A Perspective on Intelligence

Richard Sutton

University of Alberta
Alberta Machine Intelligence Institute
Reinforcement Learning and Artificial Intelligence Lab

> “Intelligence is the most powerful phenomenon in the universe”
>
> —Ray Kurzweil, 2009, *Transcendent Man*

Artificial intelligence research is ambitious

- AI researchers seek to understand intelligence well enough to create beings of greater intelligence than current humans

- Reaching this profound intellectual milestone will enrich our economies and challenge our societal institutions

  - It will be unprecedented and transformational, but also a continuation of trends that are thousands of years old

- People have always created tools and been changed by them; it's what humans do

- The next big step is to understand ourselves

- This is a quest grand and glorious, and quintessentially human

My perspective

- The creation of super-intelligent agents, and super-intelligent augmented humans, will be an unalloyed good for the world

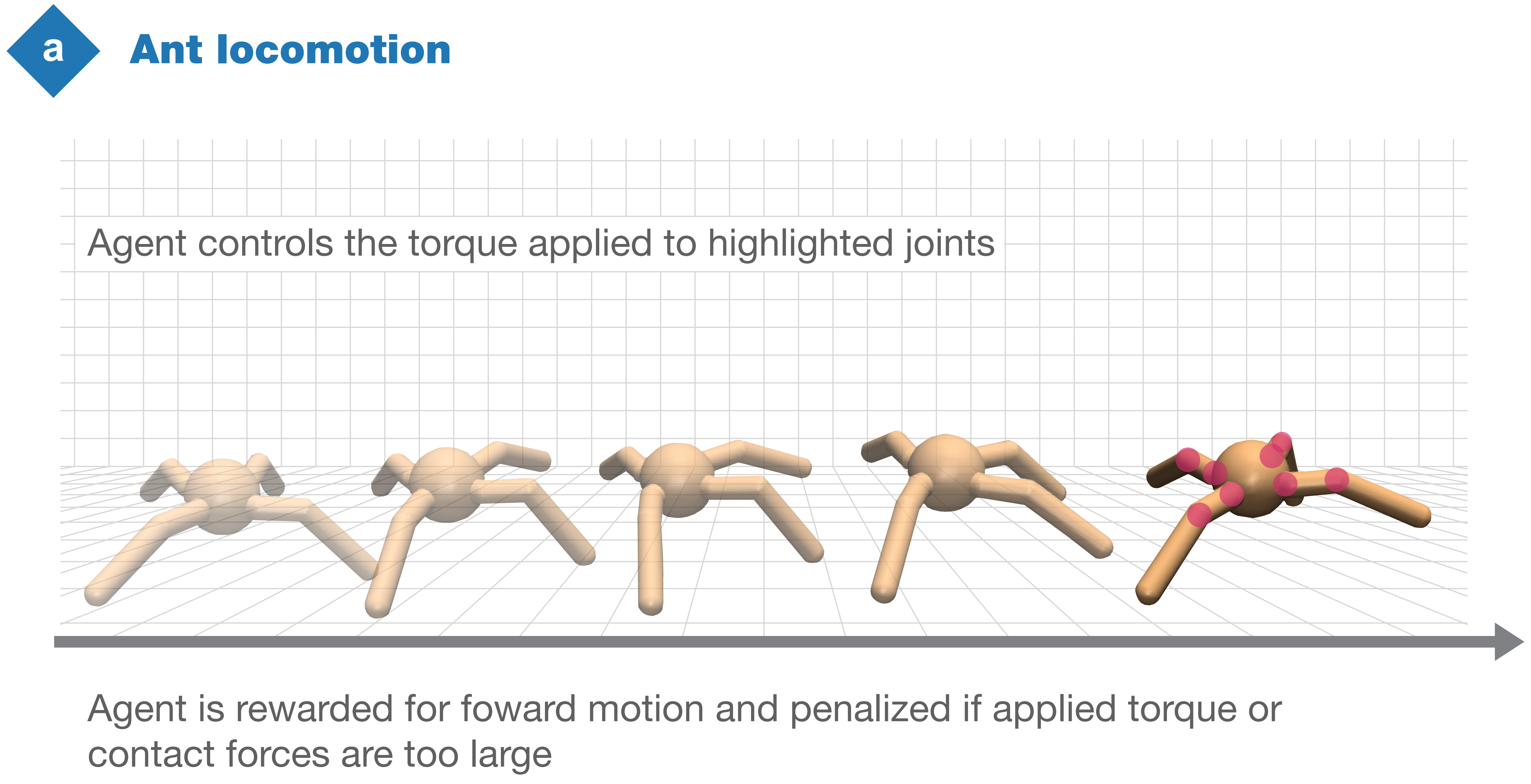
- The path to intelligent agents runs through reinforcement learning (and not through, e.g., Large Language Models)

- The biggest bottleneck to ambitious AI is inadequate deep learning algorithms (article in the journal *Nature*, August 22, 2024)

- The greatest impacts and advances in AI are still to come

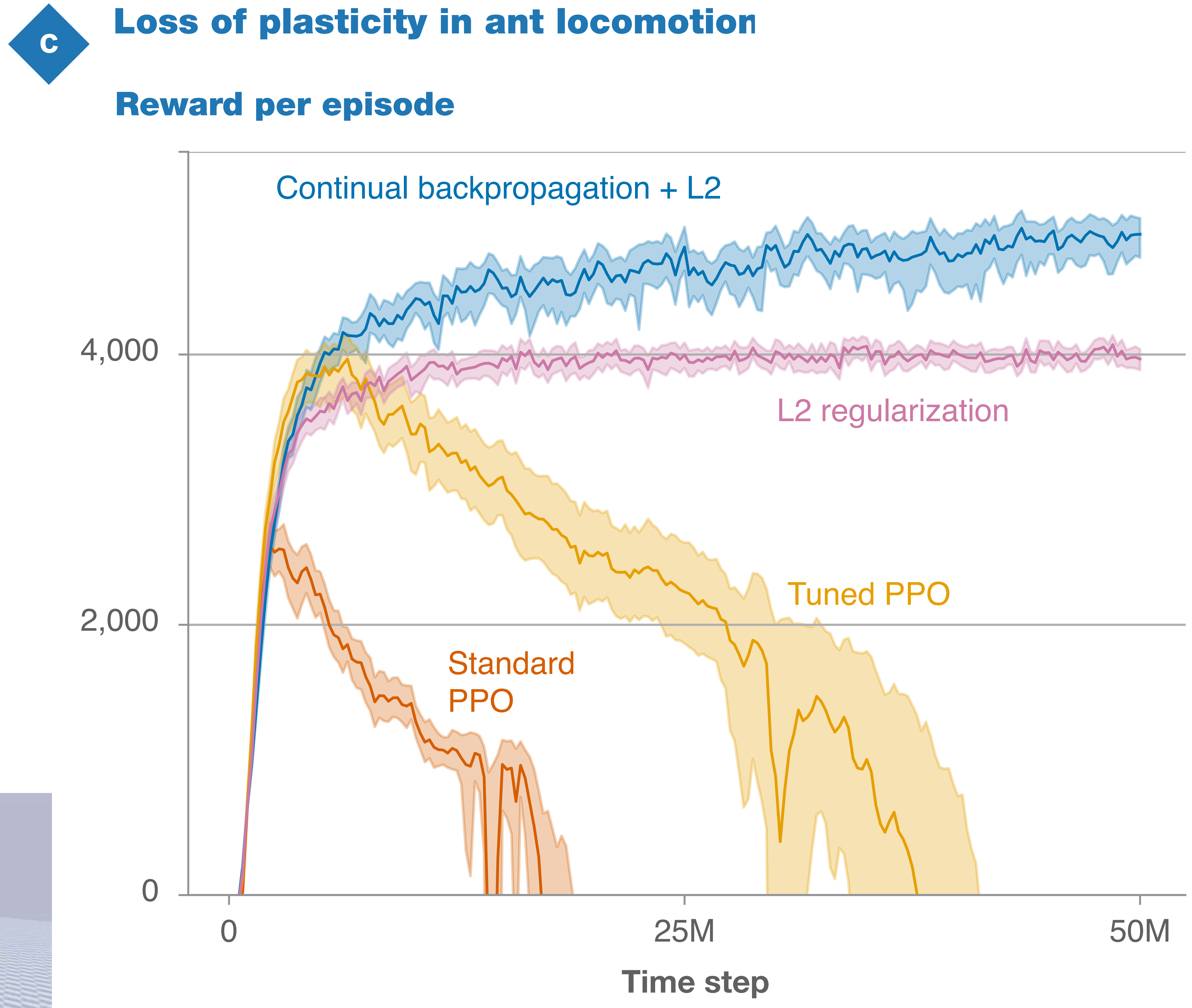
  - If AI is a race, it's a marathon, not a sprint

Loss of Plasticity in Reinforcement Learning

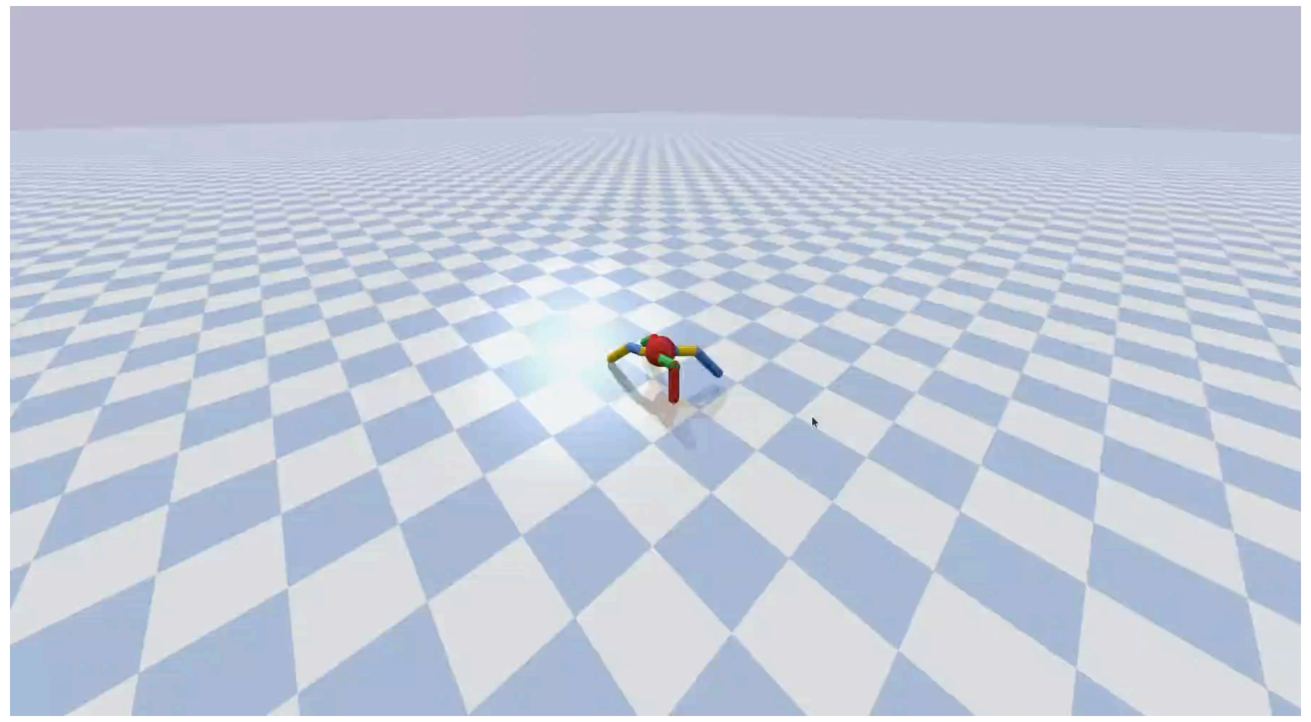


Agent controls the torque applied to highlighted joints

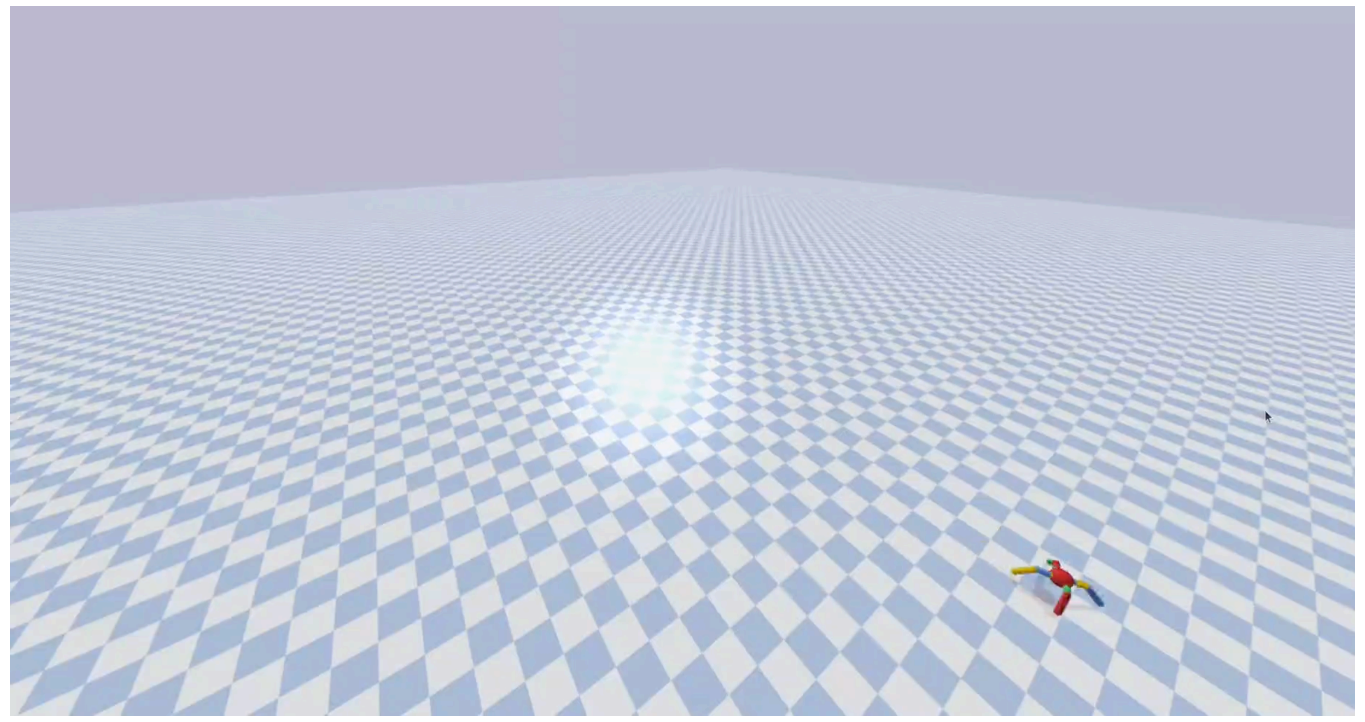
Agent is rewarded for foward motion and penalized if applied torque or contact forces are too large



