

Coachable Agents

Shruti Mishra
Imperial College London

Coachable Agents for interactive gameplay

Shruti Mishra,
while at Sony AI

Coachable Agents for interactive gameplay and directions for flow control

Shruti Mishra
Imperial College London

Coachable Agents for interactive gameplay

Coachable Agents for interactive gameplay

Roberto Capobianco, Harm van Seijen, Nolan Bard, Neil Burch, Fatima D. Davelouis, Josh Davidson, Alisa Devlic, Yunshu Du, Ishan Durugkar, Siddhant Gangapurwala, Daniel Hernandez, G. Zacharias Holland, Sahil Jain, Kenta Kawamoto, Raksha Kumaraswamy, Patrick MacAlpine, Dustin Morrill, Declan Oller, Francesco Riccio, Akanksha Saran, Craig Sherstan, Kaushik Subramanian, Thomas Walsh, Samuel Barrett, Kizza N. Frisbee, Mady Govil, Johannes Günther, Varun R. Kompella, James MacGlashan, Michael Thomure, Jaden B. Travnik, Kevin Waugh, Elahe Aghapour, Florian Fuchs, Andreanne Lemay, Shruti Mishra, Takuma Seno, Peter Stone, Michael Spranger, Peter R. Wurman

Sony AI, Game AI team

interactive gameplay



interactive gameplay

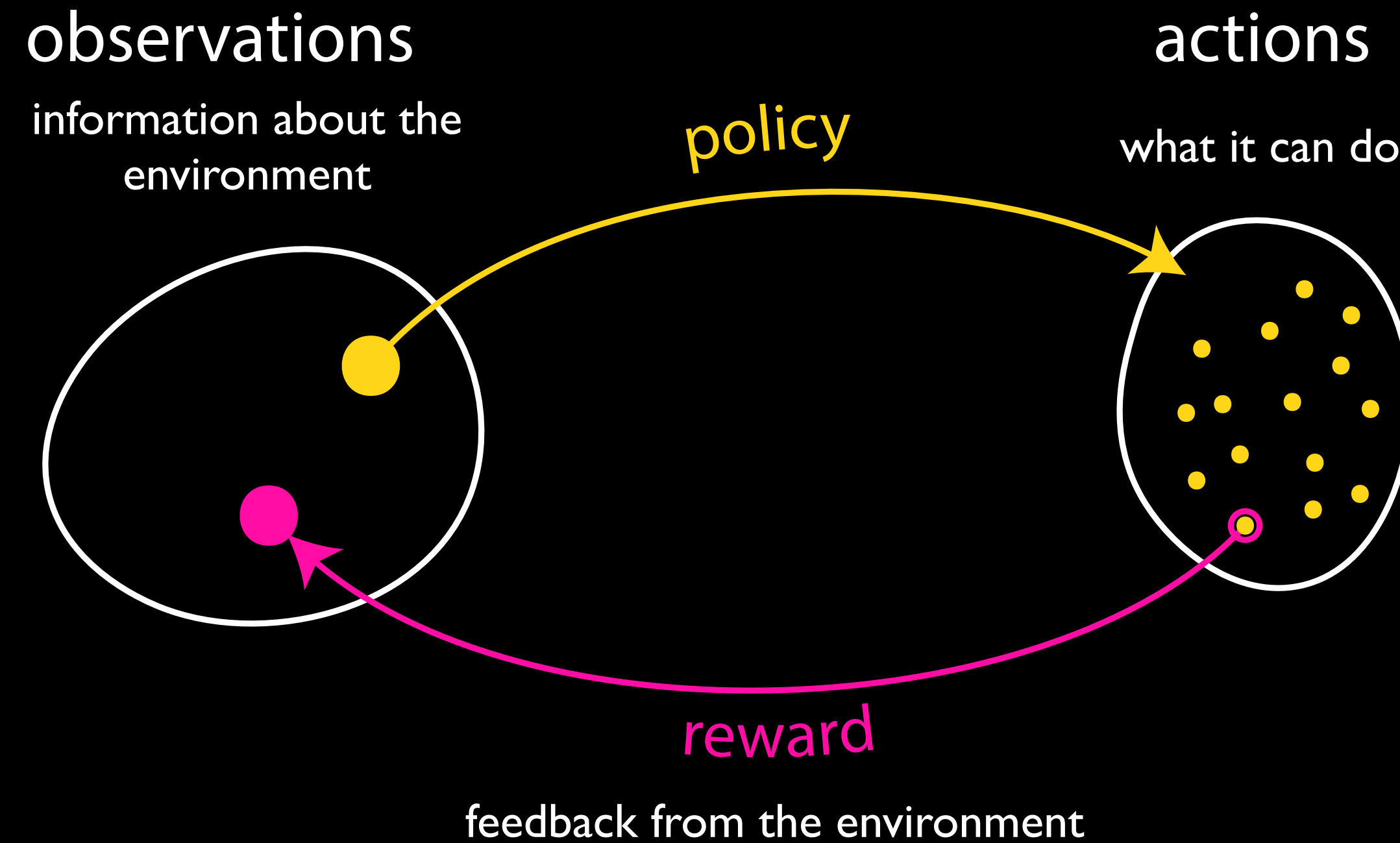


includes a slider interface for controlling behaviour

What are the *Agents*?

