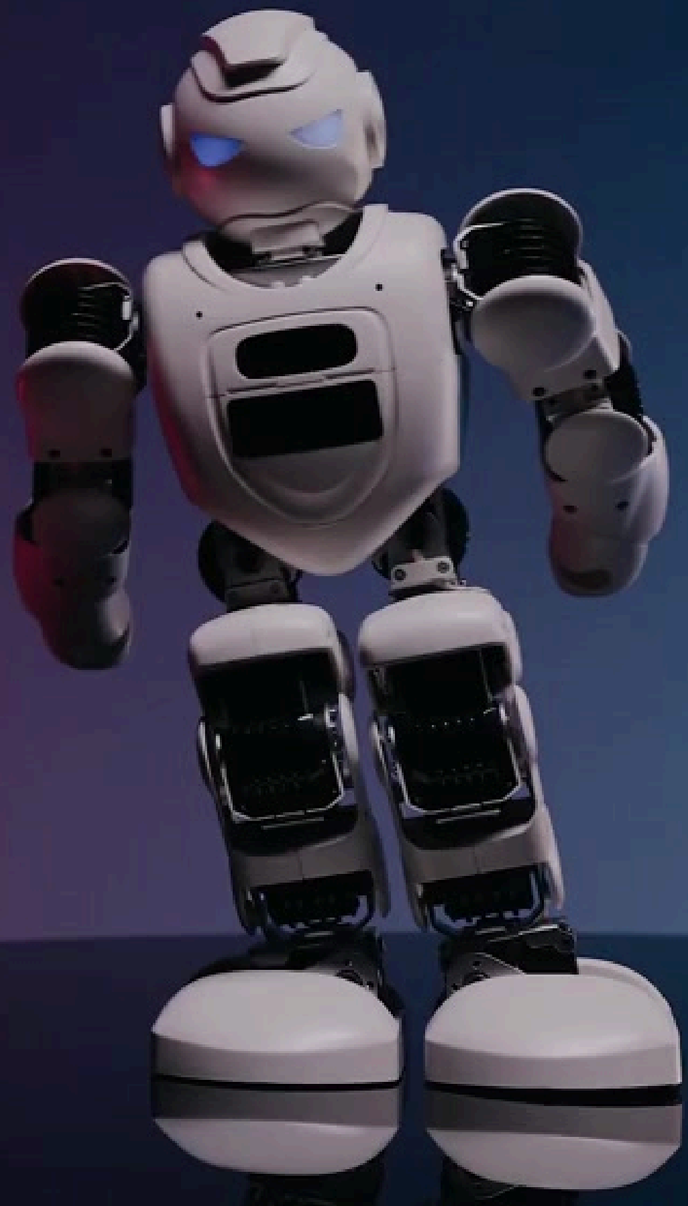
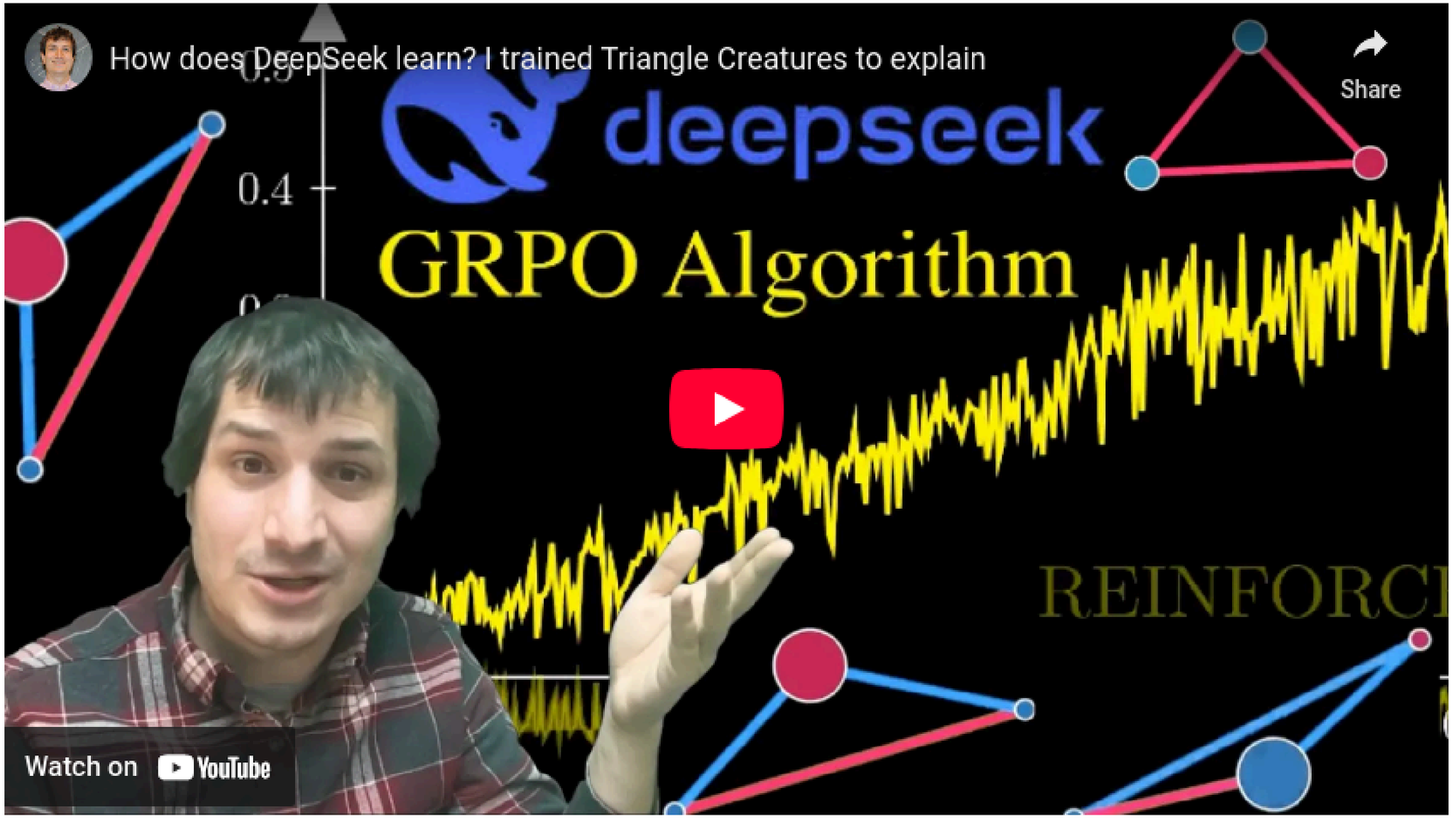
WEEK 11 RL