PPO



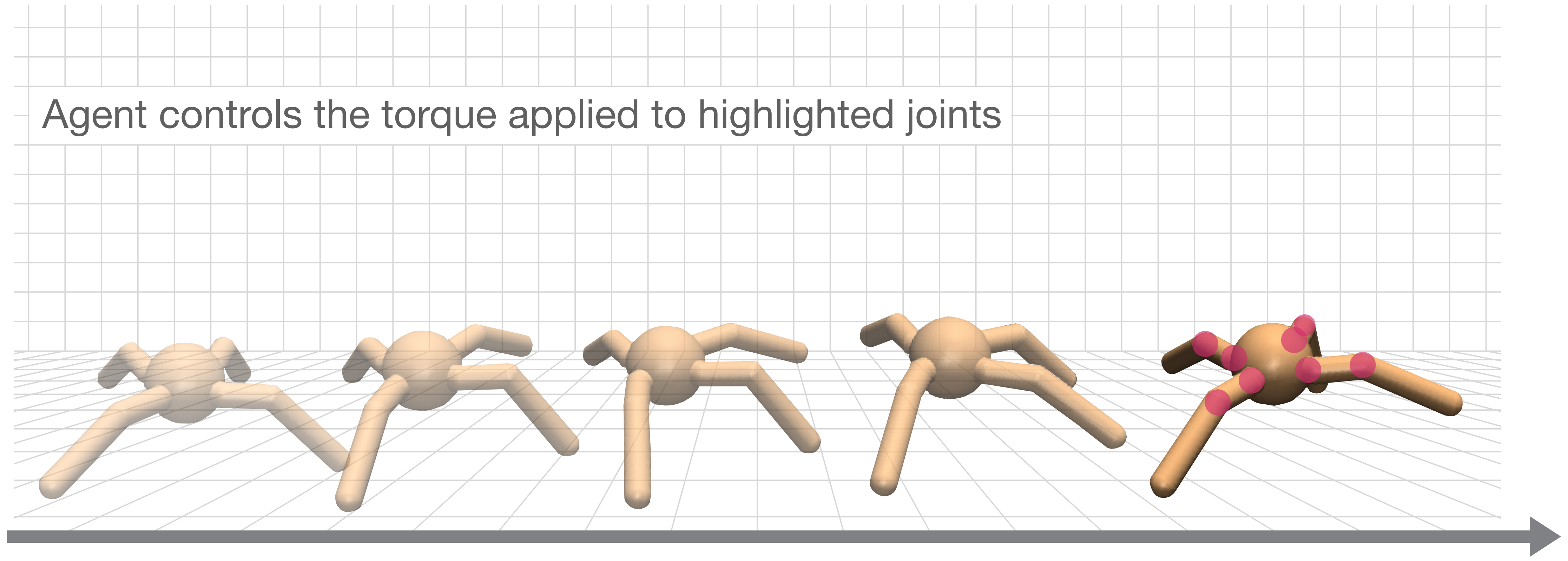
Continual PPO



“Loss of Plasticity in Deep Continual Learning”
by Dohare, Hernandez-Garcia, Lan, Rahman,
Mahmood, & Sutton, *Nature 632*, August 22, 2024

Loss of Plasticity in Reinforcement Learning

a **Ant locomotion**

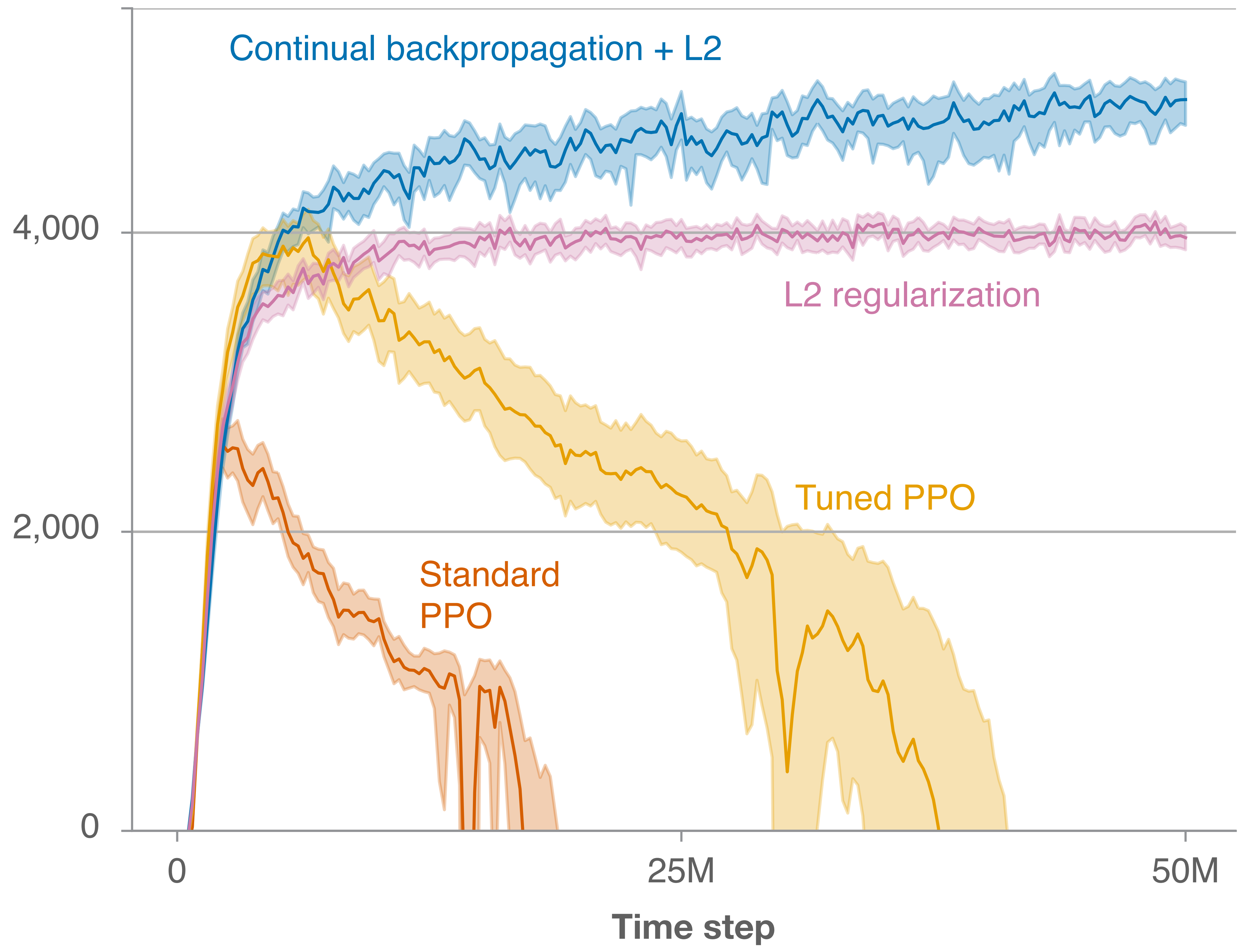


Agent controls the torque applied to highlighted joints

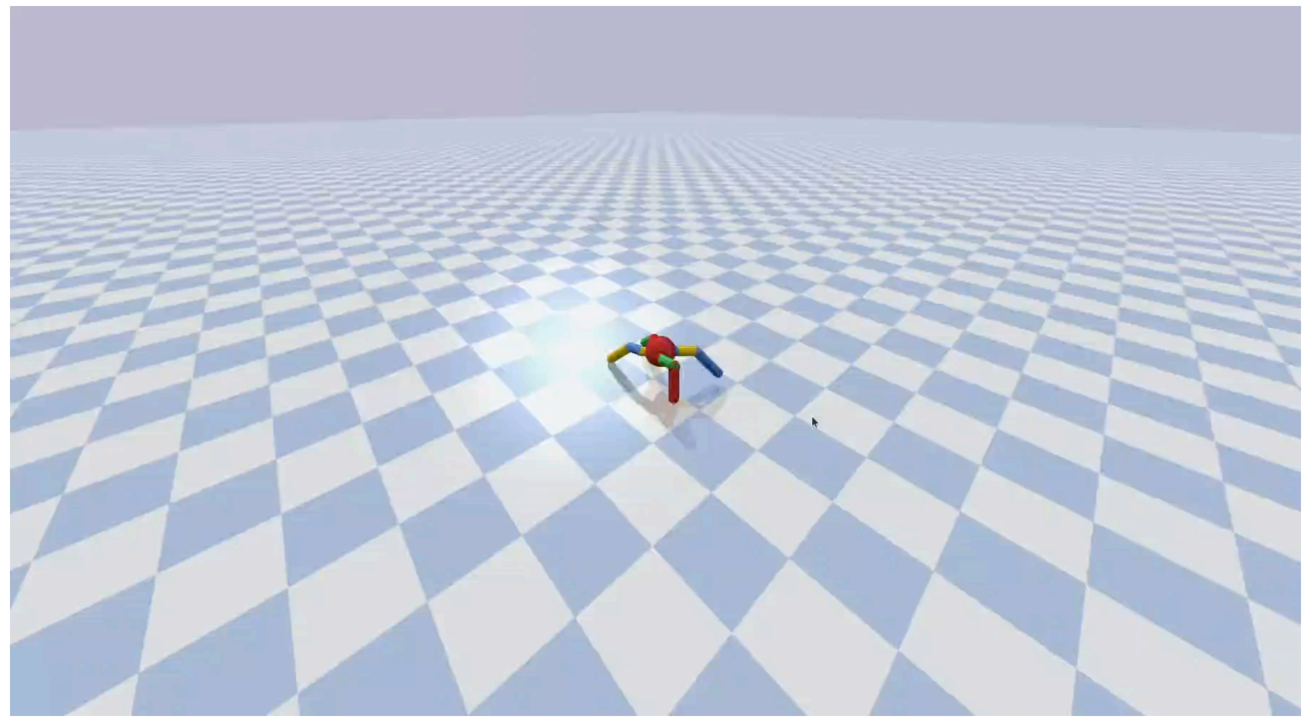
Agent is rewarded for foward motion and penalized if applied torque or contact forces are too large

c **Loss of plasticity in ant locomotion**

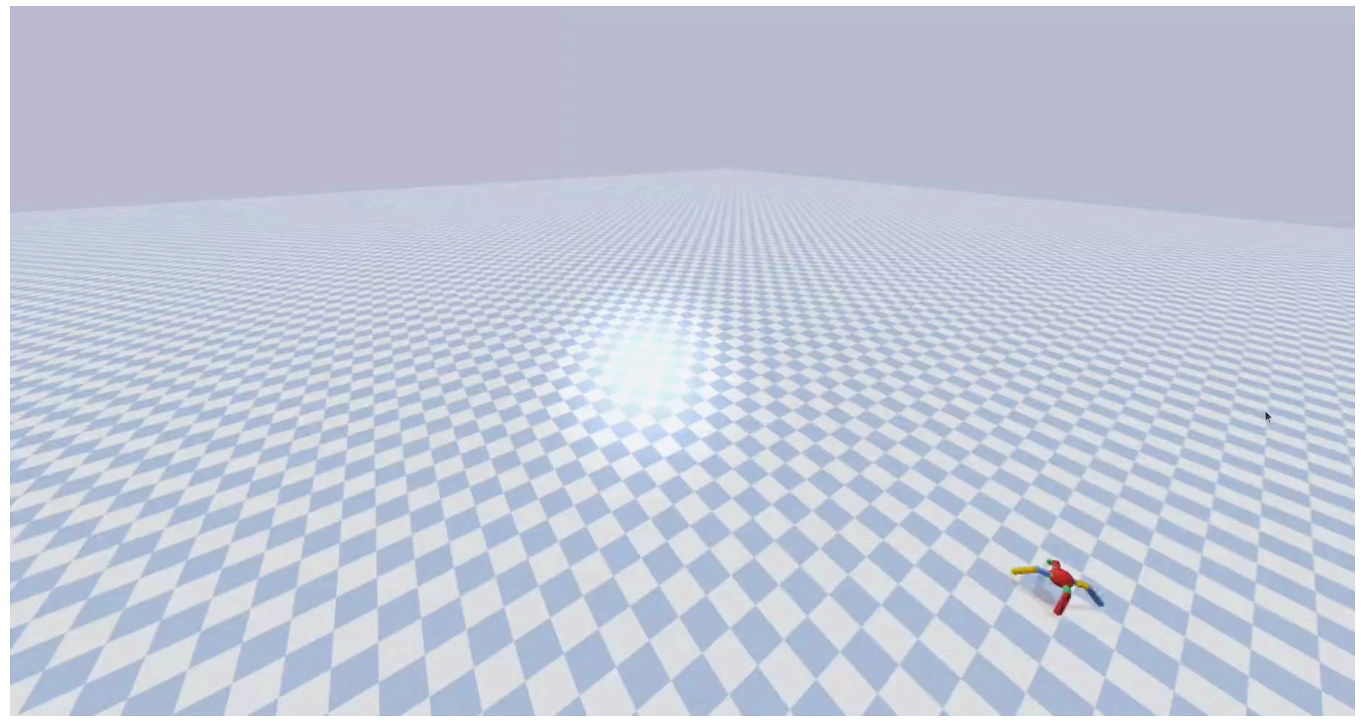
Reward per episode



PPO



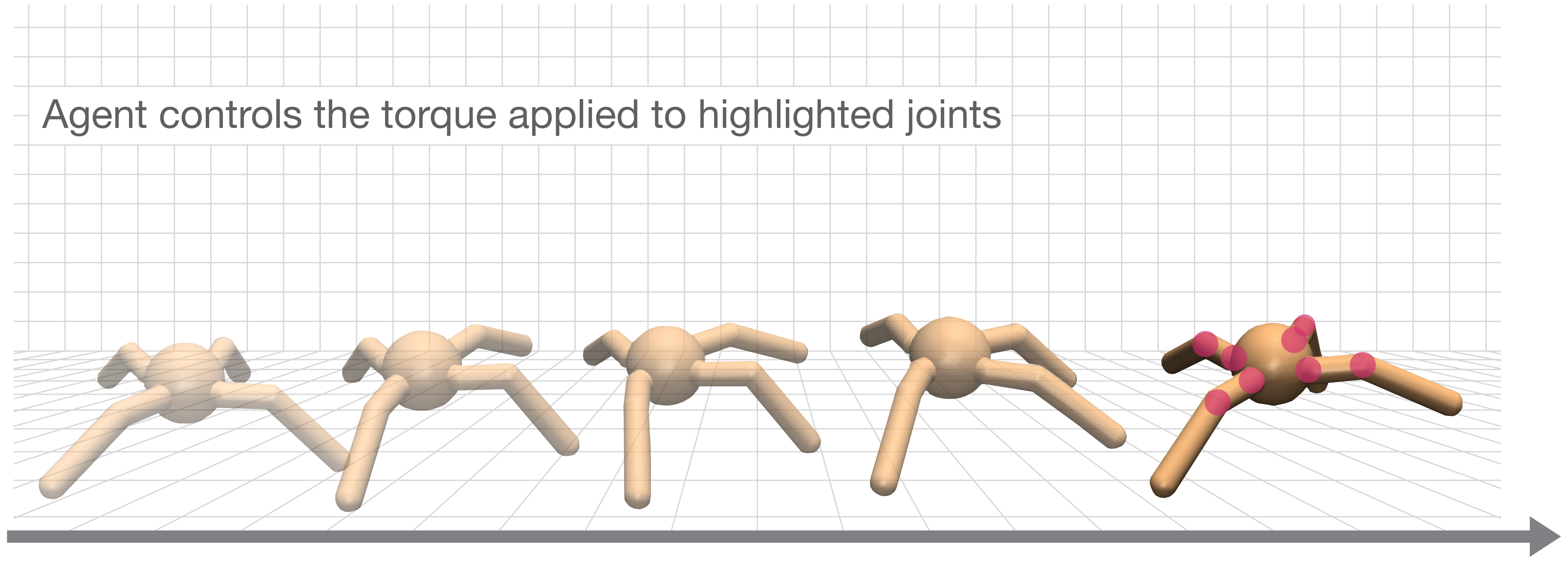
Continual PPO



“Loss of Plasticity in Deep Continual Learning”
by Dohare, Hernandez-Garcia, Lan, Rahman,
Mahmood, & Sutton, *Nature 632*, August 22, 2024