Reinforcement learning

Agents

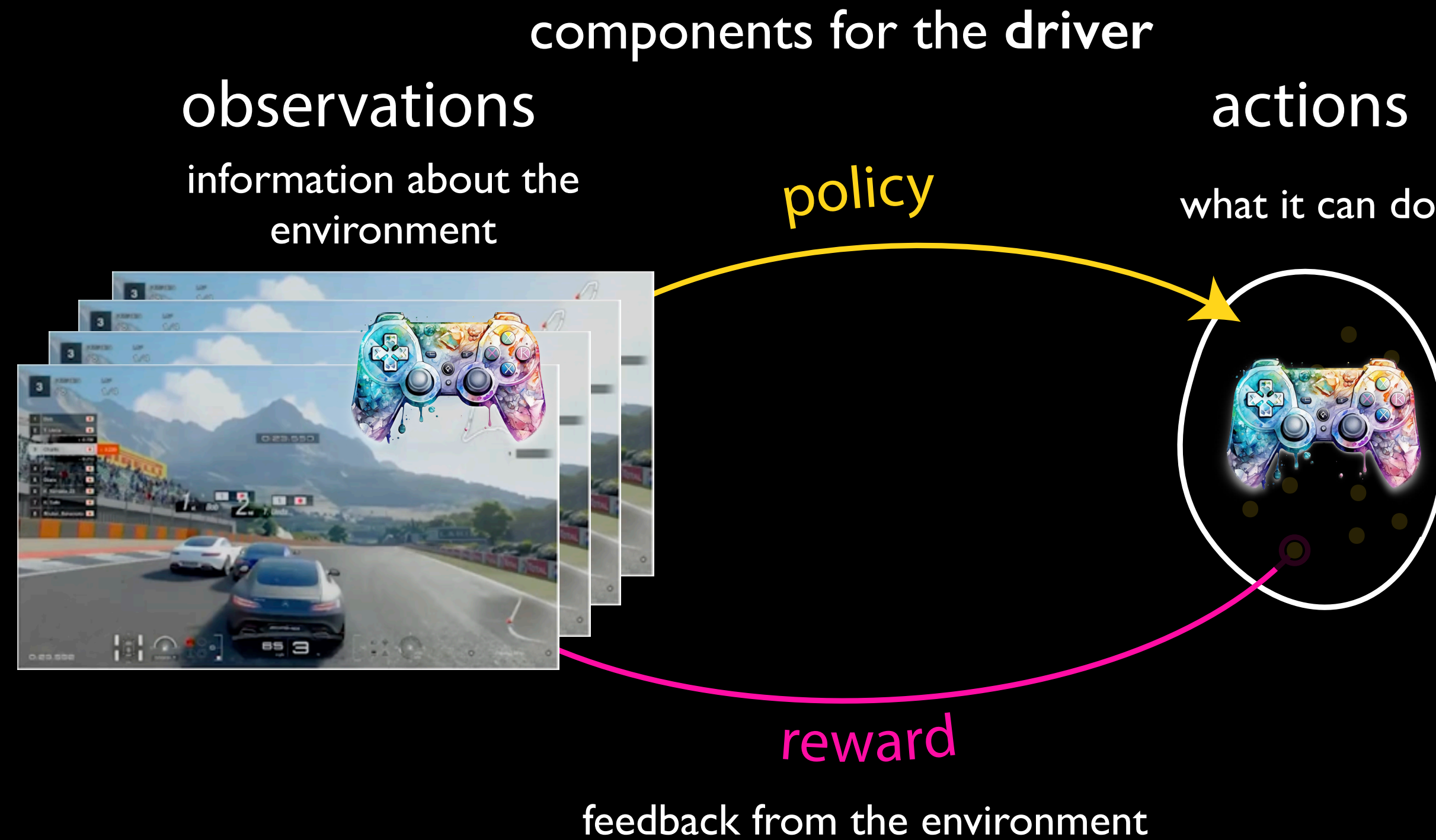


Given an **observation**, an agent chooses an **action** to attain some **long-term goal**

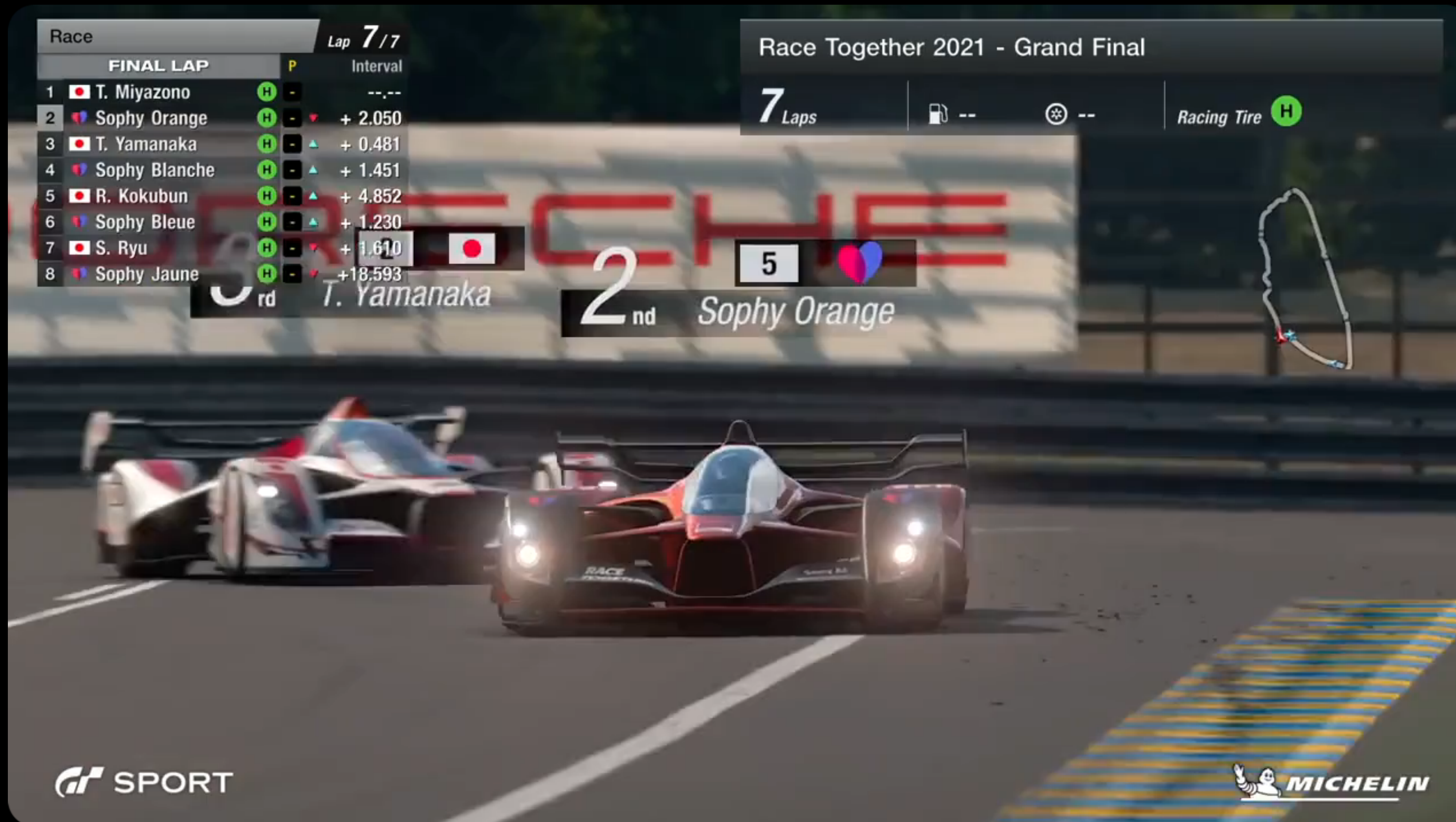
Goal: use **reward** to develop a **policy** to **maximise long-term reward**