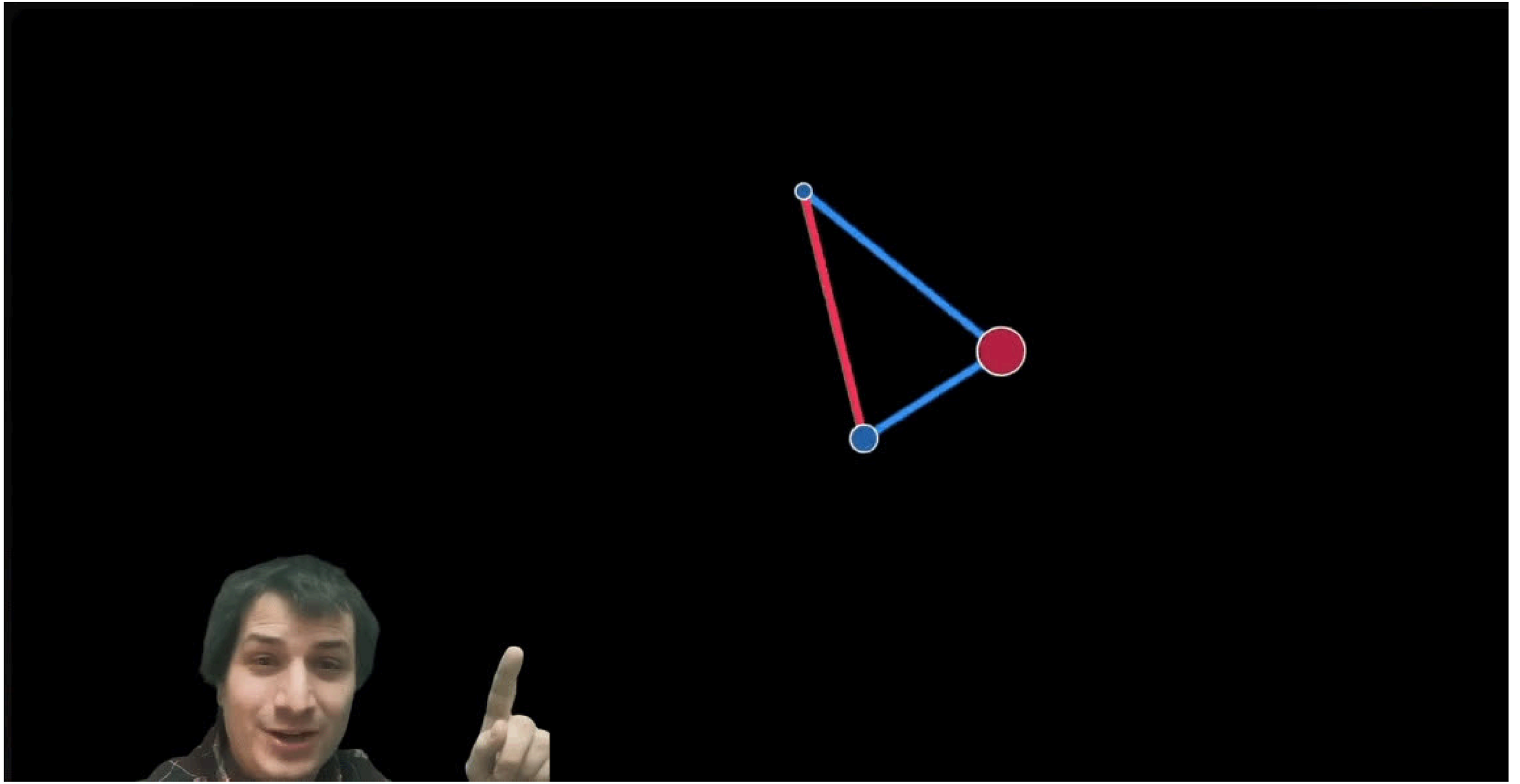
DEEP LEARNING FOR COMPUTER VISION

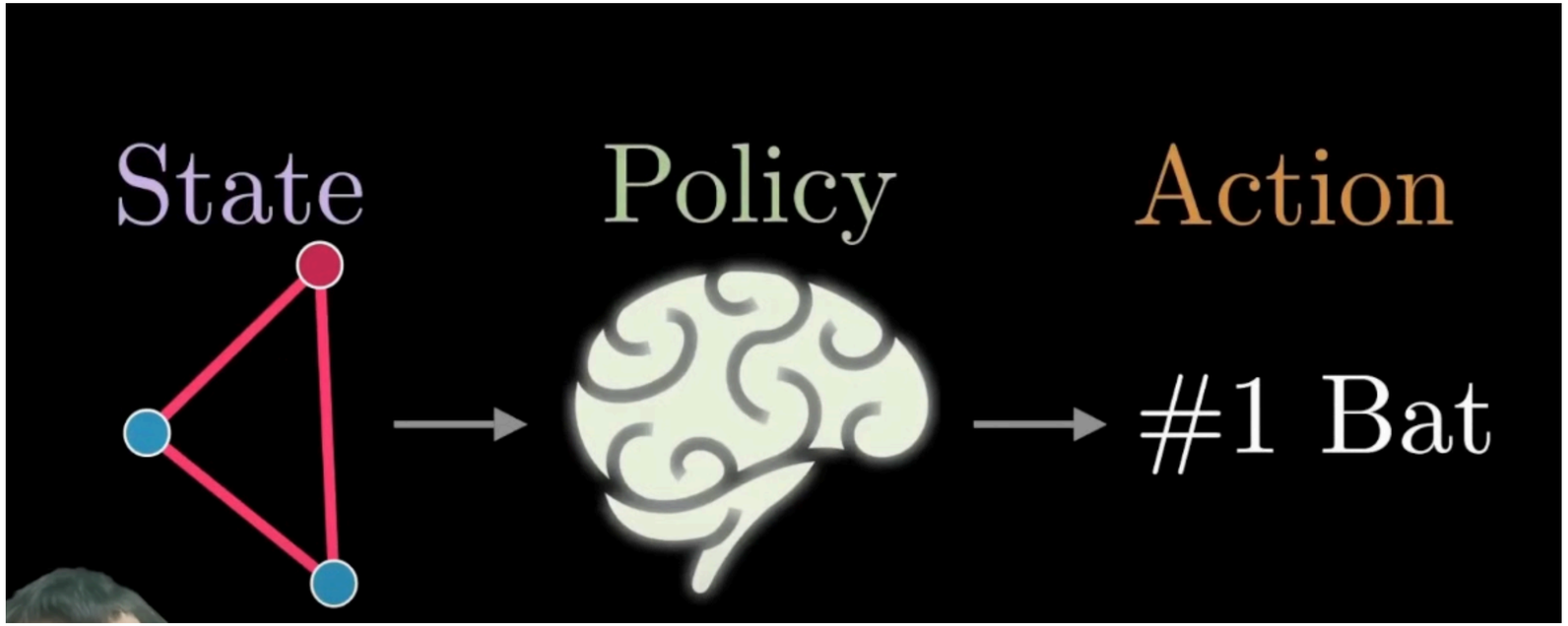
UNLIMITED