Loss of Plasticity in Reinforcement Learning

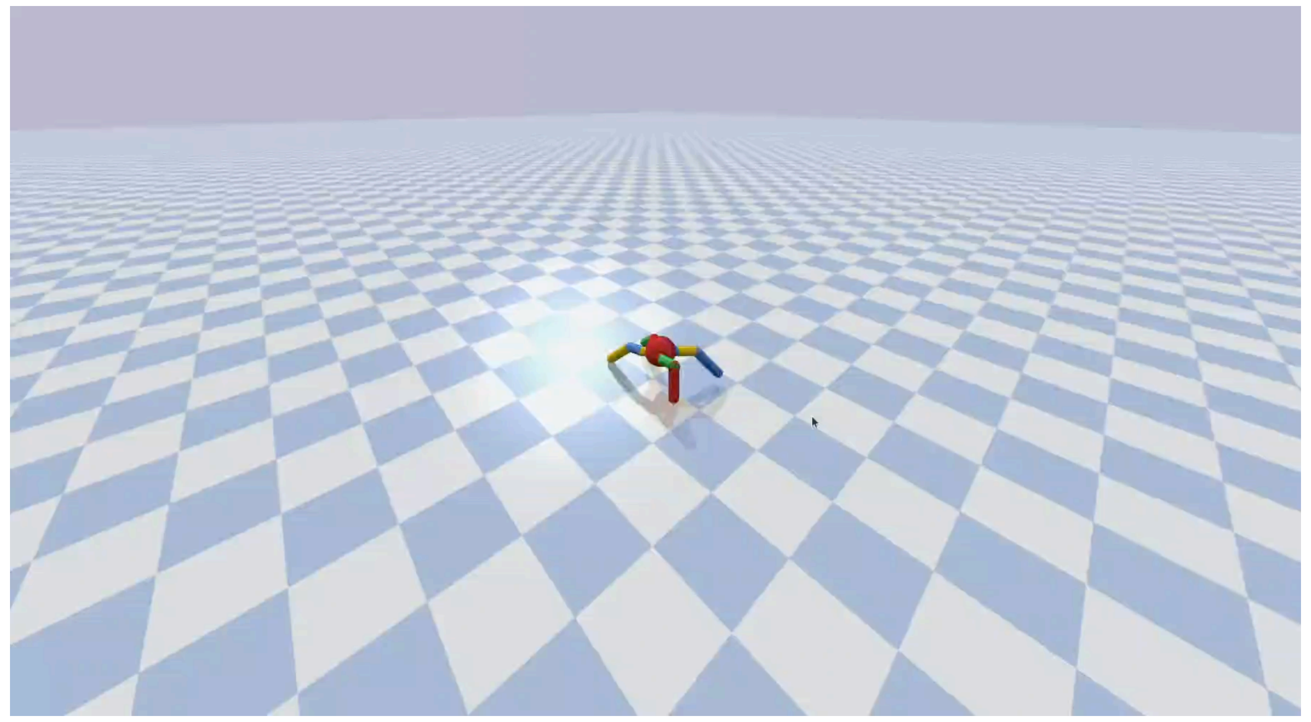
a **Ant locomotion**



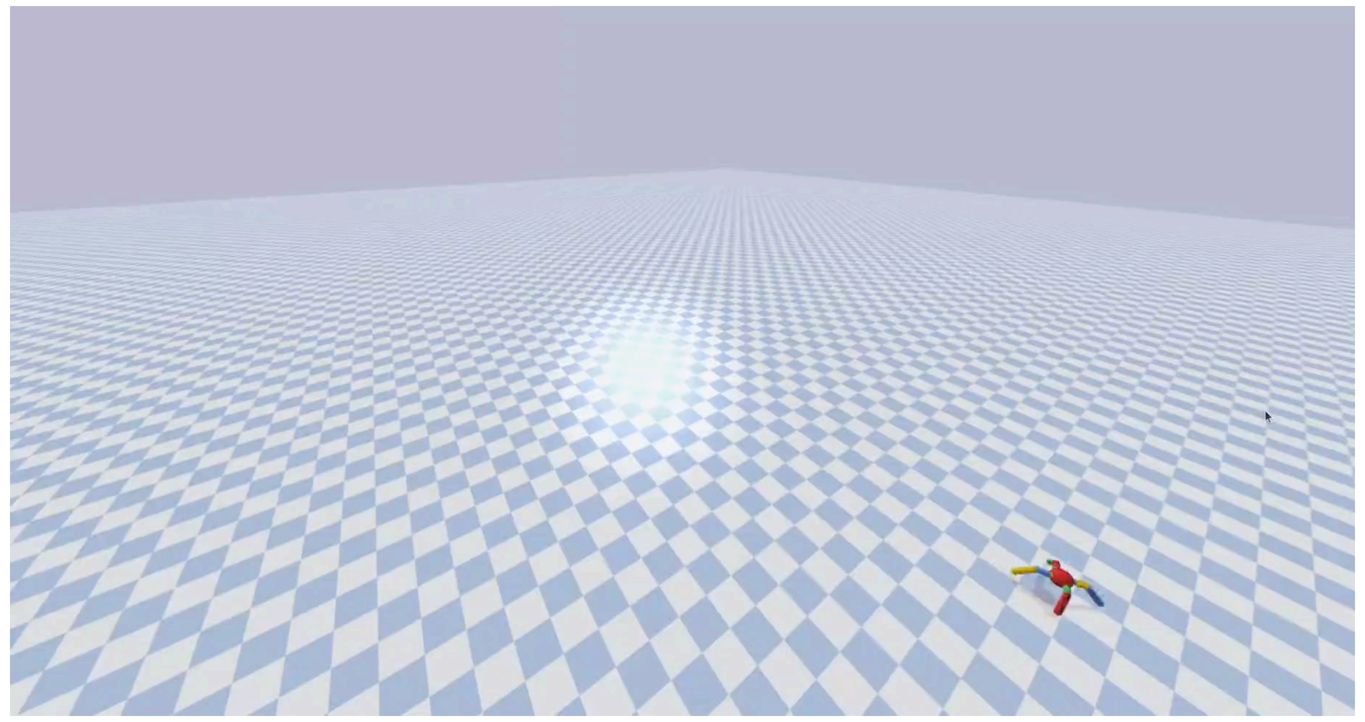
Agent controls the torque applied to highlighted joints

Agent is rewarded for foward motion and penalized if applied torque or contact forces are too large