Reinforcement learning

Agents

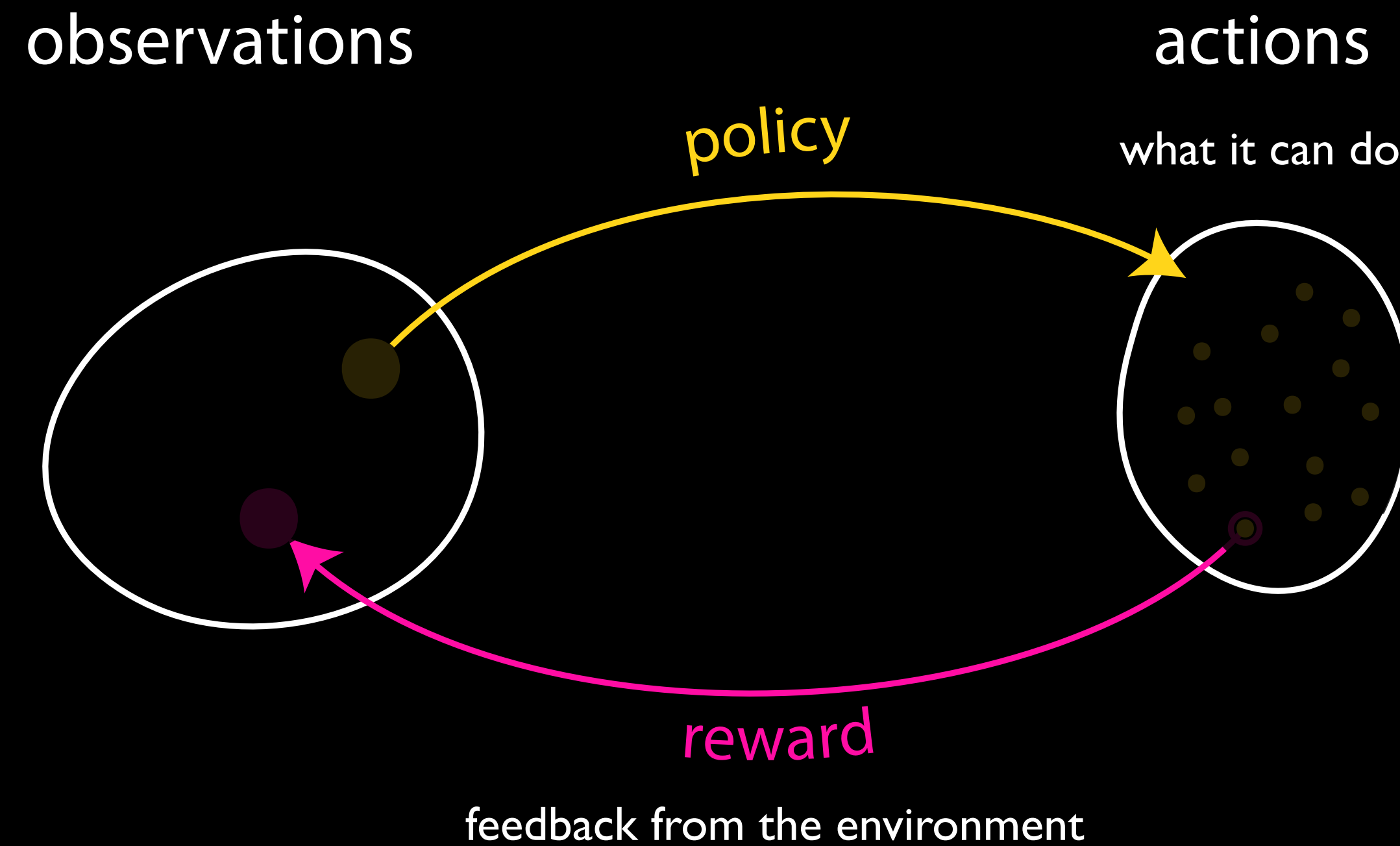


Reinforcement learning in Gran Turismo



Wurman et al., 2022, Outracing champion Gran Turismo drivers with deep reinforcement learning, Nature