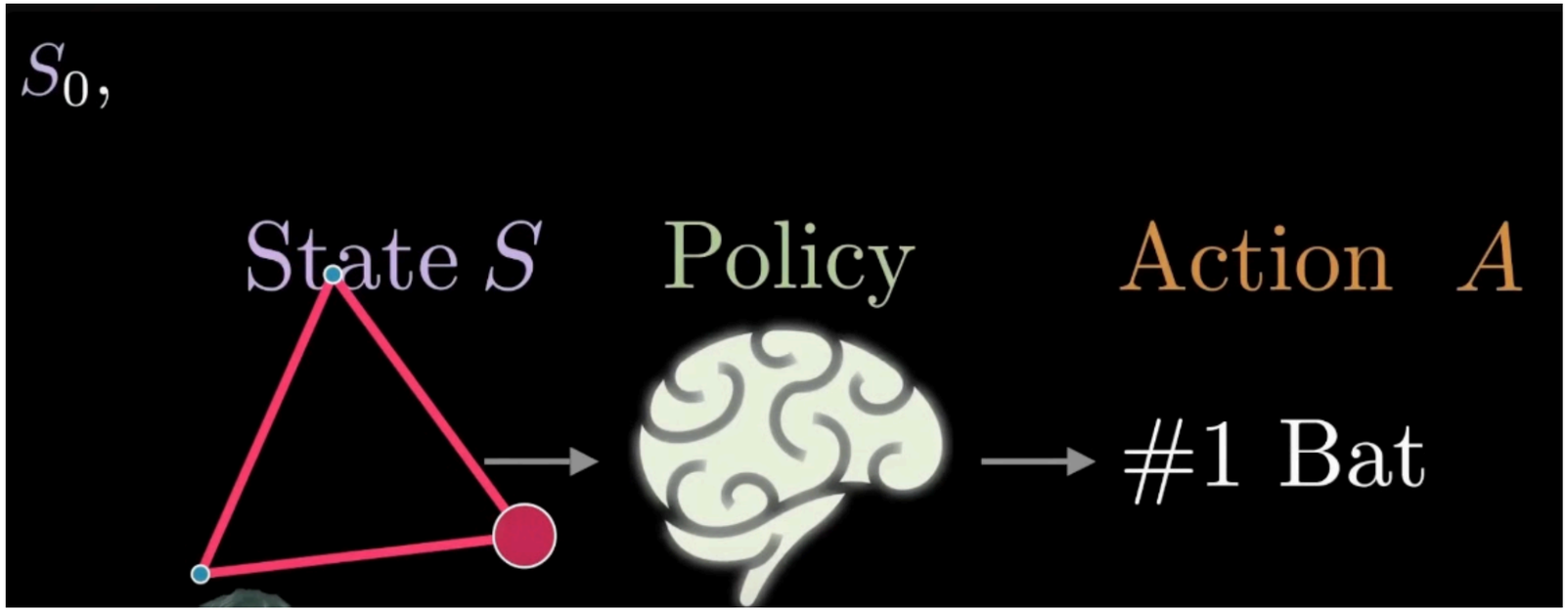


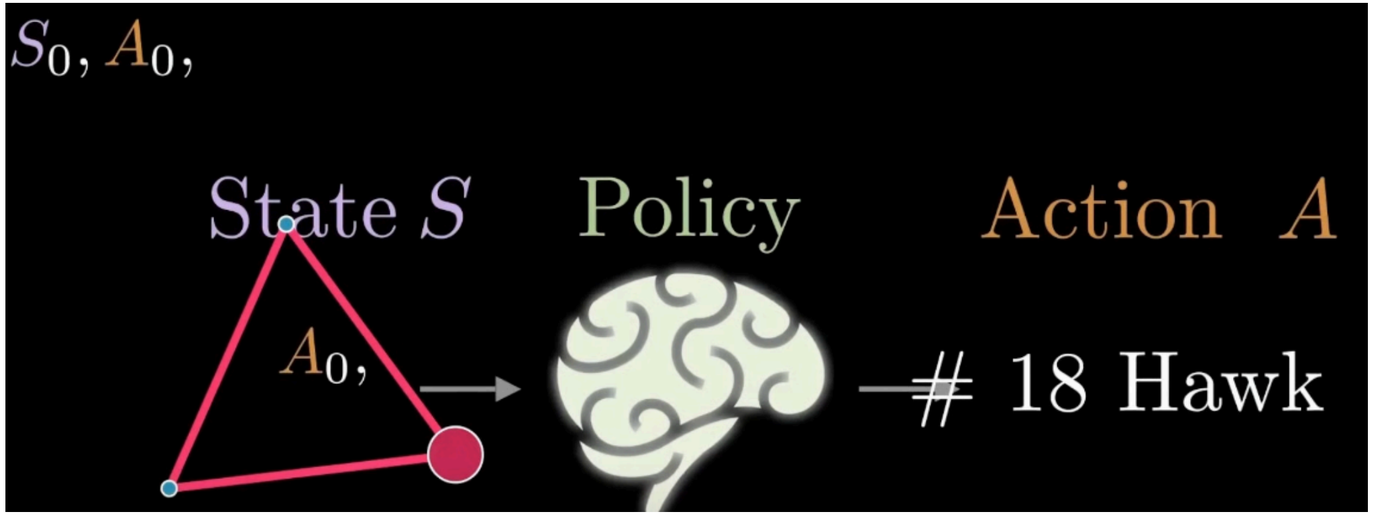
Presented by **Asst. Prof. Dr. Tuchsanai Ploysuwan**