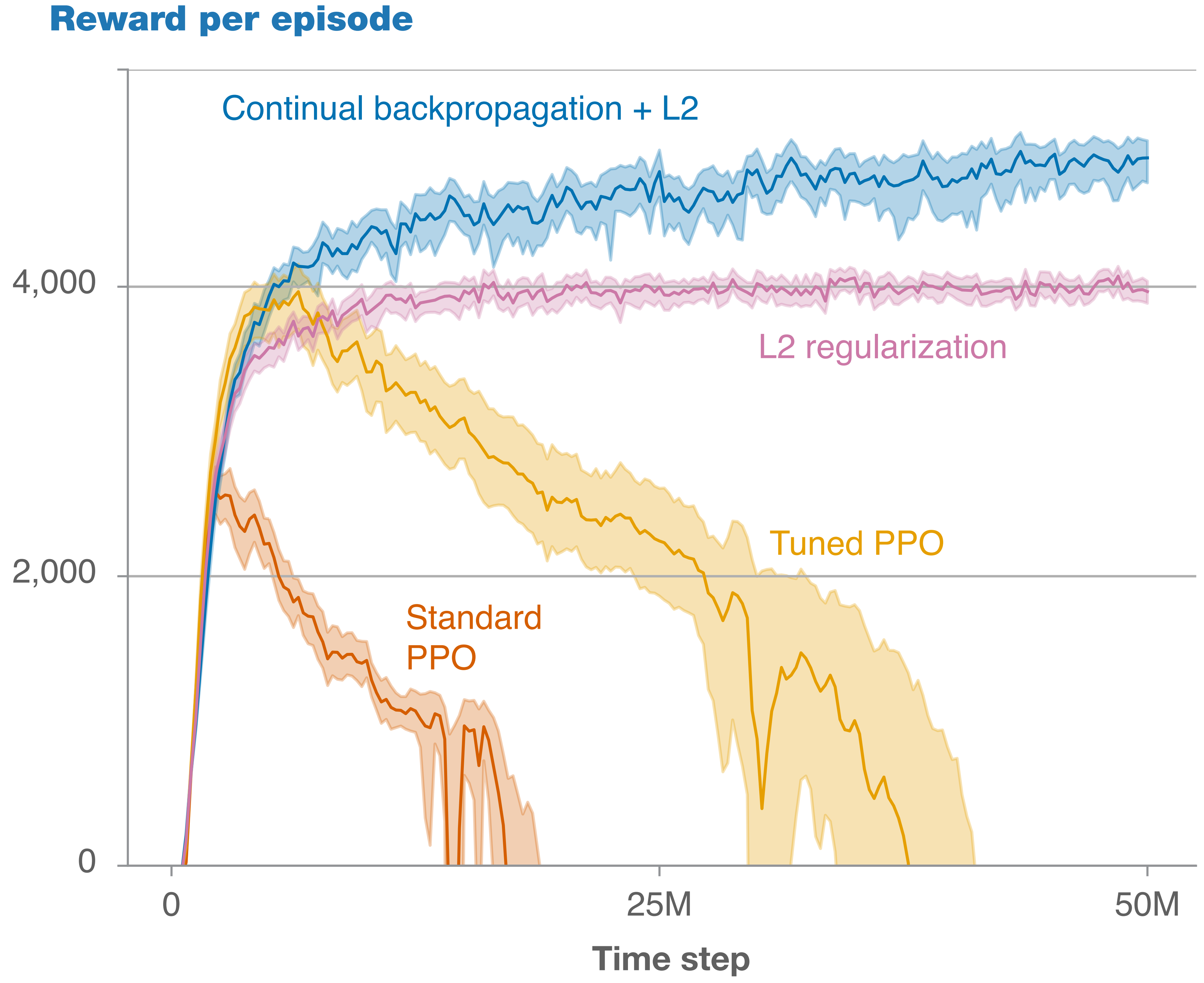
PPO



Continual PPO

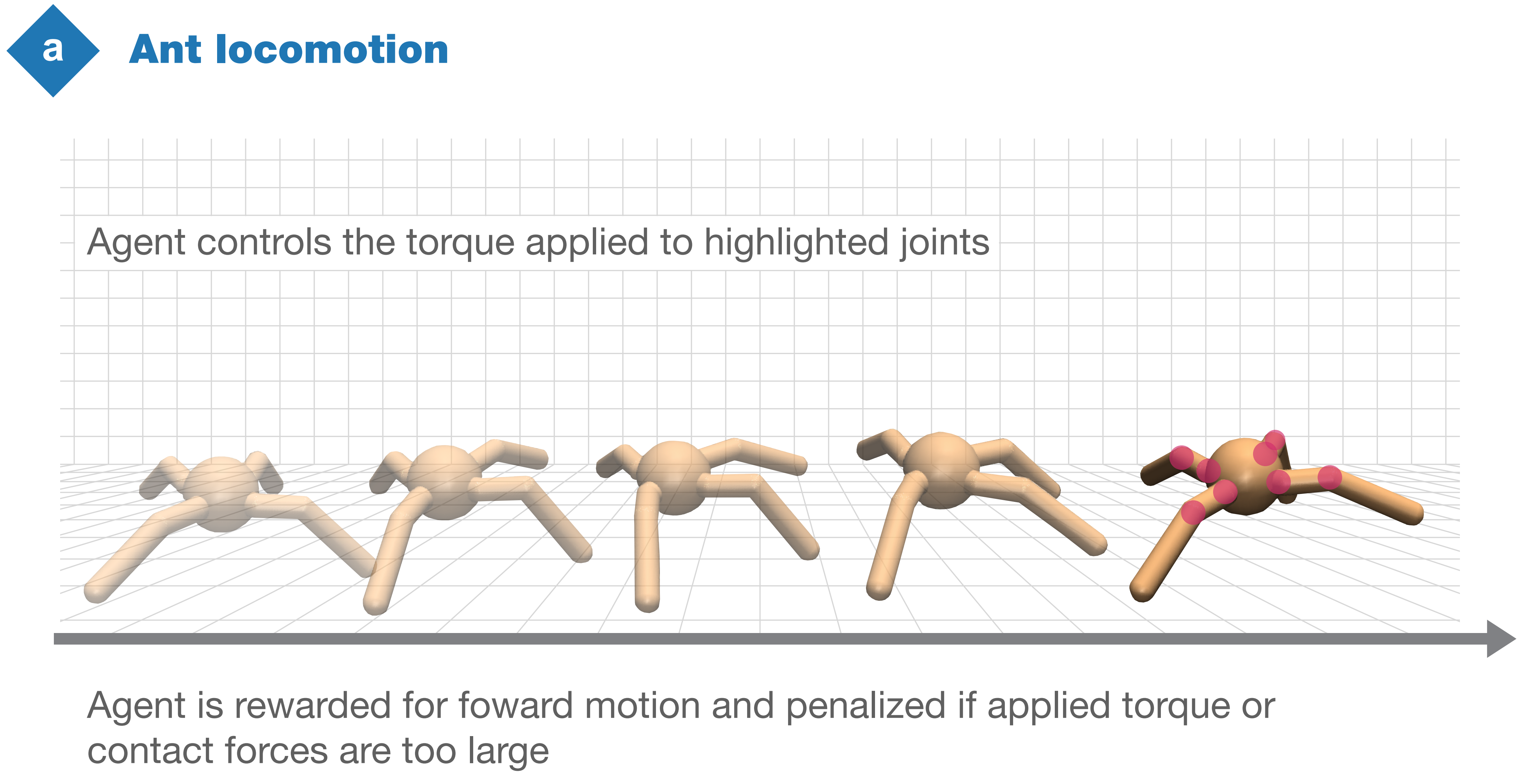


c **Loss of plasticity in ant locomotion**



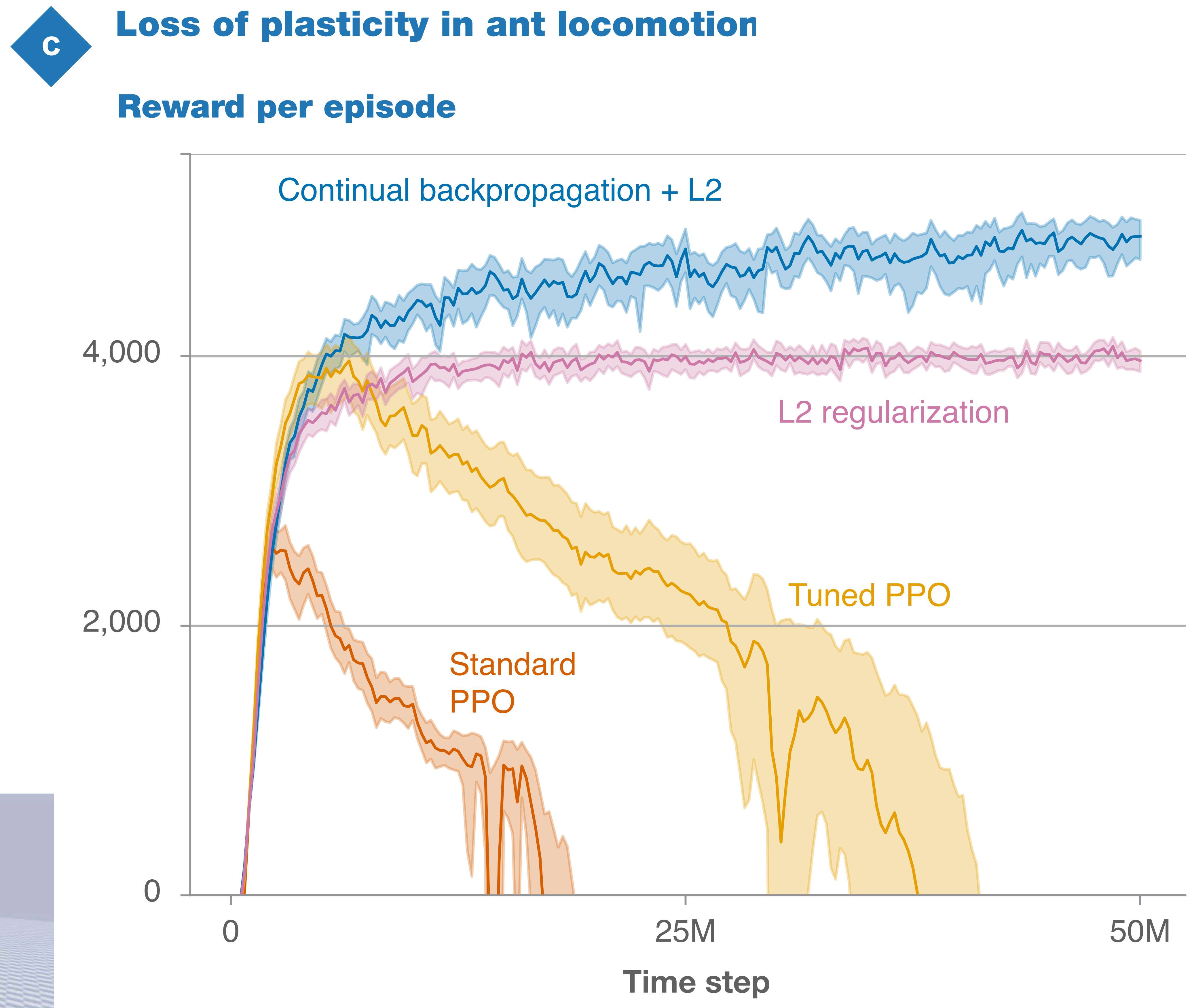
“Loss of Plasticity in Deep Continual Learning” by Dohare, Hernandez-Garcia, Lan, Rahman, Mahmood, & Sutton, *Nature 632*, August 22, 2024

Loss of Plasticity in Reinforcement Learning

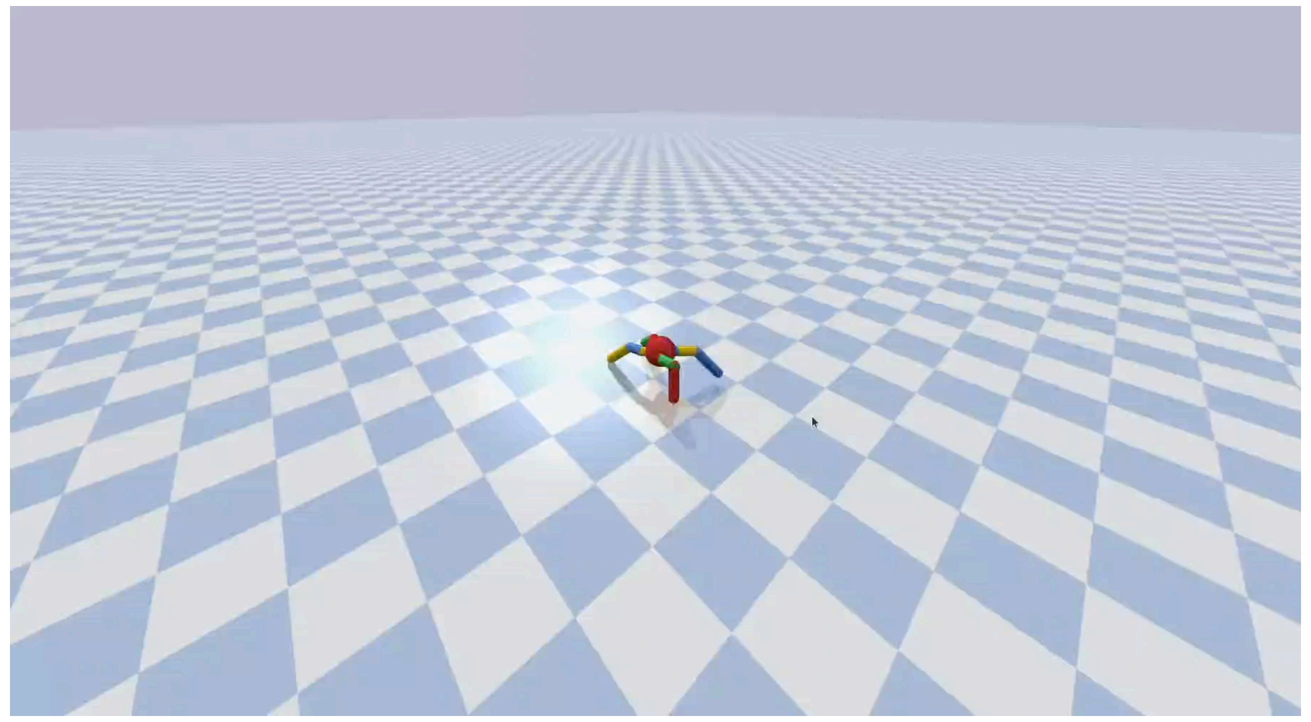


Agent controls the torque applied to highlighted joints

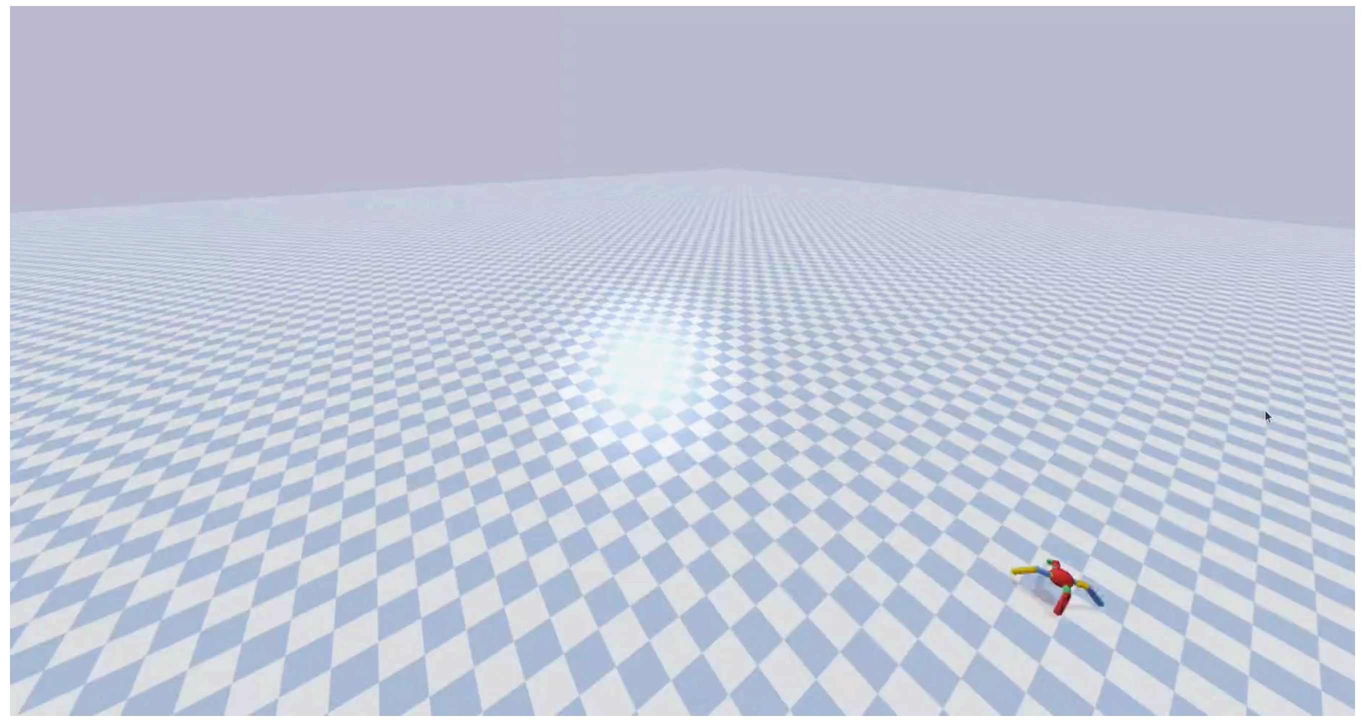
Agent is rewarded for foward motion and penalized if applied torque or contact forces are too large



PPO



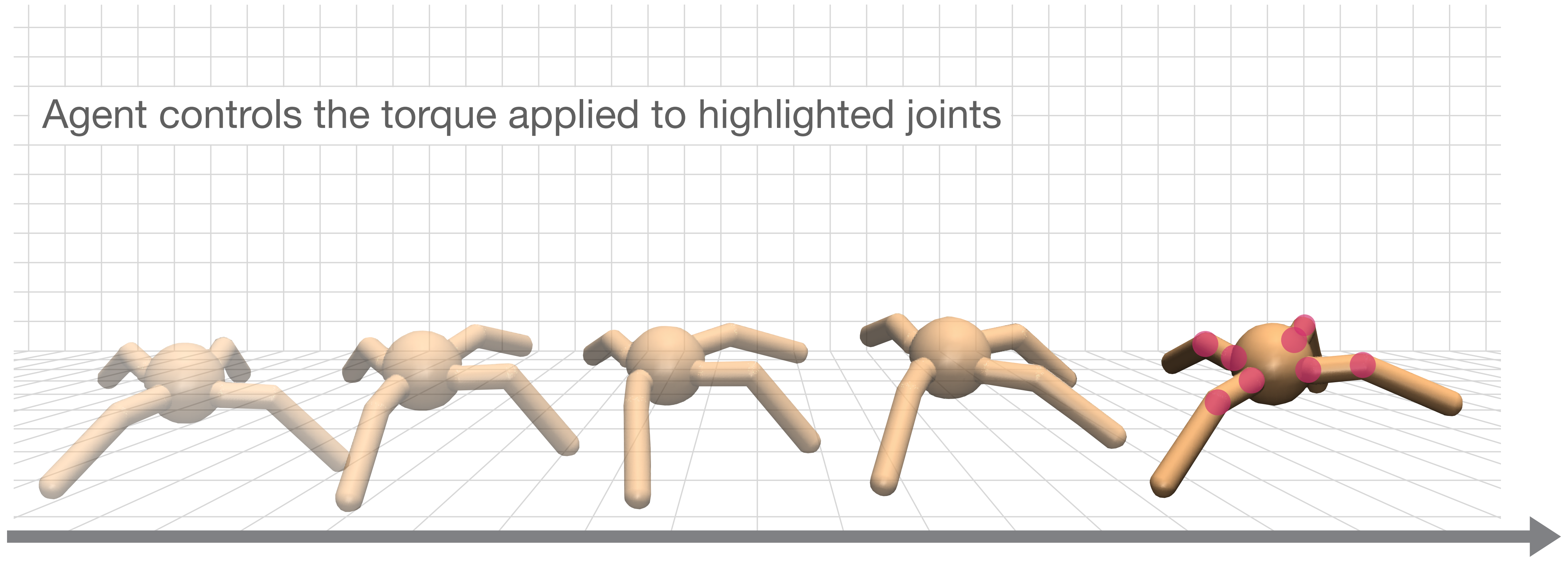
Continual PPO



“Loss of Plasticity in Deep Continual Learning”
by Dohare, Hernandez-Garcia, Lan, Rahman,
Mahmood, & Sutton, *Nature 632*, August 22, 2024