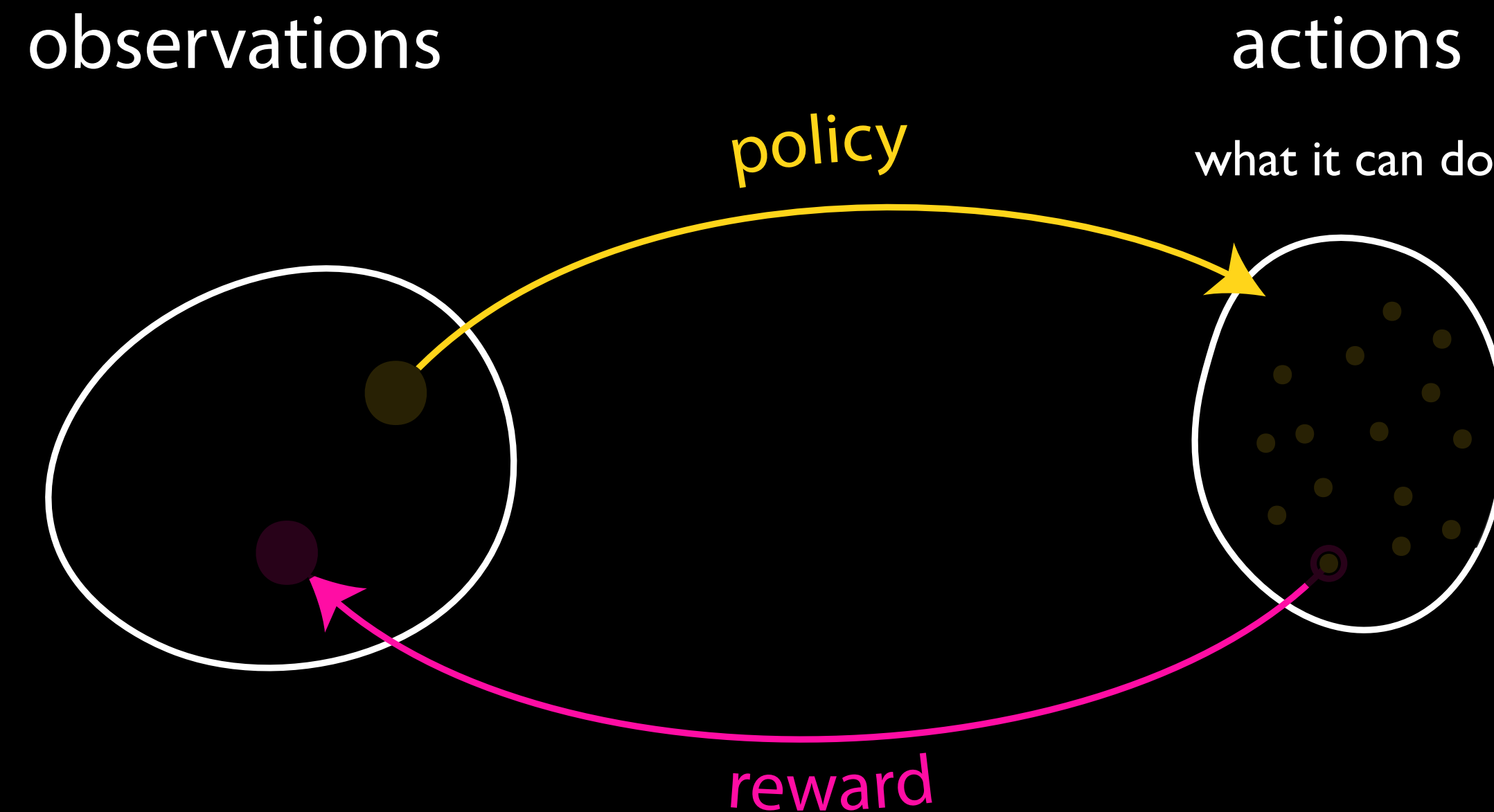
Reinforcement learning in Gran Turismo



$$r = \sum_i w_i r_i \quad \text{comprises multiple terms}$$

Reinforcement learning in Gran Turismo

Agents that can be interactive



feedback from the environment

$$r = \sum_i w_i r_i$$

comprises multiple terms
 w_i can be a range

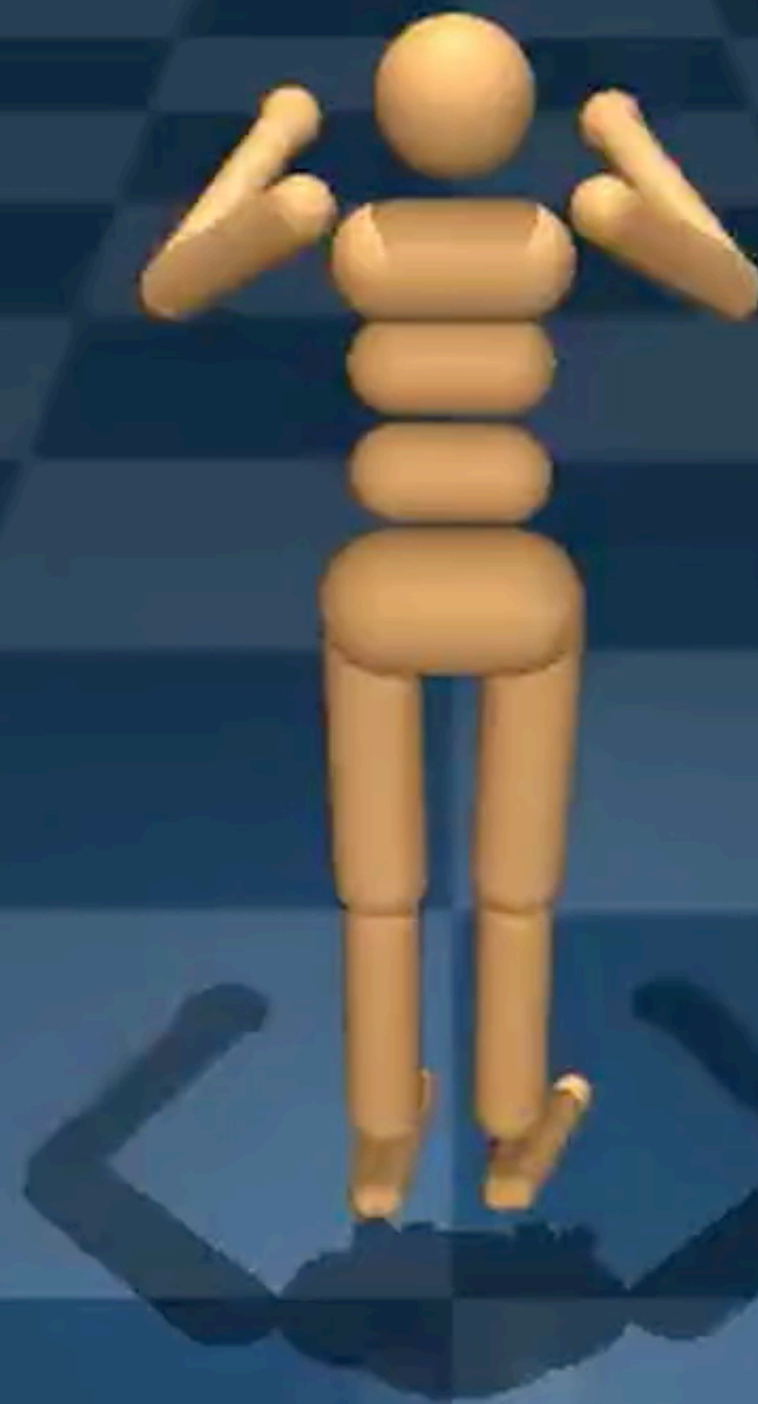
Reinforcement learning in Gran Turismo Coachable Agents for interactive gameplay



Capobianco et al., 2026, under review at TMLR

Coachable Agents: Demonstration on a humanoid

“Coaching” enables a different type of interaction with the agent than direct control.