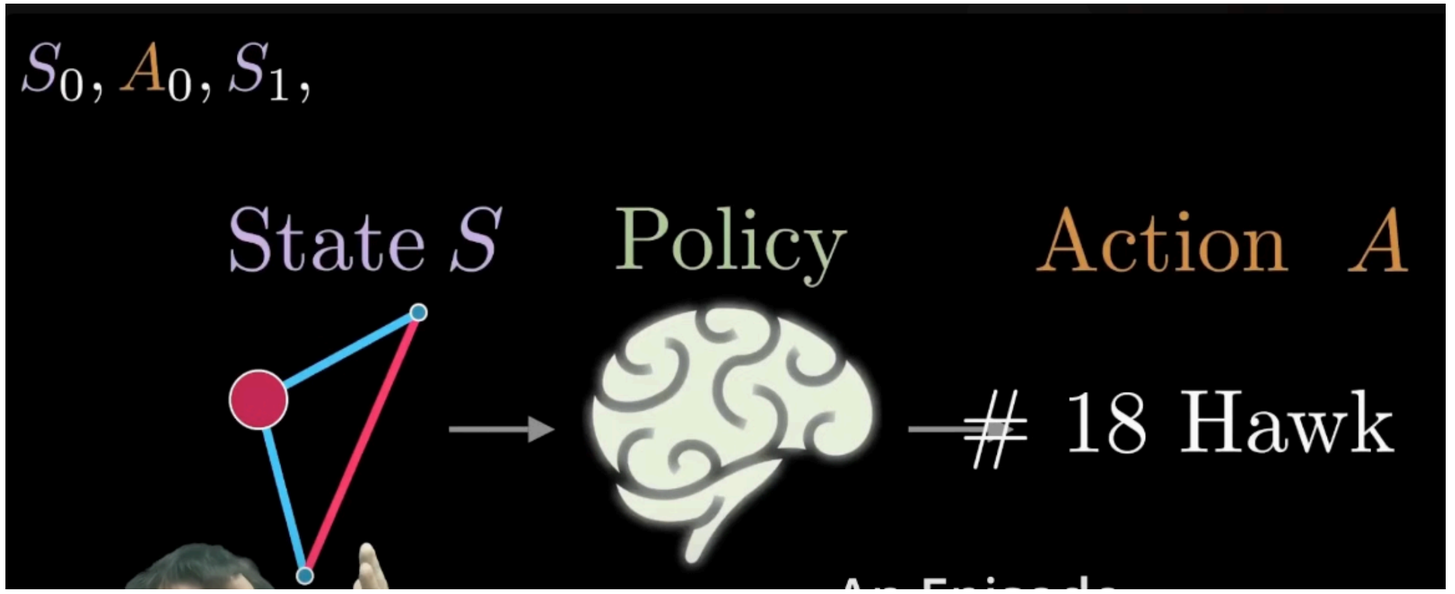

How does DeepSeek learn? I trained Triangle Creatures to explain
Share
deepseek
GRPO Algorithm
0.4
REINFORCE
Watch on YouTube