Loss of Plasticity in Reinforcement Learning

a **Ant locomotion**

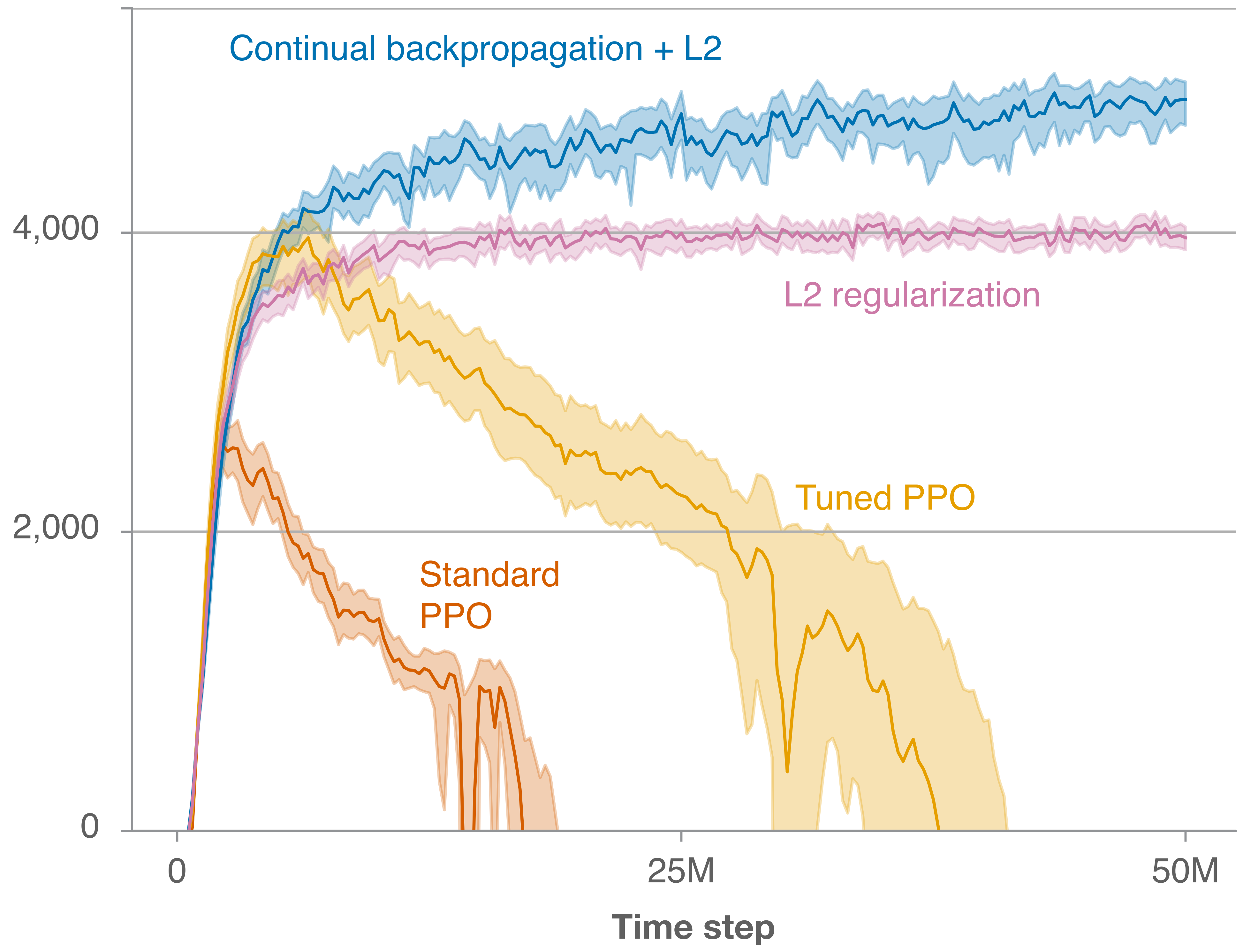


Agent controls the torque applied to highlighted joints

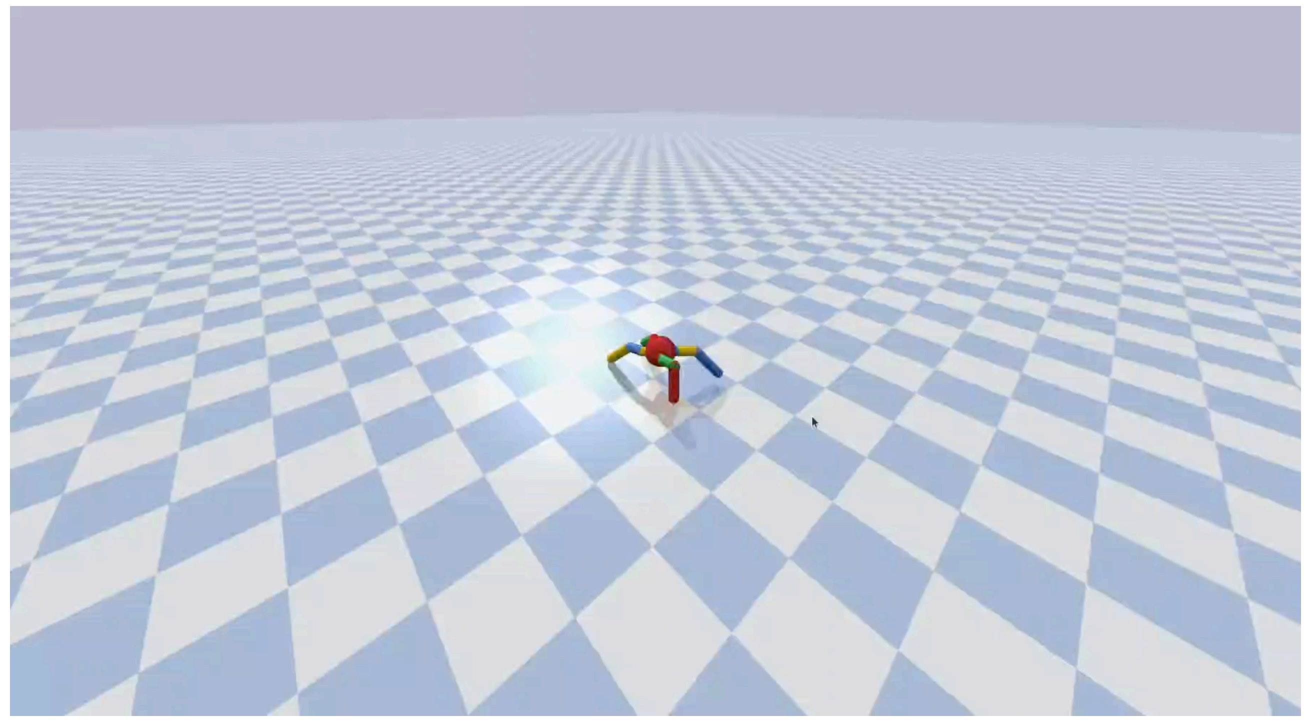
Agent is rewarded for foward motion and penalized if applied torque or contact forces are too large

c **Loss of plasticity in ant locomotion**

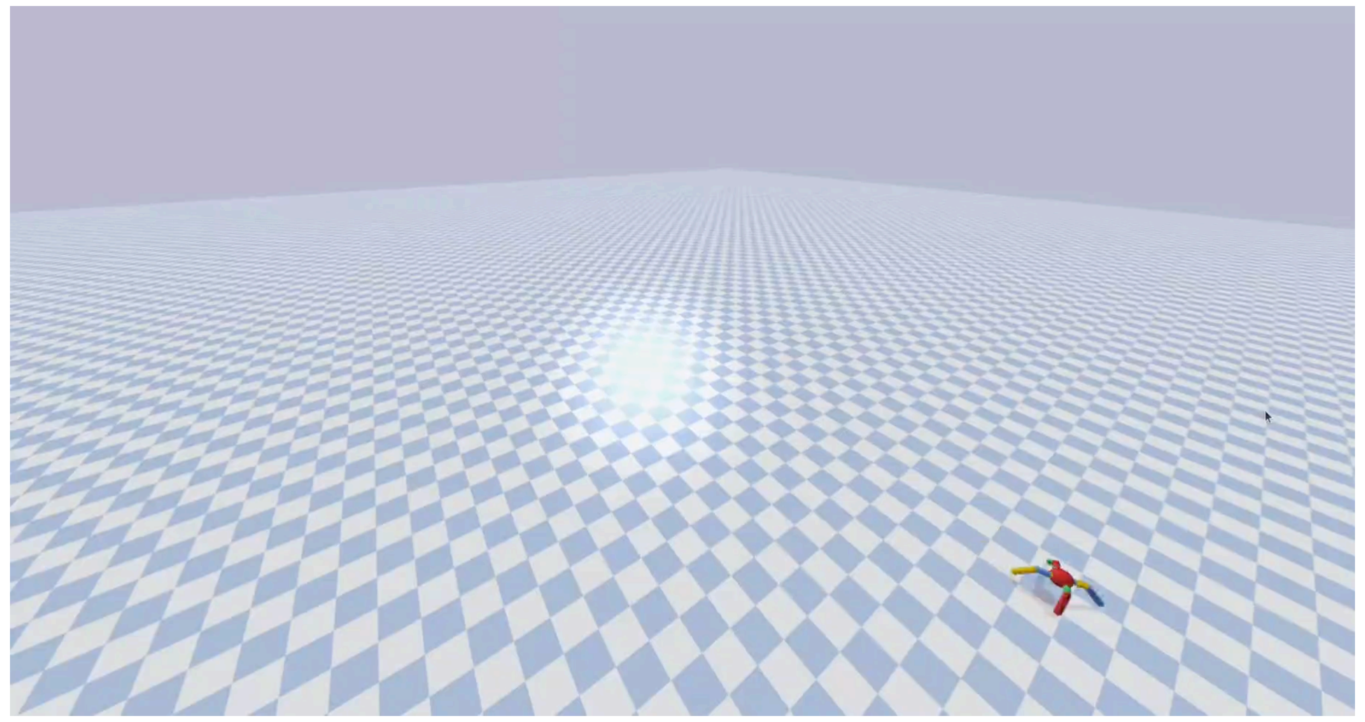
Reward per episode



PPO



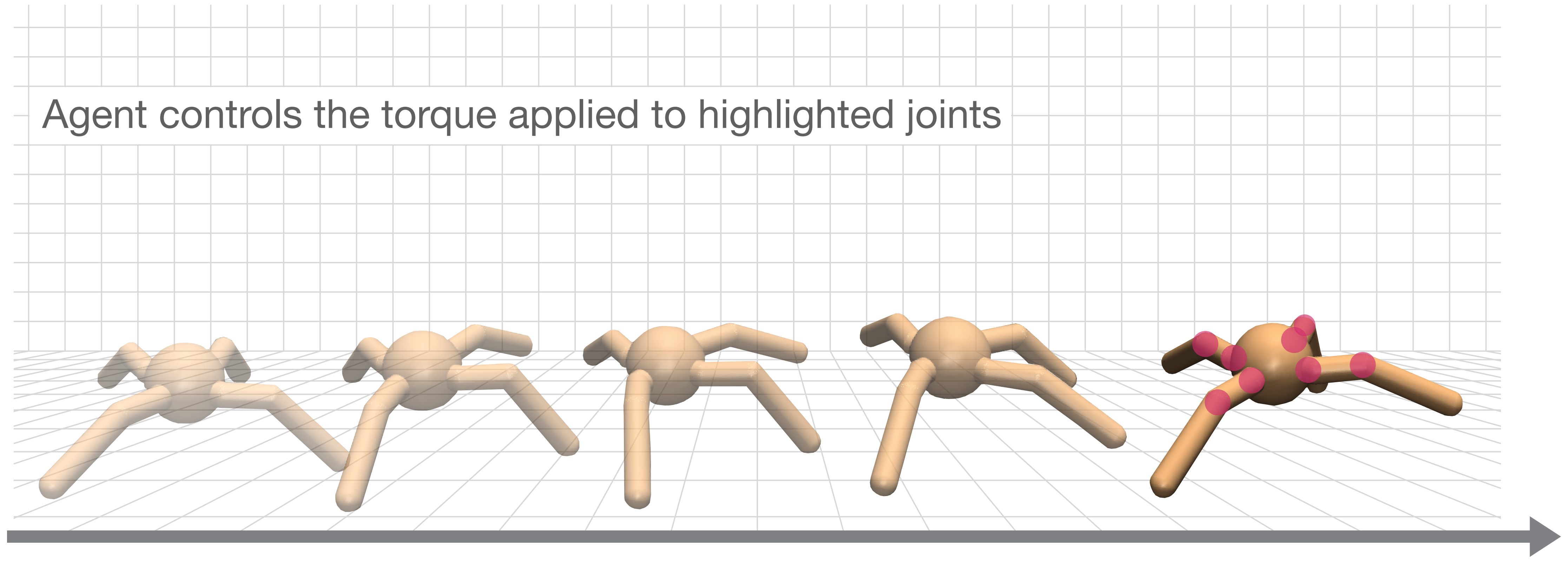
Continual PPO



“Loss of Plasticity in Deep Continual Learning” by Dohare, Hernandez-Garcia, Lan, Rahman, Mahmood, & Sutton, *Nature 632*, August 22, 2024