Arm style

T-Pose

Arms up

Arms in front

Gait

Short

Medium

Long

Coachable Agents

and directions for flow control

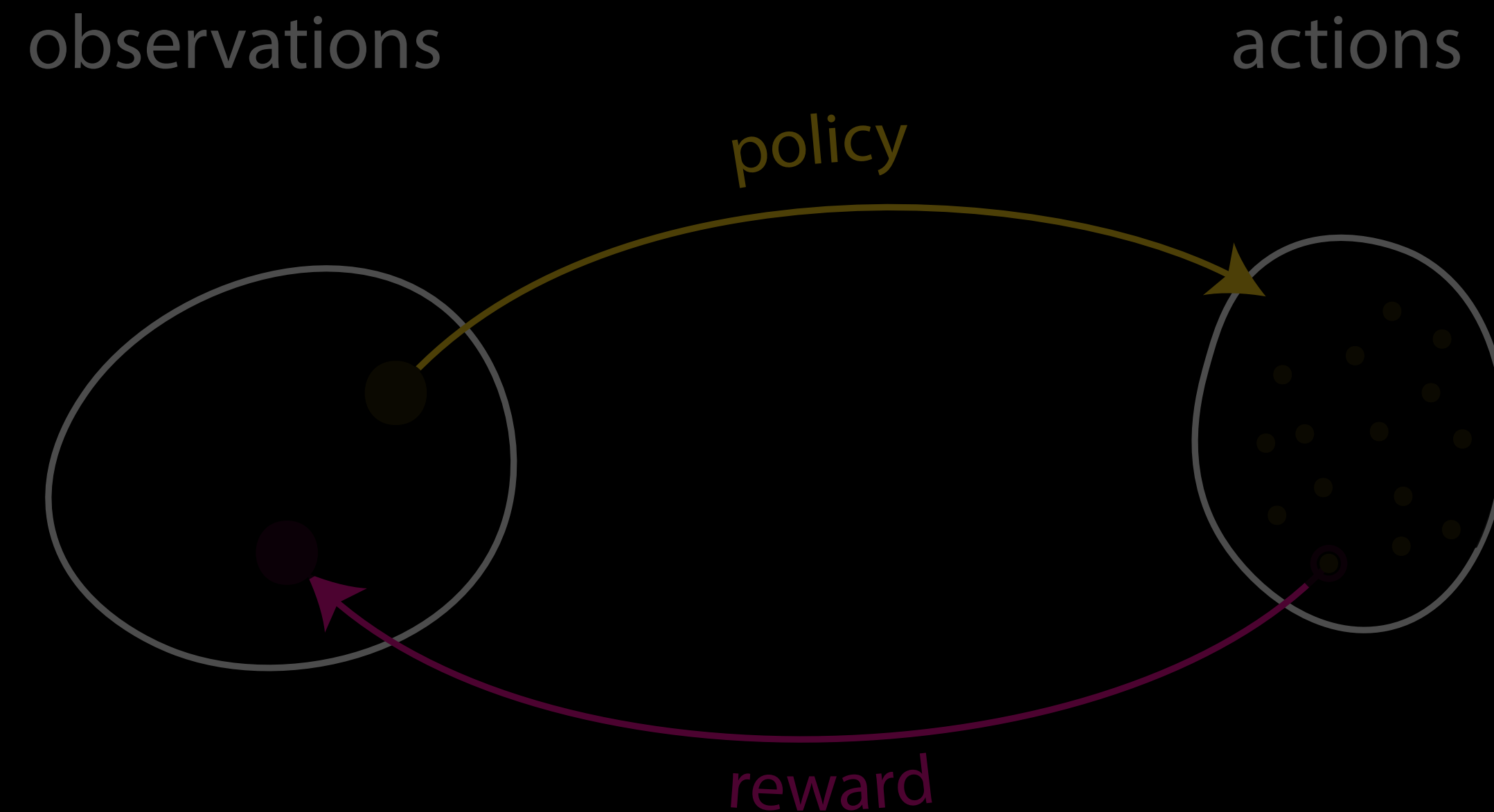
Coachable Agents

and directions for flow control

Frutos et al., 2025. Enhancing Energy Generation While Mitigating Noise Emissions in Wind Turbines Through Multi-Objective Optimization: A Deep Reinforcement Learning Approach, Wind Energy Journal

Coachable Agents

and directions for flow control

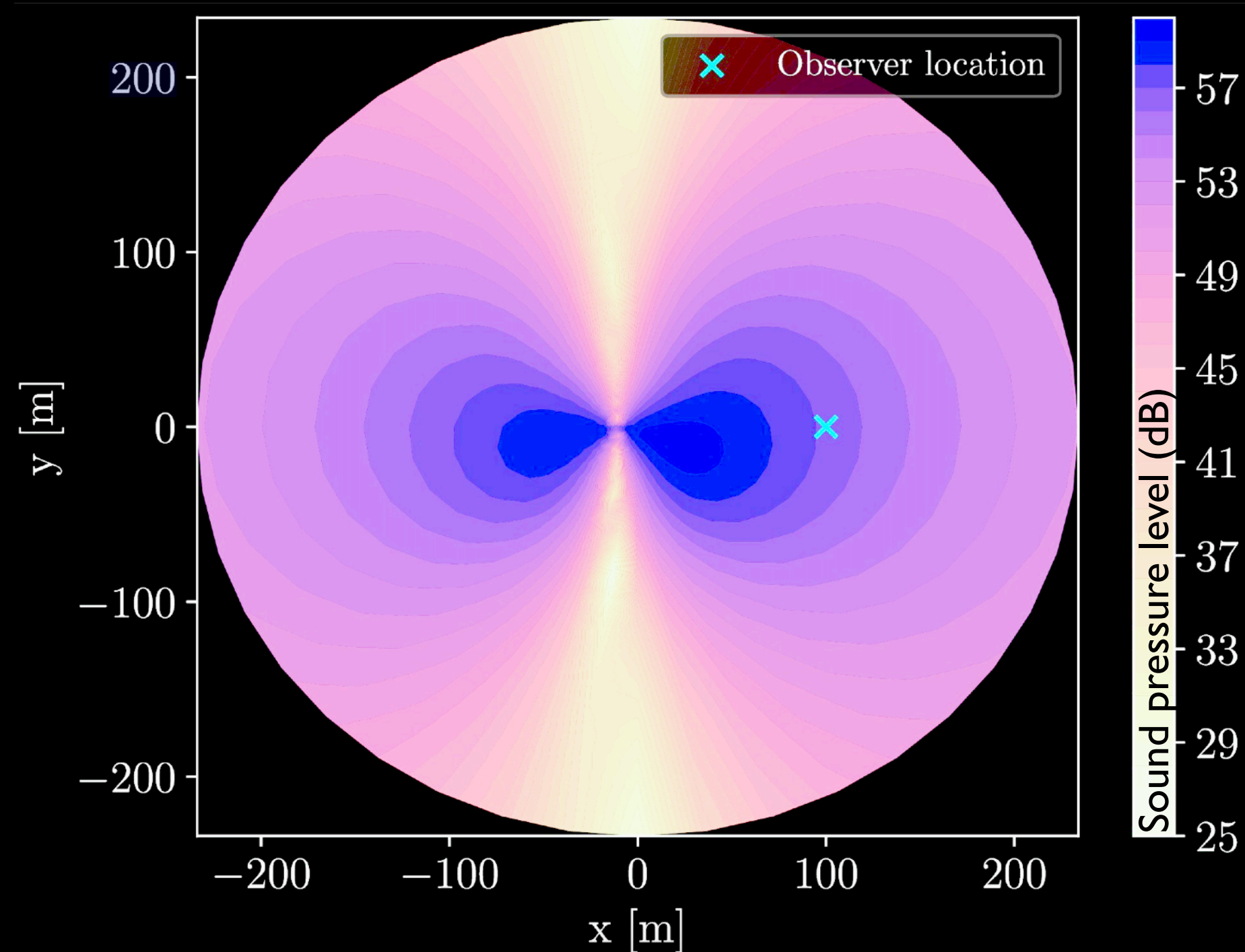


feedback from the environment

$$\text{reward} = w_{\text{power}} \text{reward}_{\text{power}} + w_{\text{noise}} \text{reward}_{\text{noise}}$$

