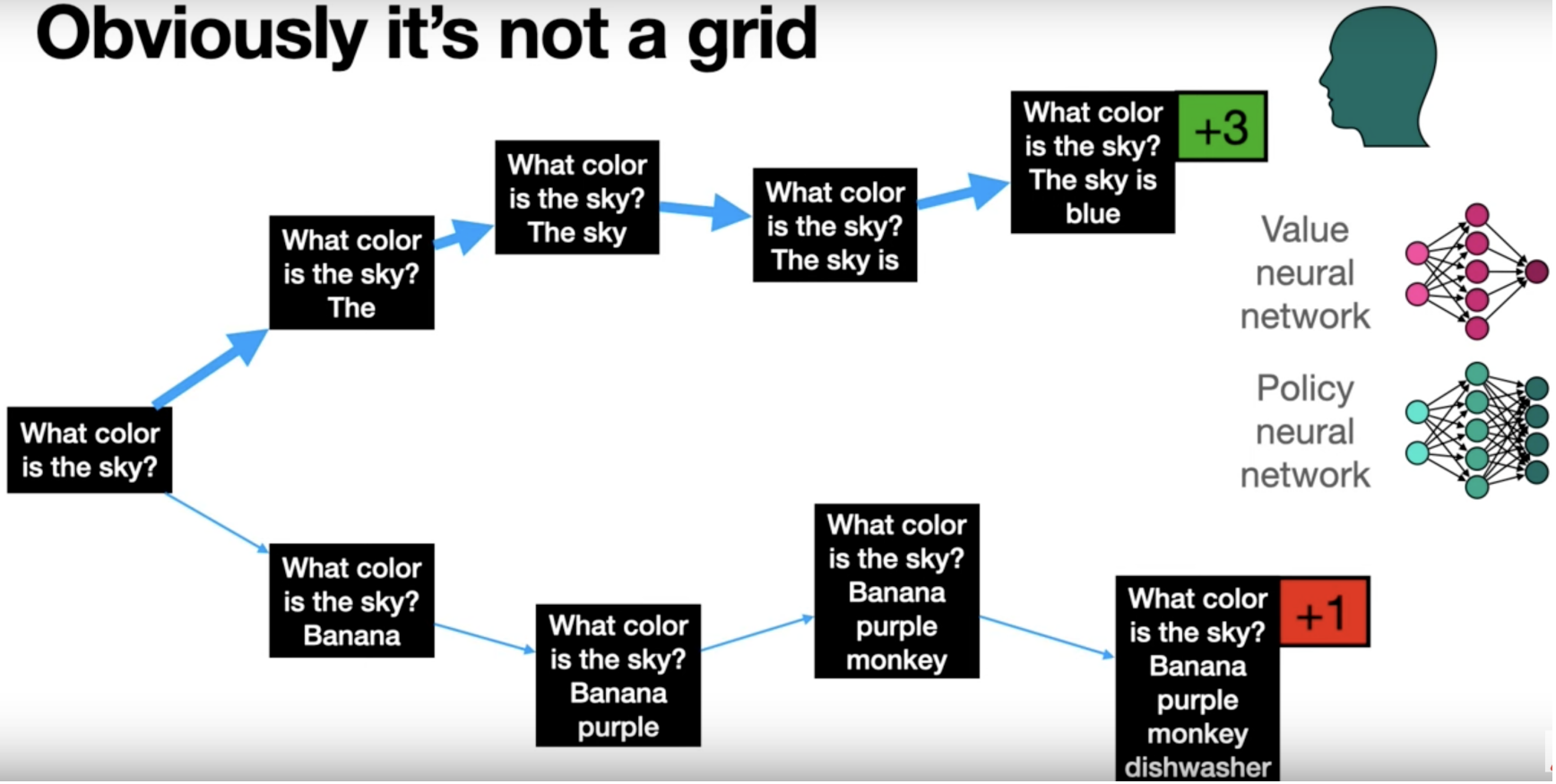
Longer paths






Obviously it's not a grid
Hello, how are you? Good
Hello, how are you?
What color is the sky?
What color is the sky? The
What color is the sky? The sky
What color is the sky? The sky is
What color is the sky? The sky is blue
The Iliad
What color is the sky? Banana
What color is the sky? Banana purple
What color is the sky? Banana purple monkey
What color is the sky? Banana purple monkey dishwasher
Once upon a time…

Obviously it's not a grid



Illustrative user prompts from InstructGPT/GPT-3 distributions

Use Case	Example
brainstorming	indie movie ideas - A guy travels to South America to become a shaman. - A documentary about the world of juggling.
brainstorming	Tell me a list of topics related to: - interior design - sustainable ecosystem - fake plants
classification	The following is a list of products and the kind of product t Product: product. Type: type Product: product. Type: type Product: product. Type:

Illustrative user prompts from InstructGPT/GPT-3 distributions

Use Case	Example
classification	This is a tweet sentiment classifier. tweet Sentiment: negative === tweet Sentiment: neutral === tweet Sentiment:

Categories for user prompts from InstructGPT/GPT-3

Brainstorming
Chat
Classification
Extract
Generation
QA, closed
QA, open
Rewrite
Summarization
Other

Illustrative user prompts from InstructGPT/GPT-3 distributions

Use Case	Example
generation	Write an outline for an essay about John von Neumann and his contributions to computing: I. Introduction, his life and background A: His early life B:
rewrite	Rephrase this for me: "I can't seem to find out how to work this darn thing." Alternate phrasing: "
rewrite	Covert my resume into a profile overview. { resume } Profile overview:

Illustrative user prompts from InstructGPT/GPT-3 distributions

Use Case	Example
chat	The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly. Human: Hello, who are you? AI: I am an AI created by OpenAI. How can I help you today? Human: I'm feeling kind of down today. AI:
closed qa	When you drop a heavy stone from a tree, what happens? A. The stone falls to the ground. B: The stone stays in the tree. C: The stone floats. D: Nothing happens. Answer:

Illustrative user prompts from InstructGPT/GPT-3 distributions

Use Case	Example
closed qa	Text: { article describing what yoga mats to buy } Question: What are the things I should consider when buying a yoga mat? Answer:
open qa	Q: Who is Leonardo da Vinci? A:

InstructGPT vs. GPT-3 (Ouyang et al. 2022)

- Labelers significantly prefer InstructGPT outputs over outputs from GPT-3.
- InstructGPT models show improvements in truthfulness over GPT-3.
- InstructGPT shows small improvements in toxicity over GPT-3, but not bias.
- InstructGPT models show promising generalization to instructions outside of the RLHF finetuning distribution.
- InstructGPT models generalize to the preferences of "held-out" labelers that did not produce any training data.
- InstructGPT still makes simple mistakes.

Direct Preference Optimization

Quote from the abstract of Rafailov et al. 2023:

In this paper we introduce a new parameterization of the reward model in RLHF that enables extraction of the corresponding optimal policy in closed form, allowing us to solve the standard RLHF problem with only a simple classification loss. The resulting algorithm, which we call Direct Preference Optimization (DPO), is stable, performant, and computationally lightweight, eliminating the need for sampling from the LM during fine-tuning or performing significant hyperparameter tuning. Our experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train.