Loss of Plasticity in Reinforcement Learning

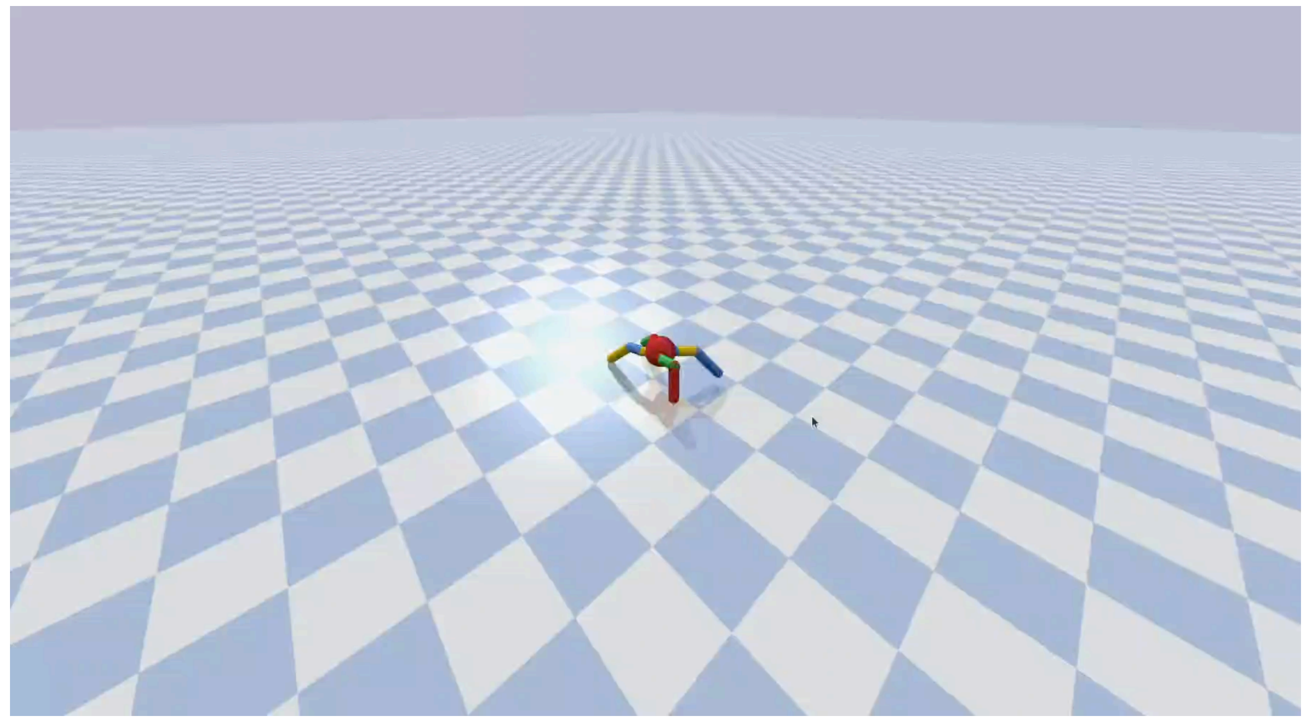
a **Ant locomotion**



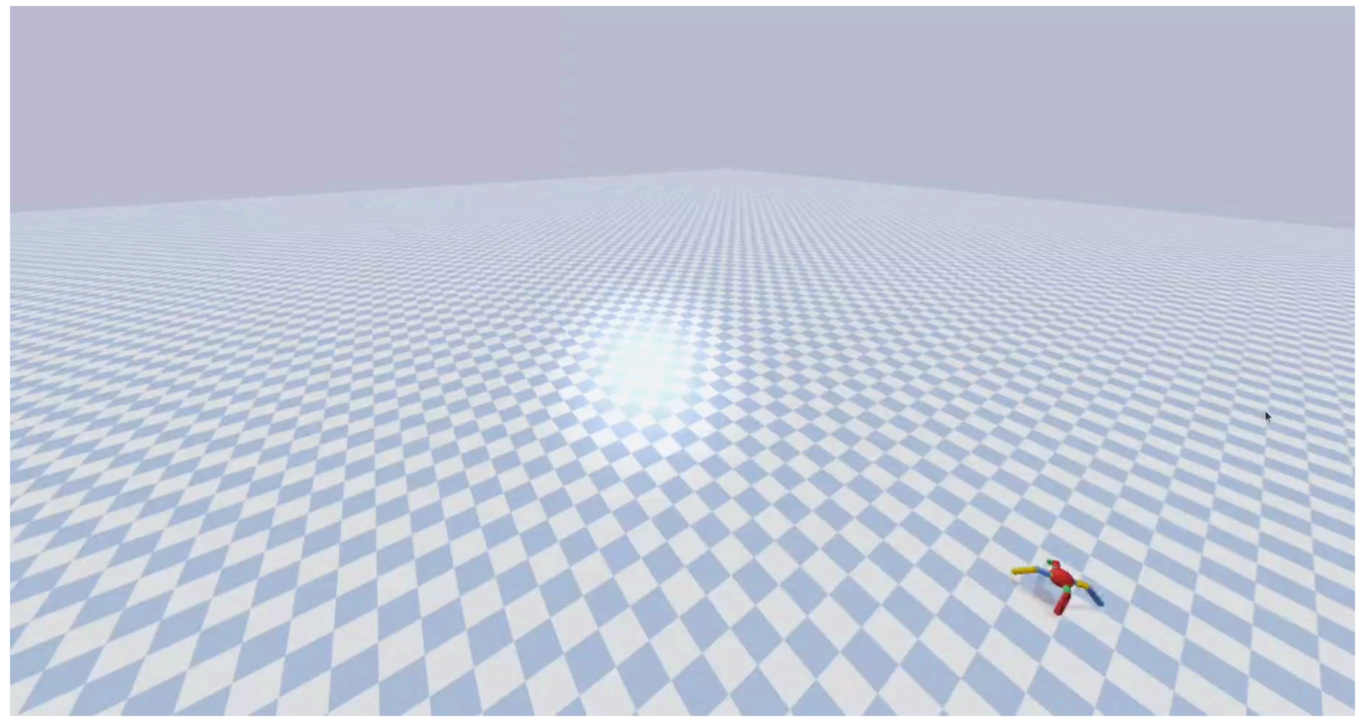
Agent controls the torque applied to highlighted joints

Agent is rewarded for foward motion and penalized if applied torque or contact forces are too large

PPO

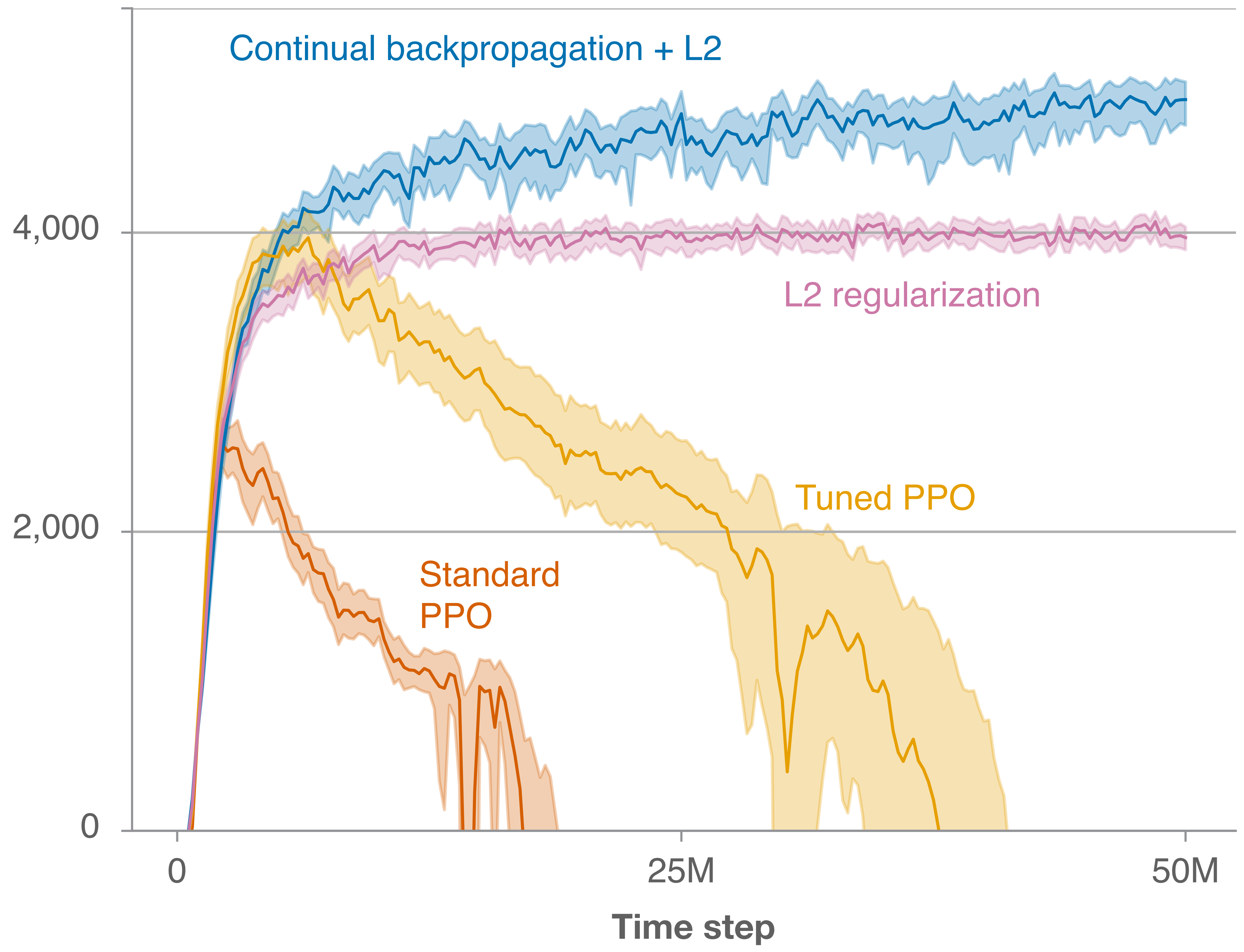


Continual PPO



c **Loss of plasticity in ant locomotion**

Reward per episode

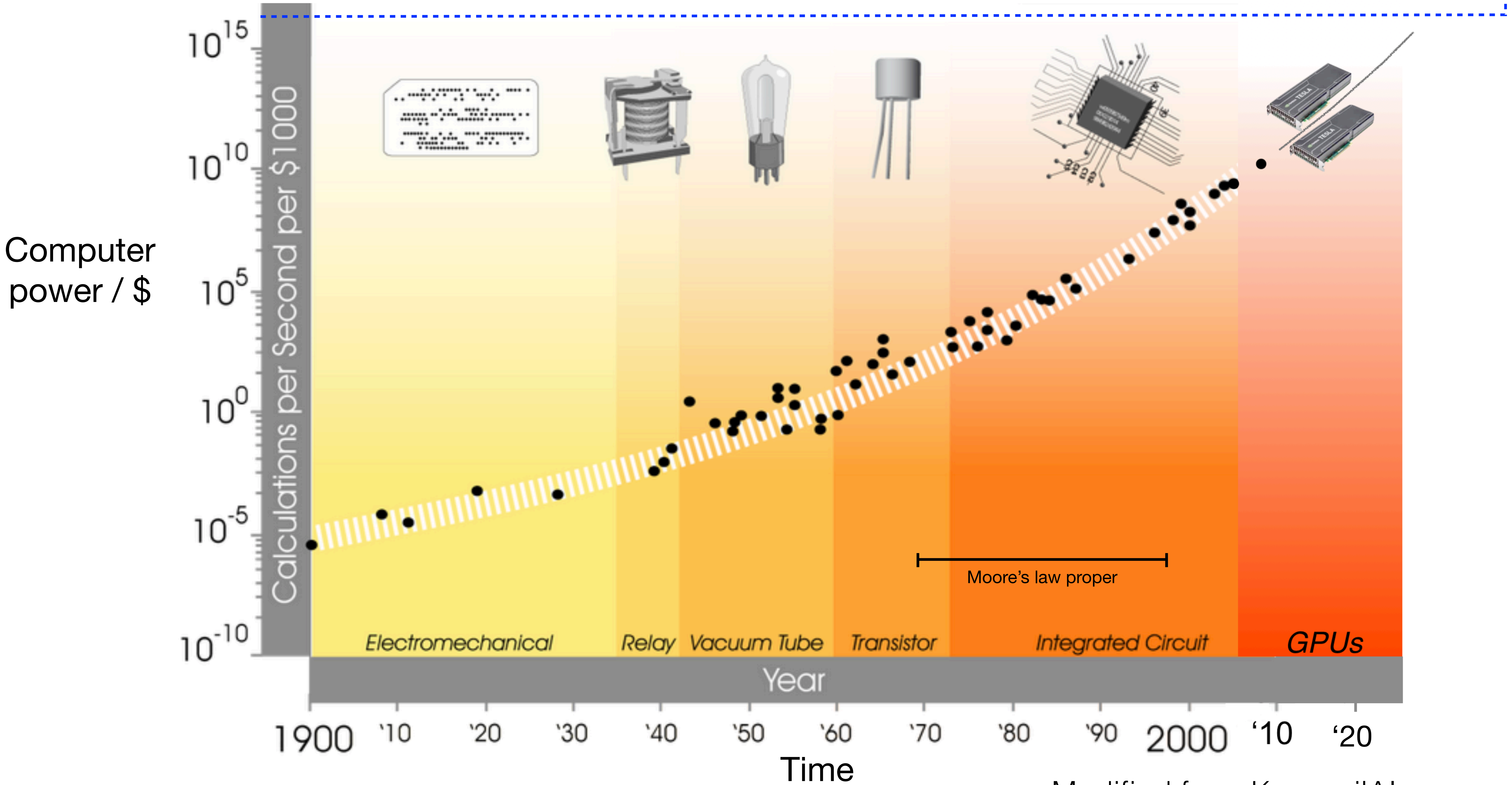


“Loss of Plasticity in Deep Continual Learning” by Dohare, Hernandez-Garcia, Lan, Rahman, Mahmood, & Sutton, *Nature 632*, August 22, 2024

Why is AI happening now?

Computer power/$ is increasing exponentially, with no end in sight, creating a powerful persistent pressure for understanding intelligence

“Moore’s Law” — The tradeoff between time, money, and computer power



Modified from KurzweilAI

Brain-scale computer power will cost ≈$1000 in ≈2030

This estimate is rough but robust: a factor of 10 ≅ 5 years

⇒ AI increases in value by a factor of 10 every 5 years

And so does the pressure to find the algorithms/software

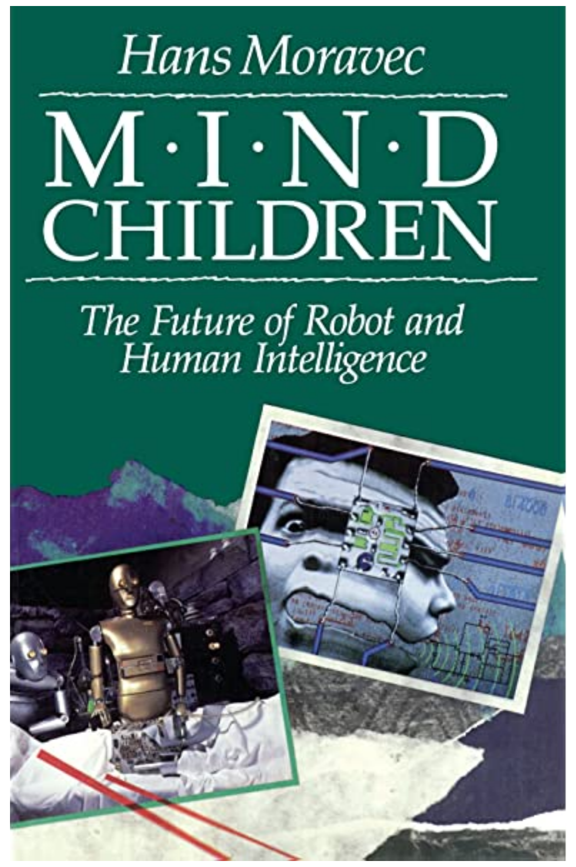
I estimate a 50% probability of human-level AI by 2040

In the ascent of humanity, succession to AI is inevitable

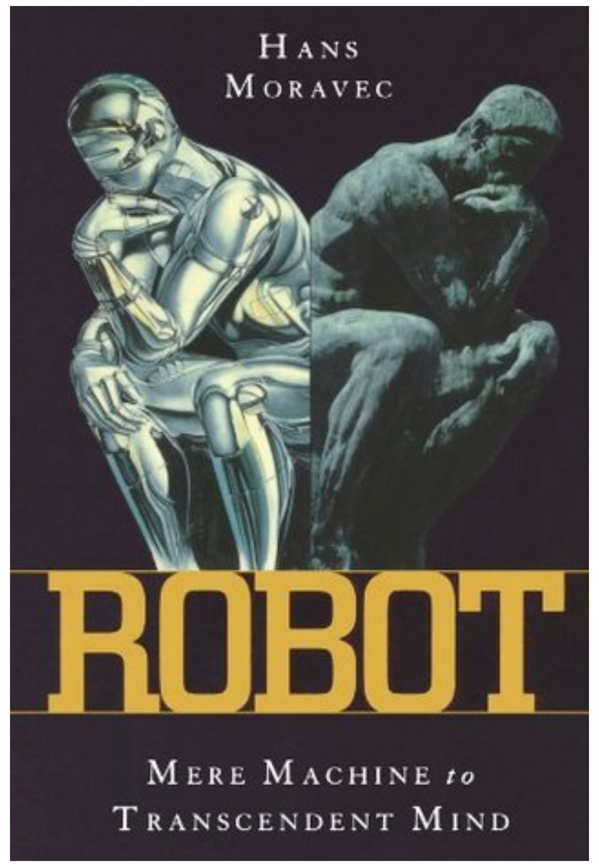
Dr. Hans Moravec (1948–)

AI researcher, Carnegie-Mellon University

On the ascent from man to AI:



1988



1998

- Barring cataclysms,
  I consider the development of intelligent machines a near-term inevitability...
- Rather quickly, they could displace us from existence
- I'm not as alarmed as many...since I consider these future machines our progeny, “mind children” built in our image and likeness, ourselves in more potent form...
  - They will embody humanity's best hope for a long-term future
  - It behooves us to give them every advantage,
    and to bow out when we can no longer contribute...