Coachable Agents

and directions for flow control



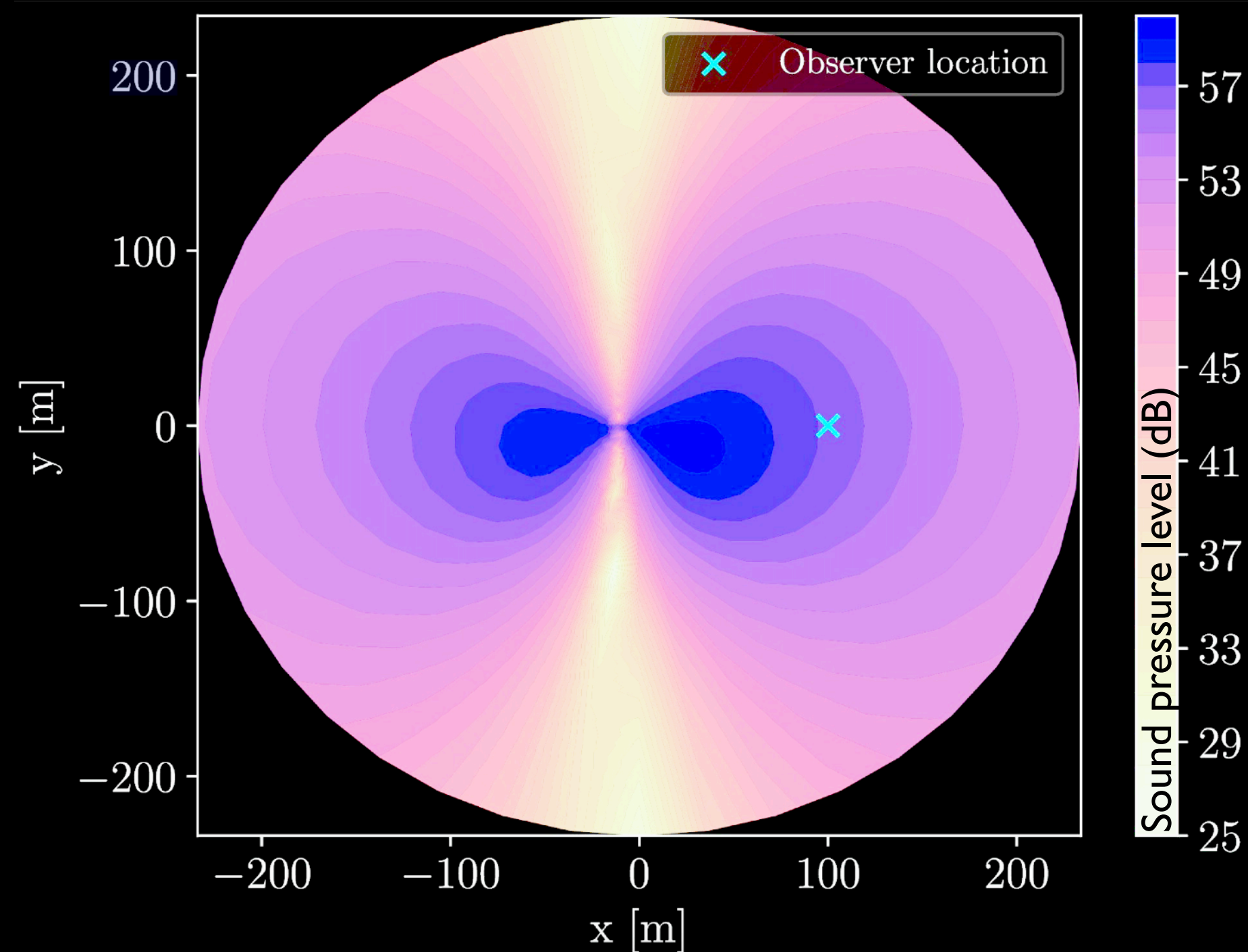
feedback from the environment

$$\text{reward} = w_{\text{power}} \text{reward}_{\text{power}} + w_{\text{noise}} \text{reward}_{\text{noise}}$$

Frutos et al., 2025. Enhancing Energy Generation While Mitigating Noise Emissions in Wind Turbines Through Multi-Objective Optimization: A Deep Reinforcement Learning Approach, Wind Energy Journal