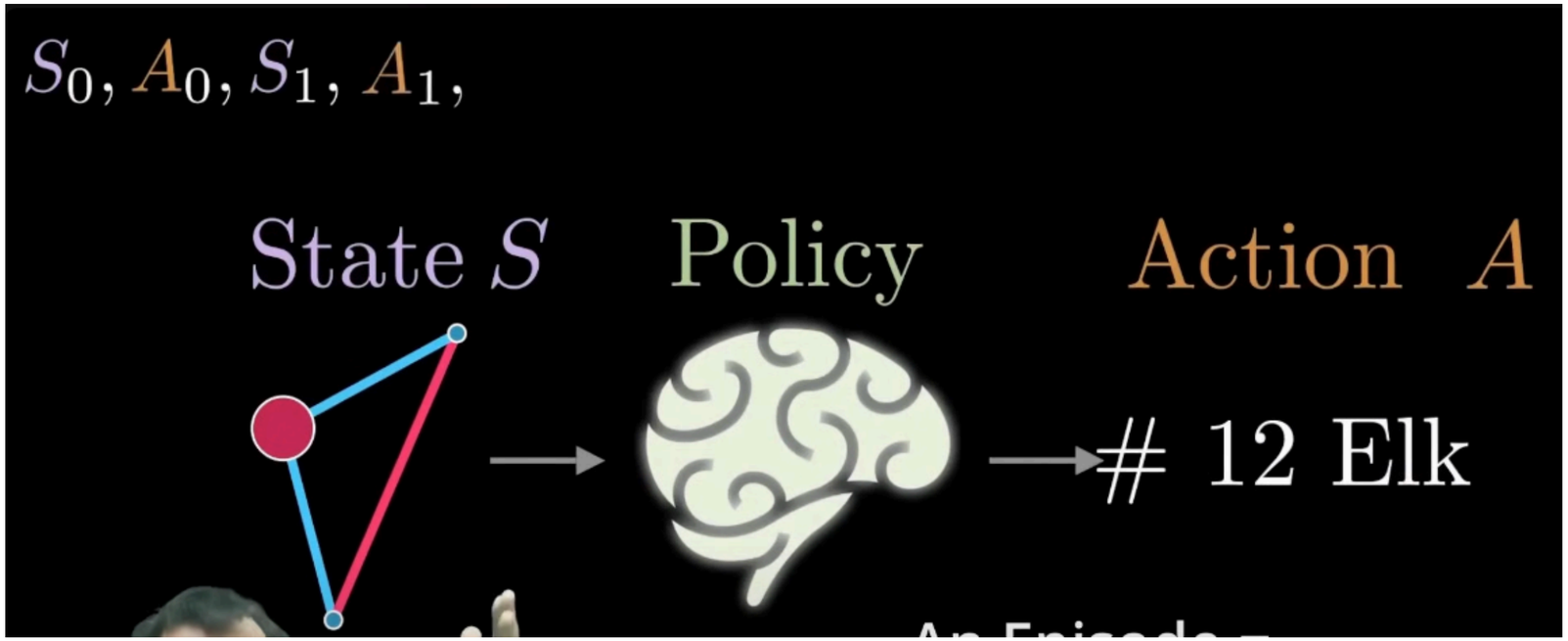


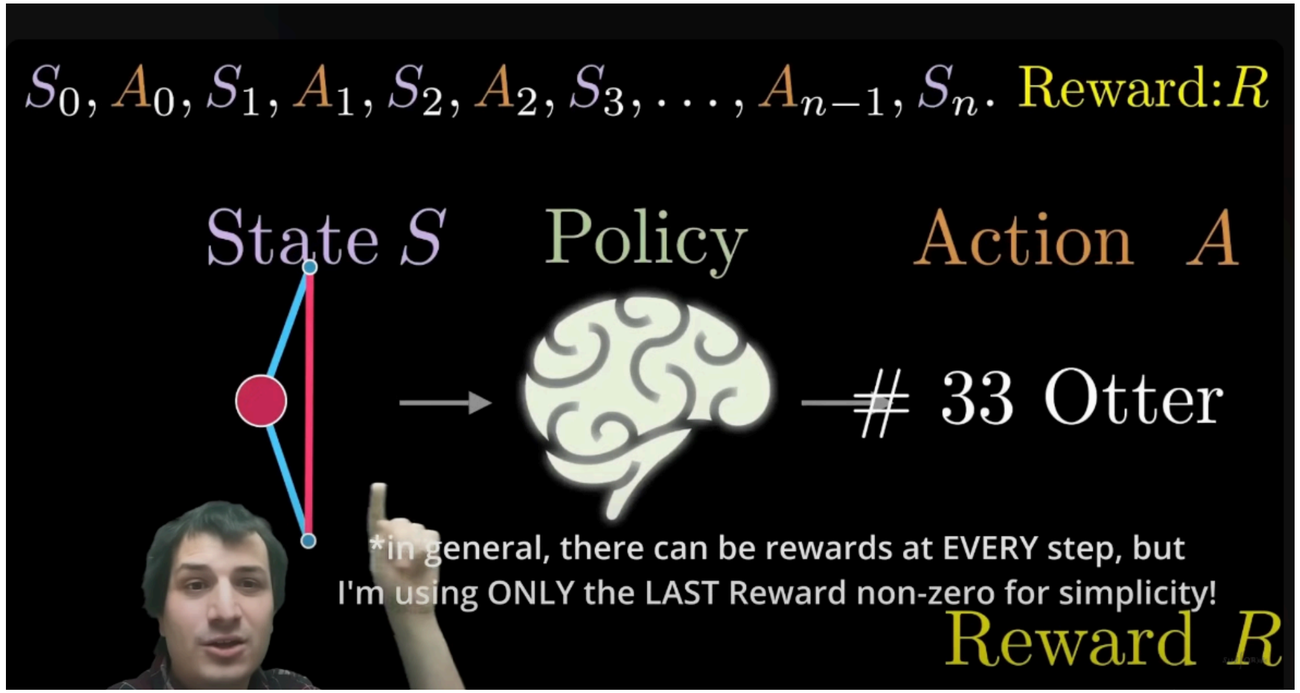

State
Policy
Action
#1 Bat


S_0,
State S
Policy
Action A
#1 Bat

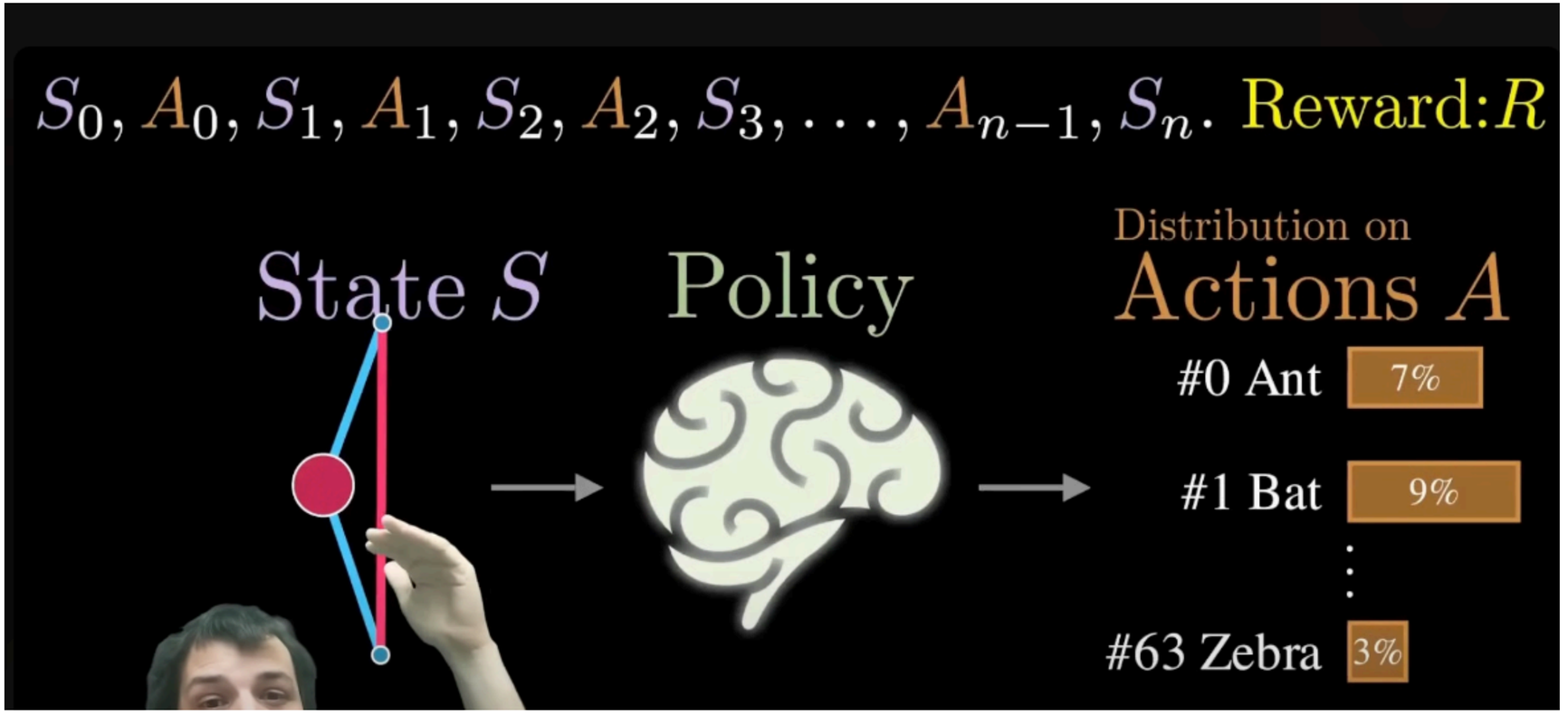

$S_0, A_0,$
State S
Policy
Action A
$A_0,$
18 Hawk


$S_0, A_0, S_1,$
State S
Policy
Action A
18 Hawk





$$\text{Reward } R = \Delta X - |\Delta Y|$$



Reward R

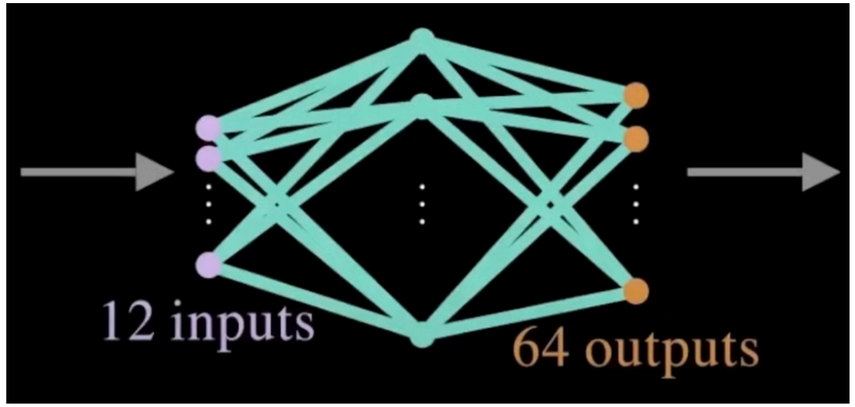