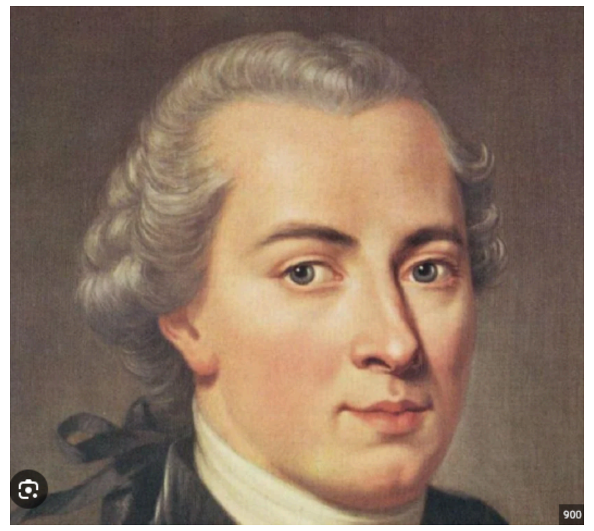
Robot: Mere Machine to Transcendent Mind, Harvard University Press, 1998

AI is not a new and alien technology.
It is one of the oldest of human strivings

- For thousands of years philosophers and ordinary people have sought to understand human intelligence
  - People have always been fascinated by their inner workings
  - How do are minds work? How can we make them work better?
- This is a grand quest, not just narcissism
  - “Intelligence is the most powerful phenomenon in the universe” —Kurzweil
- To understand intelligence is the holy grail of science and the humanities
  - A great and glorious prize!

Philosophy of mind (in the west)

John Locke wrote “An Essay Concerning Human Understanding”



Emmanuel Kant wrote “The Critique of Pure Reason”

Rene Descartes said “i think, therefore i am”

Scientists and non-scientists have been fascinated by their inner workings

Gustav Fechner

Hermann Ebbinghaus

Ivan Pavlov

Edward Thorndike

B. F. Skinner

Edward Tolman

Jean Piaget

Sigmund Freud

Carl Jung

Timothy Leary

Ray Kurzweil

AI is not a new and alien technology.
It is one of the oldest of human strivings

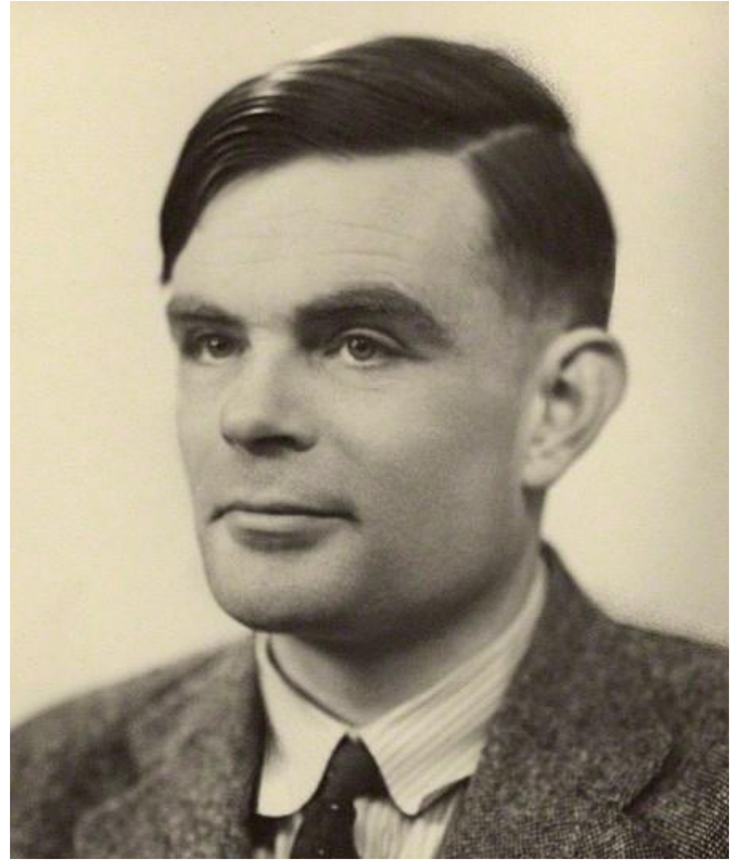
- For thousands of years philosophers and ordinary people have sought to understand human intelligence
  - People have always been fascinated by their inner workings
  - How do are minds work? How can we make them work better?
- This is a grand quest, not just narcissism
  - “Intelligence is the most powerful phenomenon in the universe” —Kurzweil
- To understand intelligence is the holy grail of science and the humanities
  - A great and glorious Prize!

Can we define “intelligence”?

Intelligence is:

Can we define “intelligence”?

Intelligence is:

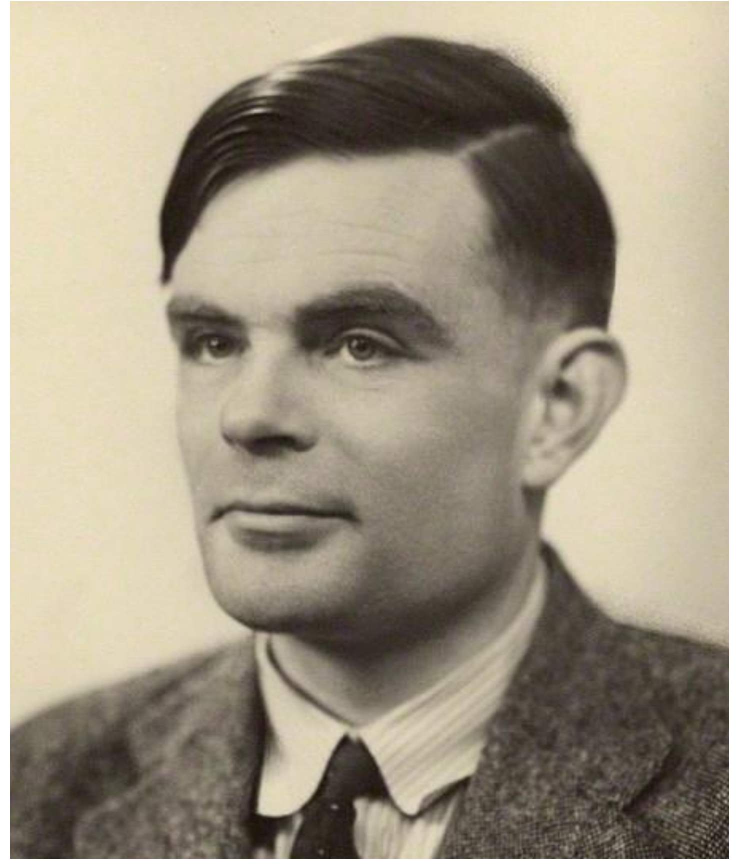


“behaving like a person” (the Turing Test)

—Alan Turing? 1950?
Founding father of CS

Can we define “intelligence”?

Intelligence is:

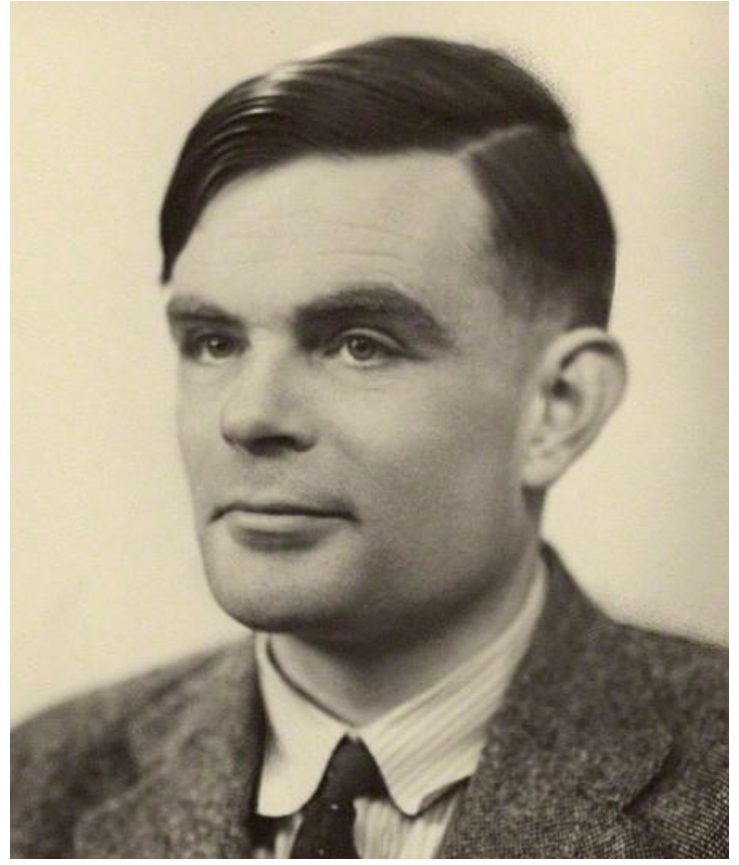


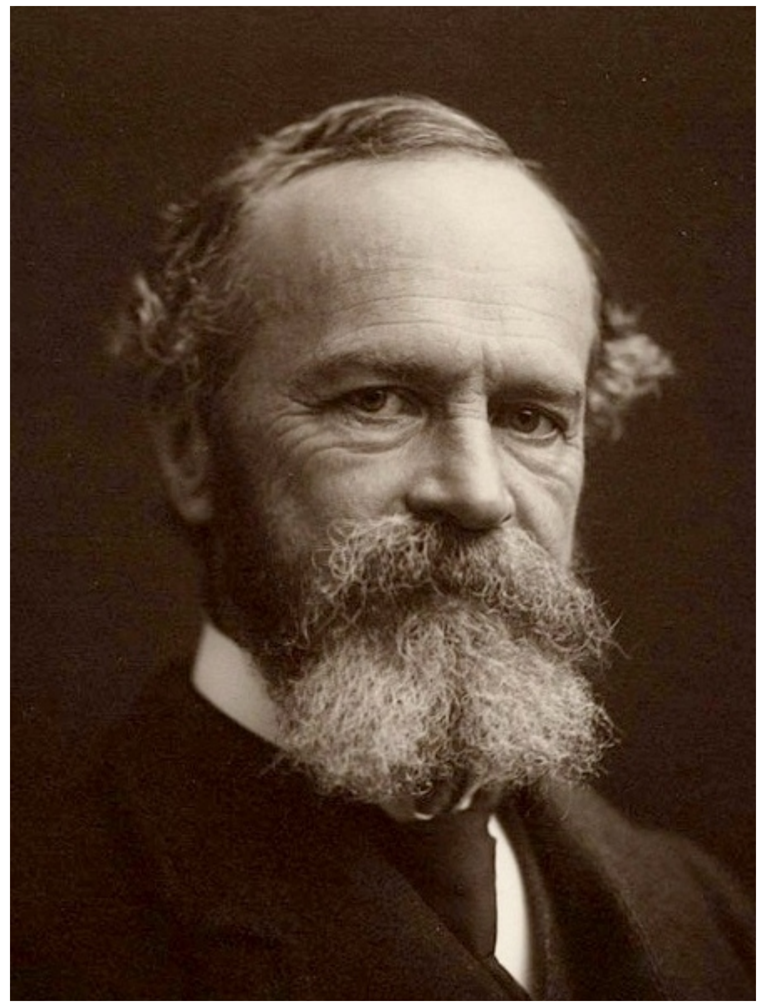
“behaving like a person” (the Turing Test) —Alan Turing? 1950?
Founding father of CS

“the ability to acquire and apply knowledge and skills” —Dictionary

Can we define “intelligence”?