Coachable Agents

and directions for flow control

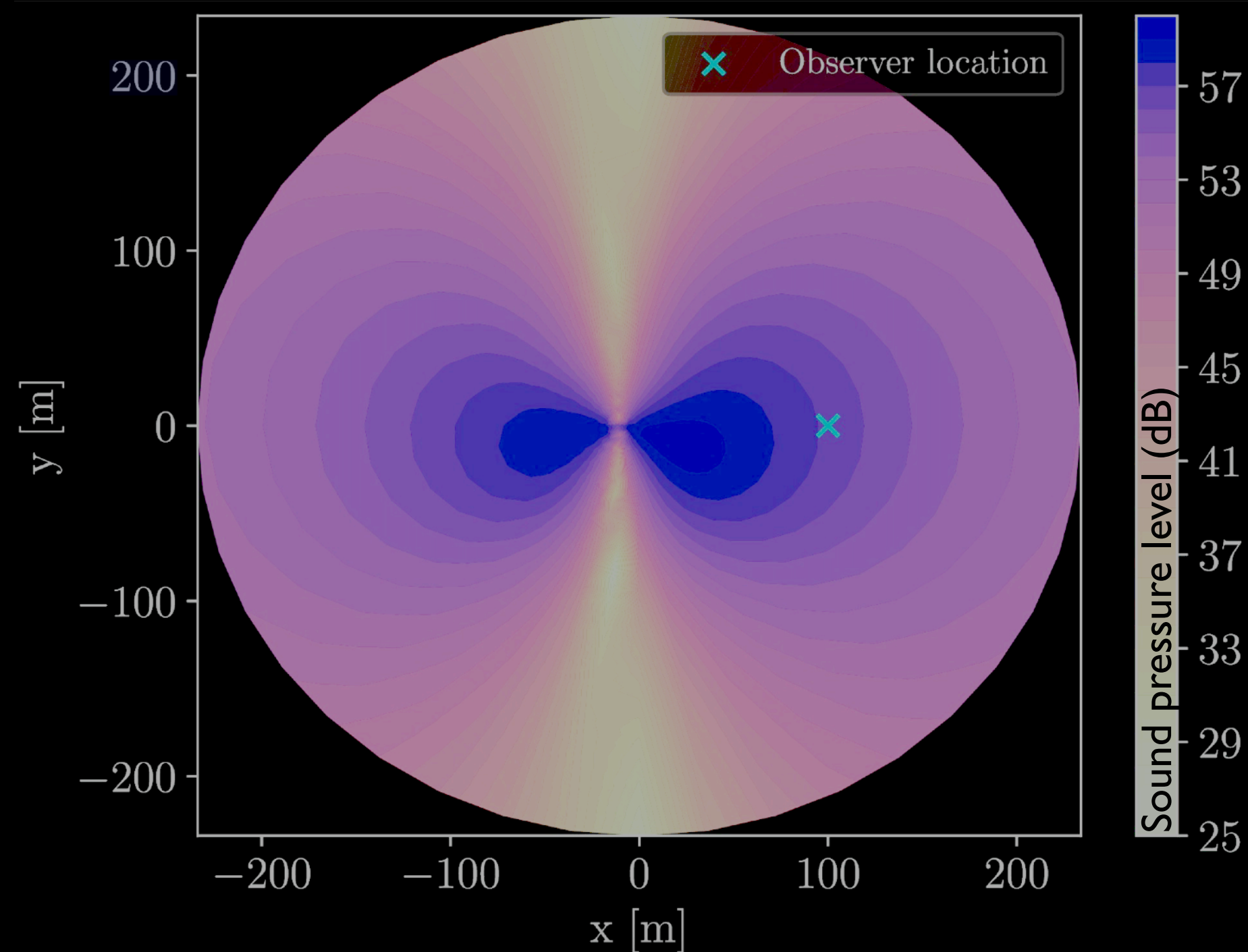


Adapting to a different observer location requires retraining.

$$\text{reward} = w_{\text{power}} \text{reward}_{\text{power}} + w_{\text{noise}} \text{reward}_{\text{noise}}$$

Coachable Agents

and directions for flow control



Adapting to a different observer location requires retraining.

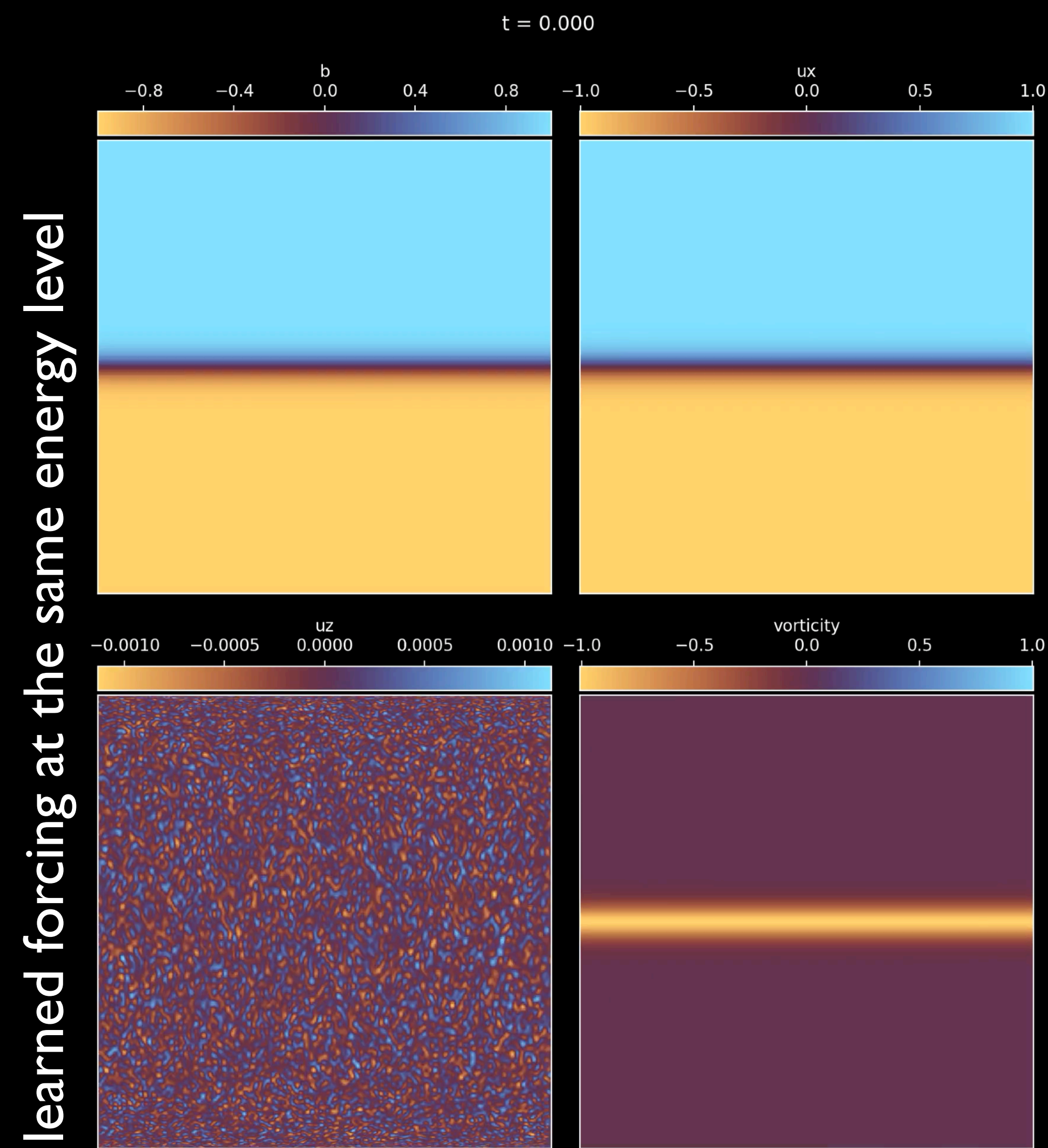
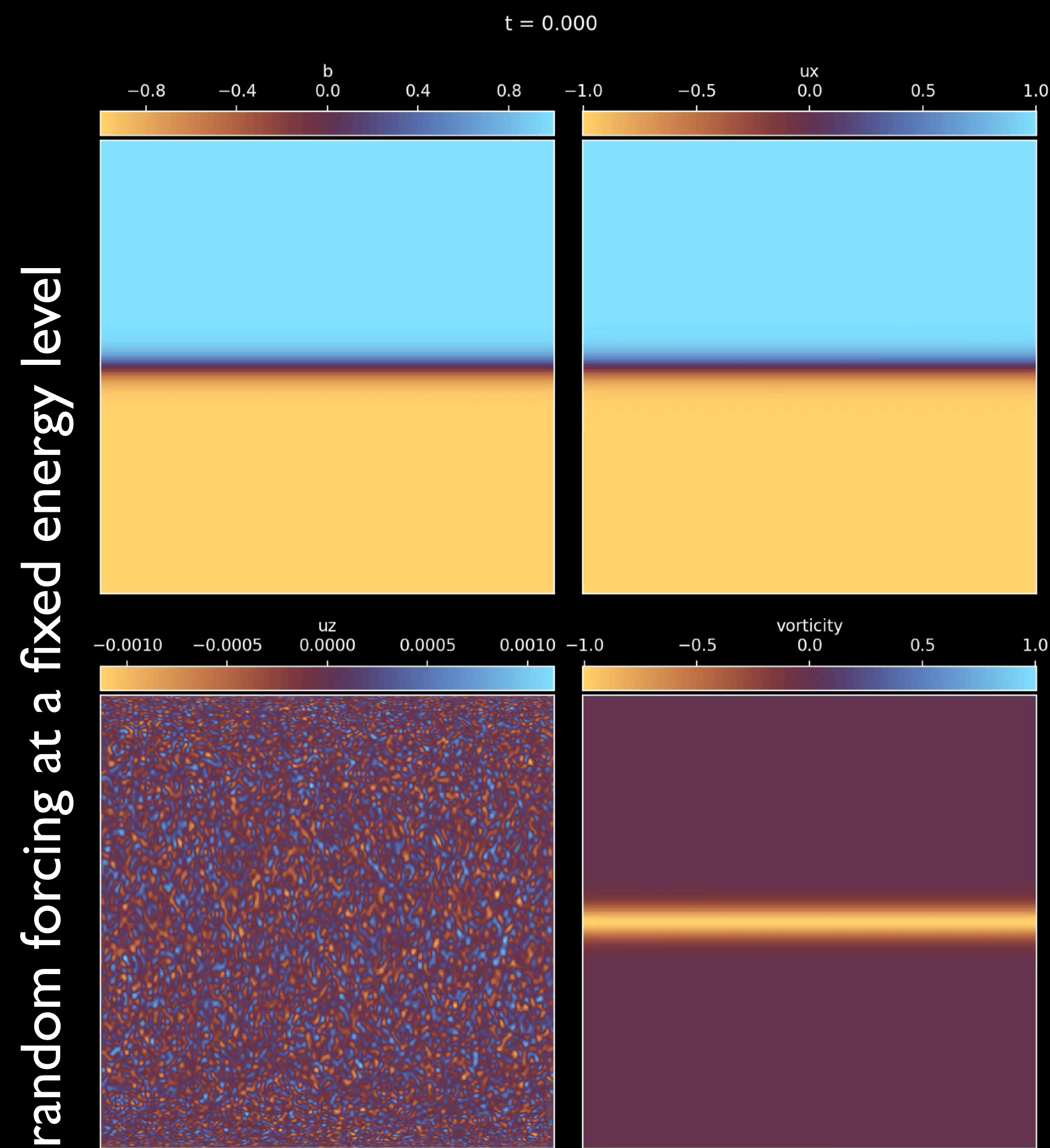
$$\text{reward} = w_{\text{power}} \text{reward}_{\text{power}} + w_{\text{noise}} \text{reward}_{\text{noise}}$$