$= \Delta X - |\Delta Y|$

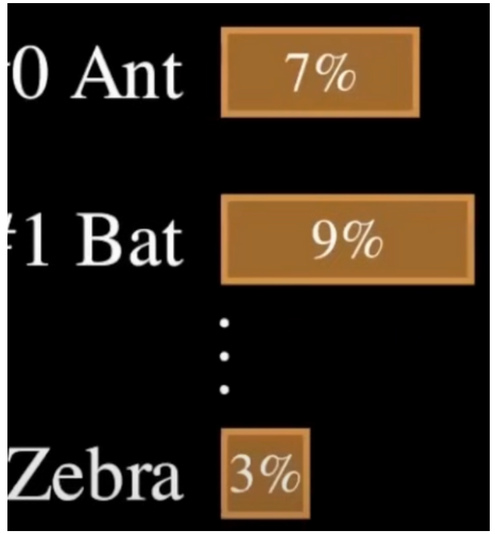
$$S_0, A_0, S_1, A_1, S_2, A_2, S_3, \ldots, A_{n-1}, S_n. \text{ Reward:} R$$

$$\pi(A|S, \theta) = \mathbb{P}(\text{Choose } A \text{ when at } S, \text{ params} = \theta)$$

Policy π

State S





Params θ

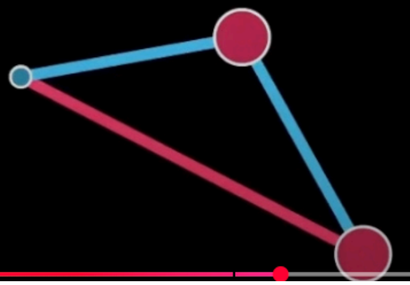
Reward R
$= \Delta X - |\Delta Y|$

The REINFORCE Algorithm