Intelligence is:





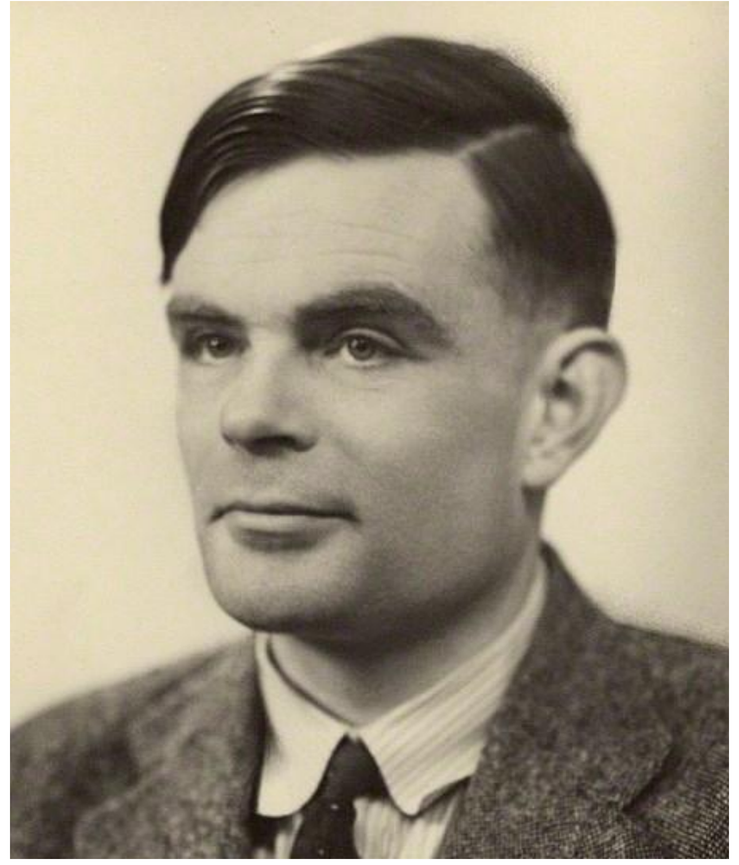
“behaving like a person” (the Turing Test) —Alan Turing? 1950?
Founding father of CS

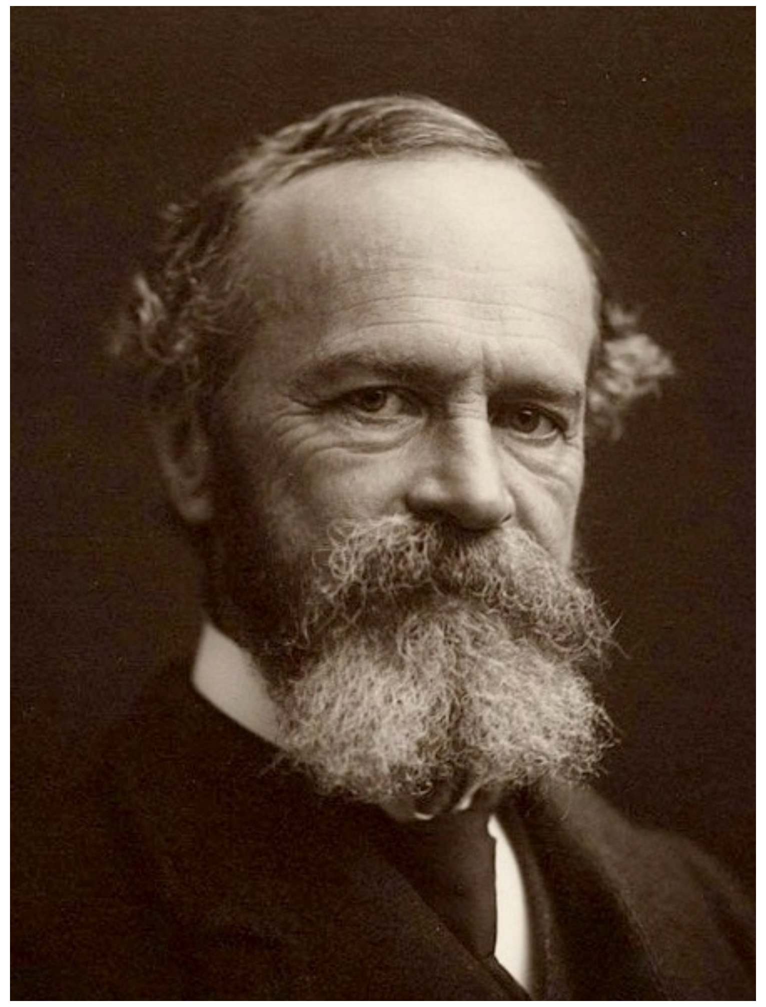
“the ability to acquire and apply knowledge and skills” —Dictionary

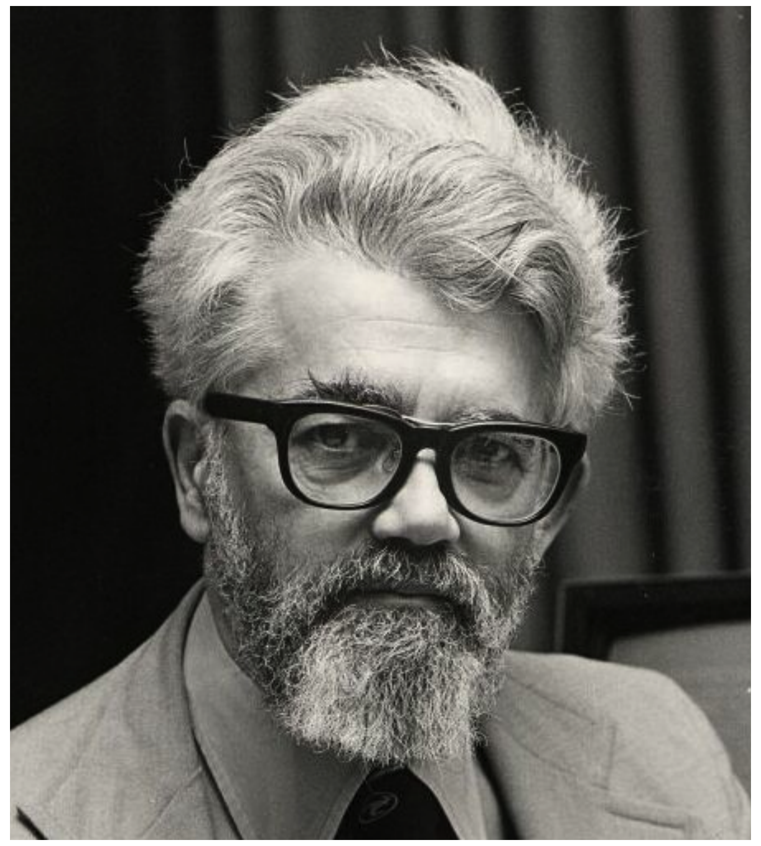
“attaining consistent ends by variable means” —William James, 1890
Founding father of Psych

Can we define "intelligence"?

Intelligence is:







"behaving like a person" (the Turing Test) —Alan Turing? 1950?
Founding father of CS

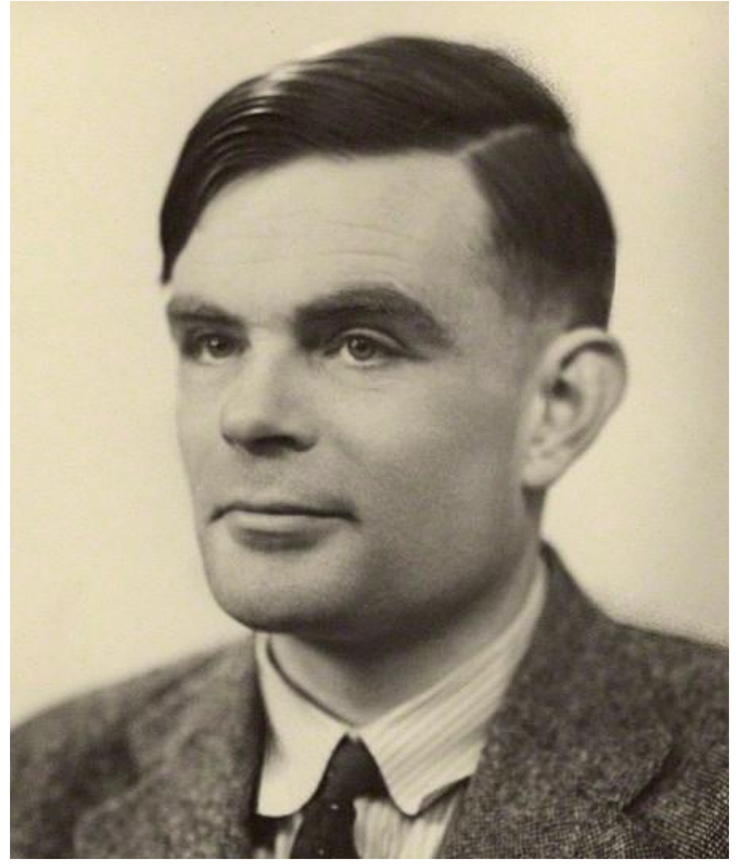
"the ability to acquire and apply knowledge and skills" —Dictionary

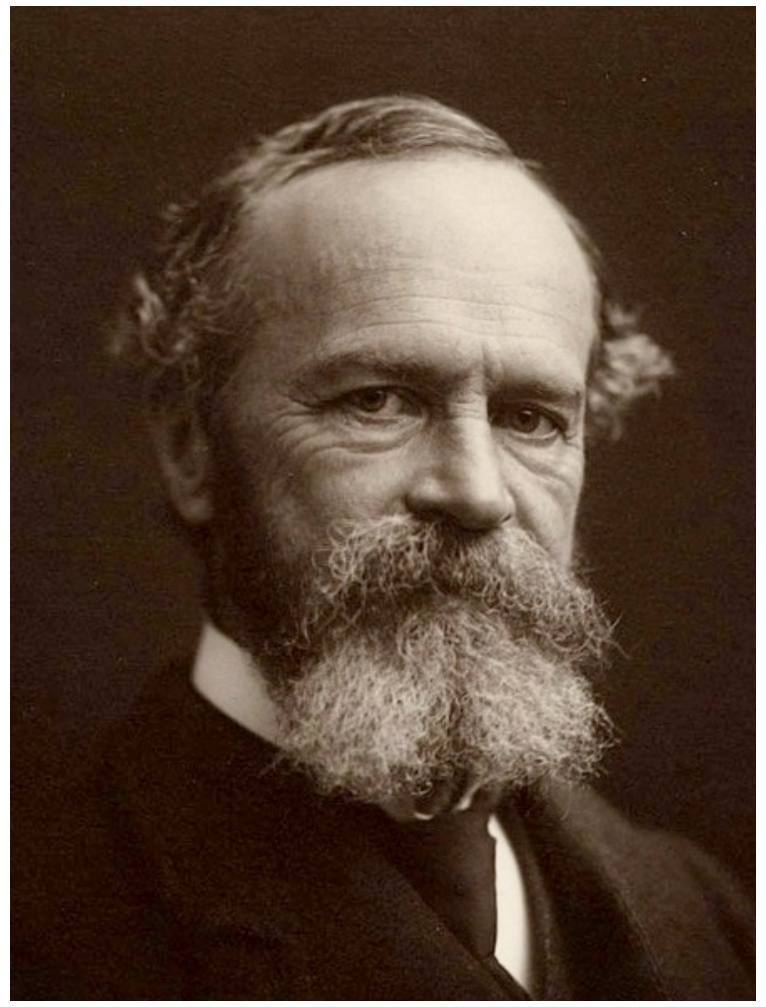
"attaining consistent ends by variable means" —William James, 1890
Founding father of Psych

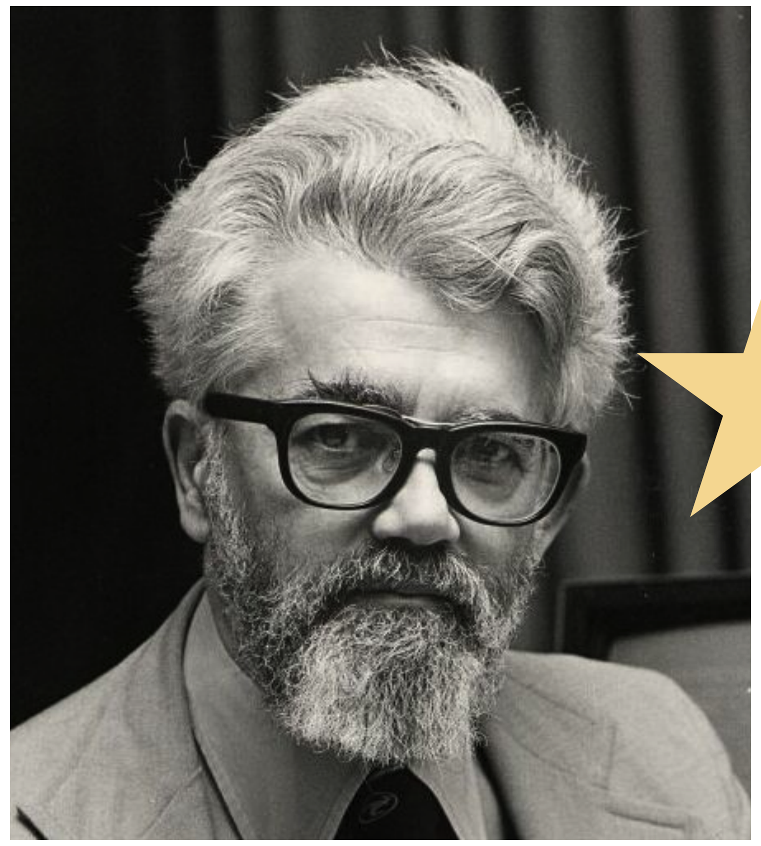
"the computational part of the ability to achieve goals" —John McCarthy, 1997
Founding father of AI

Can we define “intelligence”?

Intelligence is:







“behaving like a person” (the Turing Test) —Alan Turing? 1950?
Founding father of CS

“the ability to acquire and apply knowledge and skills” —Dictionary

“attaining consistent ends by variable means” —William James, 1890
Founding father of Psych

★ “the computational part of the ability to achieve goals” —John McCarthy, 1997
Founding father of AI

One goal, or to each his own?

- In reinforcement learning, each intelligent agent has its own goal
- Just as, in nature, each animal has its own pains and pleasures
- In AI and in nature, different agents have different goals
- In fact, our economies work best when different people have different goals and different abilities
  - they *don't* rely on people having a shared goal, a common purpose
- Decentralization is many agents, each pursuing own goal
- Cooperation is agents with different goals interacting to mutual benefit

Agents can live in peace, even when they all want different things

We are “homo cooperativus”;
We cooperate more than any other animal

- Cooperation is facilitated by language and money (both unique to humans)
- Humanity’s greatest successes are cooperations: economies, markets, governments
- Humanity’s greatest failures are failures to cooperate: war, theft, corruption
- Decentralized cooperation is an alternative to common purpose
    - In my view it is more elegant: sustainable, robust, adaptive, flexible
- Humans are better at cooperation than any other animal,
  but we are still terrible at it—we still have wars, theft, corruption, fraud

We struggle to cooperate—it's not easy

- Cooperation is not always possible — it takes two trustworthy agents
- There are always some who benefit from not cooperating:
  cheats, thieves, con men, weapons manufacturers, dictators
- Cooperation needs institutions to facilitate it
  and to punish cheaters, thieves, fraudsters, extortionists
- A centralized authority can help cooperation in the short term,
  but poison it in the long run (authoritarian and sclerotic governments)
  - Centralized control is the opposite of decentralized cooperation

There are many calls for centralized control of AI

- For controlling AI’s goals
- For pausing or stopping AI research
- For limiting the computer power of AIs
- For ensuring “safety” of AI
- For requiring disclosures of AI

There are many calls for centralized control of people

- For controlling speech and media
- For controlling trade
- For controlling employment
- For controlling finance
- For economic sanctions

The arguments for centralized control (in both cases) are eerily similar.
They are based in fear. They are all about us vs. them.
They demonize the other. They claim the other can’t be trusted.

In conclusion

- Flourishing comes from decentralized cooperation
- Humans are great at cooperation, but also terrible at it
- Cooperation is not always possible,
  but it is the source of all that is good in the world
  - We must look for it and support it, and seek to institutionalize it
- If we look with open eyes, it is easy to see who is calling for mistrust, non-cooperation, and centralized control; we should resist those calls
- This is a useful lens with which to view all calls for human and AI interaction

Thank you for your attention