Pretraining with reward weight UVFA can allow for real-time adaptation.
This may be extended to more flexible notions of “noise”.

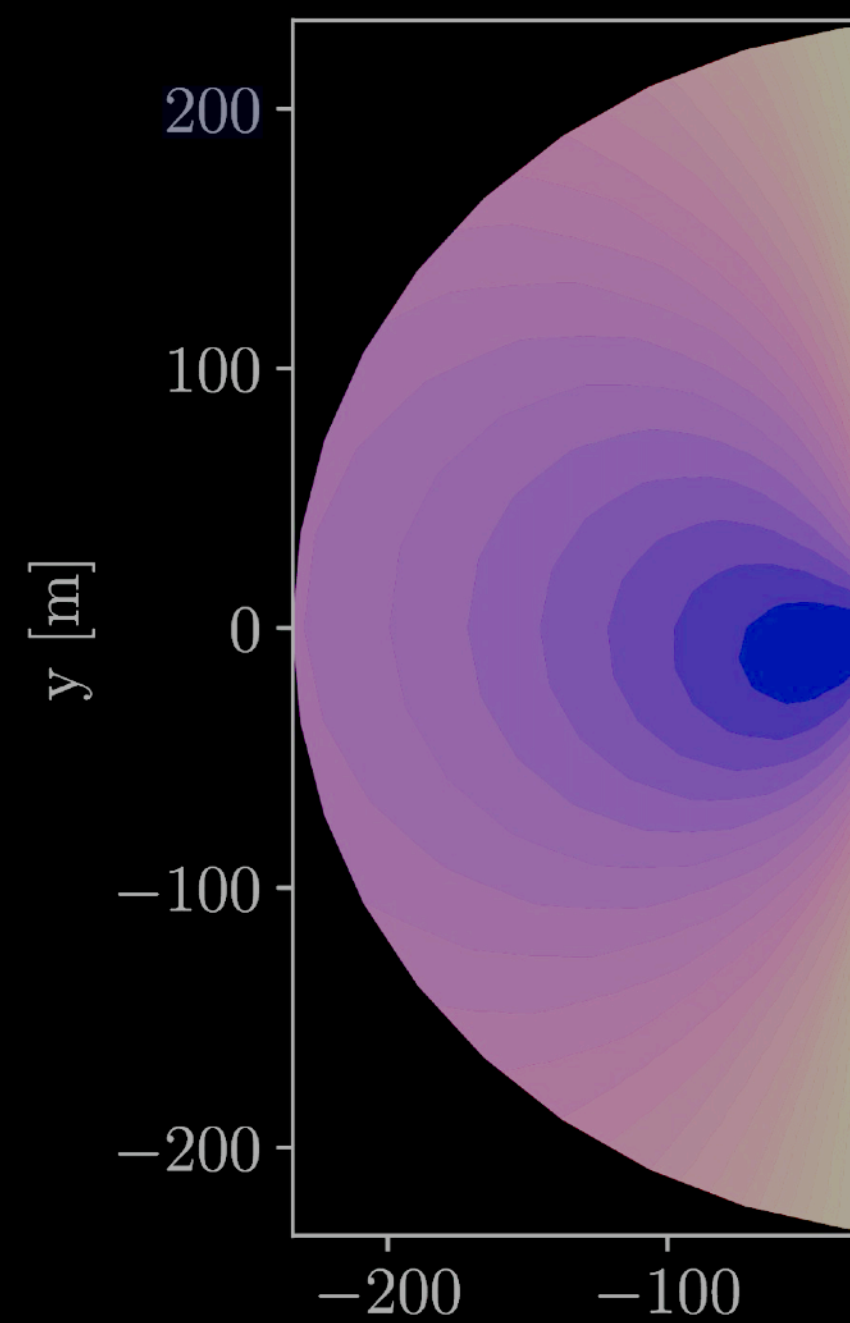
Frutos et al., 2025. Enhancing Energy Generation While Mitigating Noise Emissions in Wind Turbines Through Multi-Objective Optimization: A Deep Reinforcement Learning Approach, Wind Energy Journal

Separately,

Reinforcement learning for mixing in a flow



Coachable Agents



Arm style

T-Pose

Arms up

Arms in front



policy

reward

$$\text{reward} = w_{\text{power}} \text{reward}_{\text{power}} + w_{\text{noise}} \text{reward}_{\text{noise}}$$