State S
Params θ
Policy π
Actions A
$\pi(A|S,\theta) = \mathbb{P}$
Reward R

How to improve θ?

- Observe $S_0, A_0, S_1, A_1, \ldots, S_n$ and R
- If R is large, then tweak θ to make actions A_t more likely. i.e. $\uparrow \pi(A_t|S_t,\theta)$



#0 Ant 7%

#1 Bat 9%

⋮

#63 Zebra 3%

The REINFORCE Algorithm

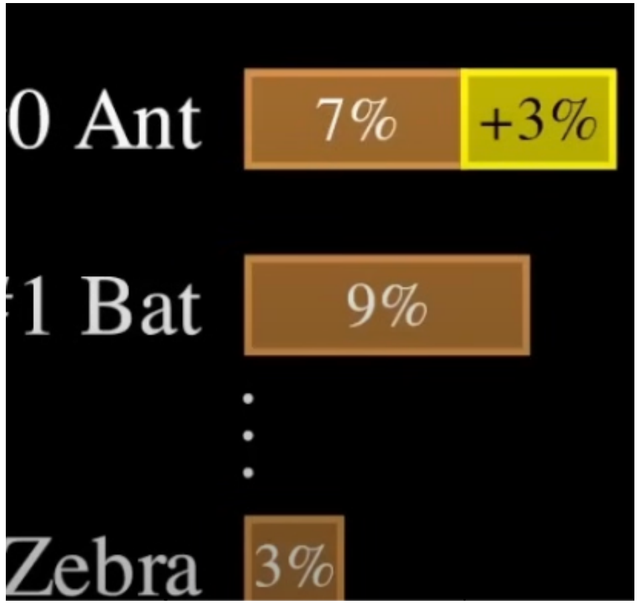
How to improve θ?

- Observe $S_0, A_0, S_1, A_1, \ldots, S_n$ and R
- If R is large, then tweak θ to make actions A_t more likely. i.e. $\uparrow \pi(A_t|S_t, \theta)$

State S
Params θ
Policy π
Actions A
$\pi(A|S, \theta) = \mathbb{P}$
Reward R

$$\nabla_\theta \log \pi(A_t|S_t, \theta)$$

$$R\nabla_\theta \log \pi(A_t|S_t, \theta)$$



```python
def log_policy_pi(params_theta, state_S, action_A):
    #"forward" is the neural network a.k.a. "the brain"
    log_probs = forward(params_theta, state_S) #log-probs of ALL possible acti
    return log_probs[action_A] #Plug in the action we took!

#Do the derivative using JAX. (Does gradient w.r.t. the 0th input by default)
grad_log_pi = jax.grad(log_policy_pi)

#Update the paramaters!
params_theta += reward_R * grad_log_pi(params_theta, state_S, action_A)
```

$$\theta_{\text{new}} = \theta + R\nabla_\theta \log \pi(A_t|S_t, \theta)$$


Avg. Reward: $\Delta X - |\Delta Y|$ (Body Lengths Per Action)
0.5
0.4
0.3
0.2
DeepSeek's
GRPO Algorithm
REINFORCE Algorithm
100
200
300
400

Improvement: The GRPO Algorithm

How to improve θ? Reward size R is relative!

- Observe $S_0, A_0, S_1, A_1, \ldots, S_n$ and R
- If R is large, then tweak θ to make actions A_t more likely. i.e. $\uparrow \pi(A_t|S_t, \theta)$

State S
Params θ
Policy π
Actions A
$\pi(A|S, \theta) = \mathbb{P}$
Reward R

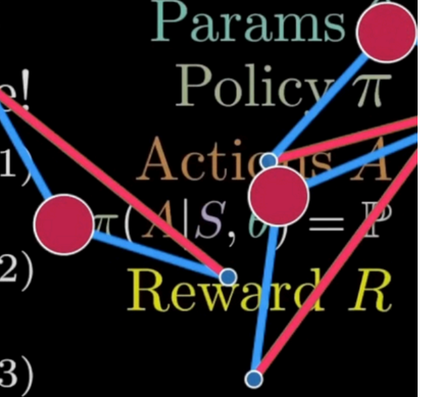
$$\theta_{\text{new}} = \theta + R\nabla_\theta \log \pi(A_t|S_t, \theta)$$

Improvement: The GRPO Algorithm

How to improve θ? Reward size R is relative!

- Observe $S_0, A_0^{(1)}, S_1^{(1)}, A_1^{(1)}, \ldots, S_n^{(1)}$ and $R^{(1)}$
- Observe $S_0, A_0^{(2)}, S_1^{(2)}, A_1^{(2)}, \ldots, S_n^{(2)}$ and $R^{(2)}$
- Observe $S_0, A_0^{(3)}, S_1^{(3)}, A_1^{(3)}, \ldots, S_n^{(3)}$ and $R^{(3)}$

State S
Params θ
Policy π
Actions A
$\pi(A|S, \theta) = \mathbb{P}$
Reward R

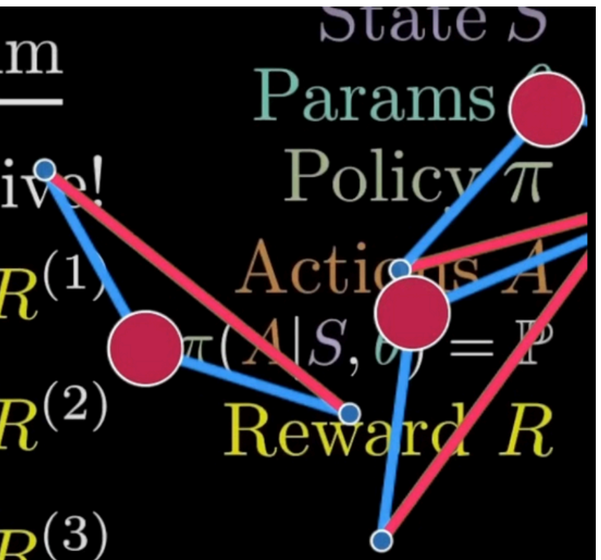


$$\theta_{\text{new}} = \theta + R\nabla_\theta \log \pi(A_t|S_t, \theta)$$

Improvement: The GRPO Algorithm

How to improve θ? Reward size R is relative!

- Observe $S_0, A_0^{(1)}, S_1^{(1)}, A_1^{(1)}, \ldots, S_n^{(1)}$ and $R^{(1)}$
- Observe $S_0, A_0^{(2)}, S_1^{(2)}, A_1^{(2)}, \ldots, S_n^{(2)}$ and $R^{(2)}$
- Observe $S_0, A_0^{(3)}, S_1^{(3)}, A_1^{(3)}, \ldots, S_n^{(3)}$ and $R^{(3)}$



$$\theta_{\text{new}} = \theta + \text{Advantage}^{(1)} \nabla_\theta \log \pi(A_t^{(1)} | S_t^{(1)}, \theta)$$

$$\text{Advantage}^{(1)} = \frac{R^{(1)} - \text{avg}\{R^{(1)}, R^{(2)}, R^{(3)}\}}{\text{std}\{R^{(1)}, R^{(2)}, R^{(3)}\}}$$