Imitation Learning as Inferring Latent Expert *Values*

Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

Setting

Solution

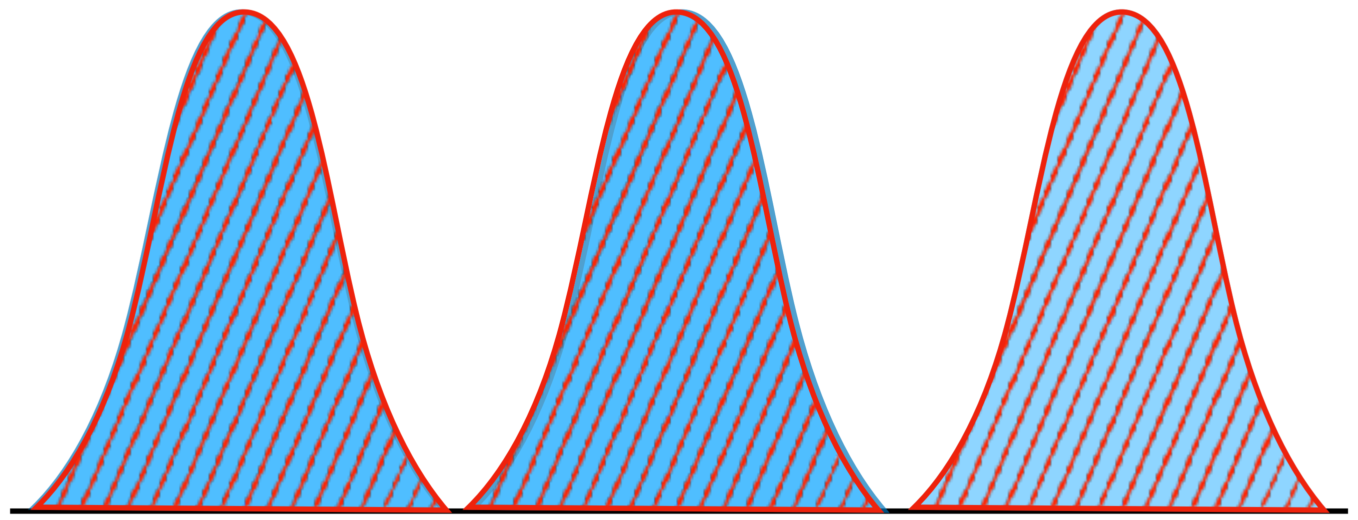
Easy

Expert is realizable

$\pi^E \in \Pi$

As $N \to \infty$, drive down

$\epsilon = 0$ (or Bayes error)



Nothing special.

Collect lots of data and

do Behavior Cloning

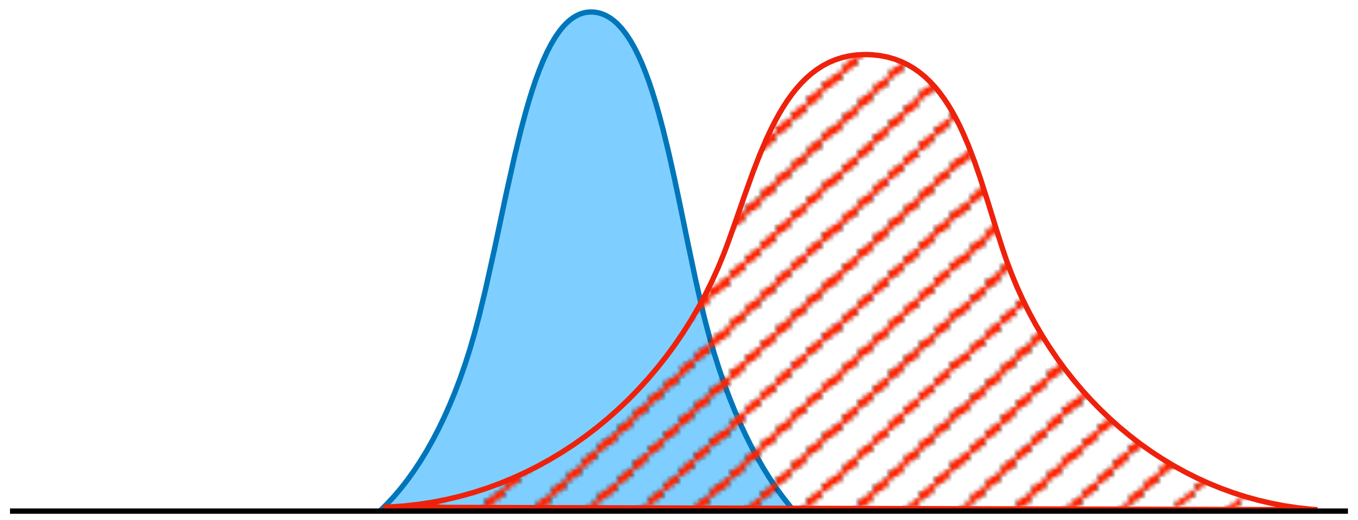
Hard

Non-realizable expert +

limited expert support

Even as $N \to \infty$,

behavior cloning $O(\epsilon T^2)$



Requires interactive expert

(DAGGER) to provide

labels $\Rightarrow O(\epsilon T)$

Two Core Ideas

Data

"What is the distribution of states?"

Loss

"What is the metric to match to human?"

Two Core Ideas

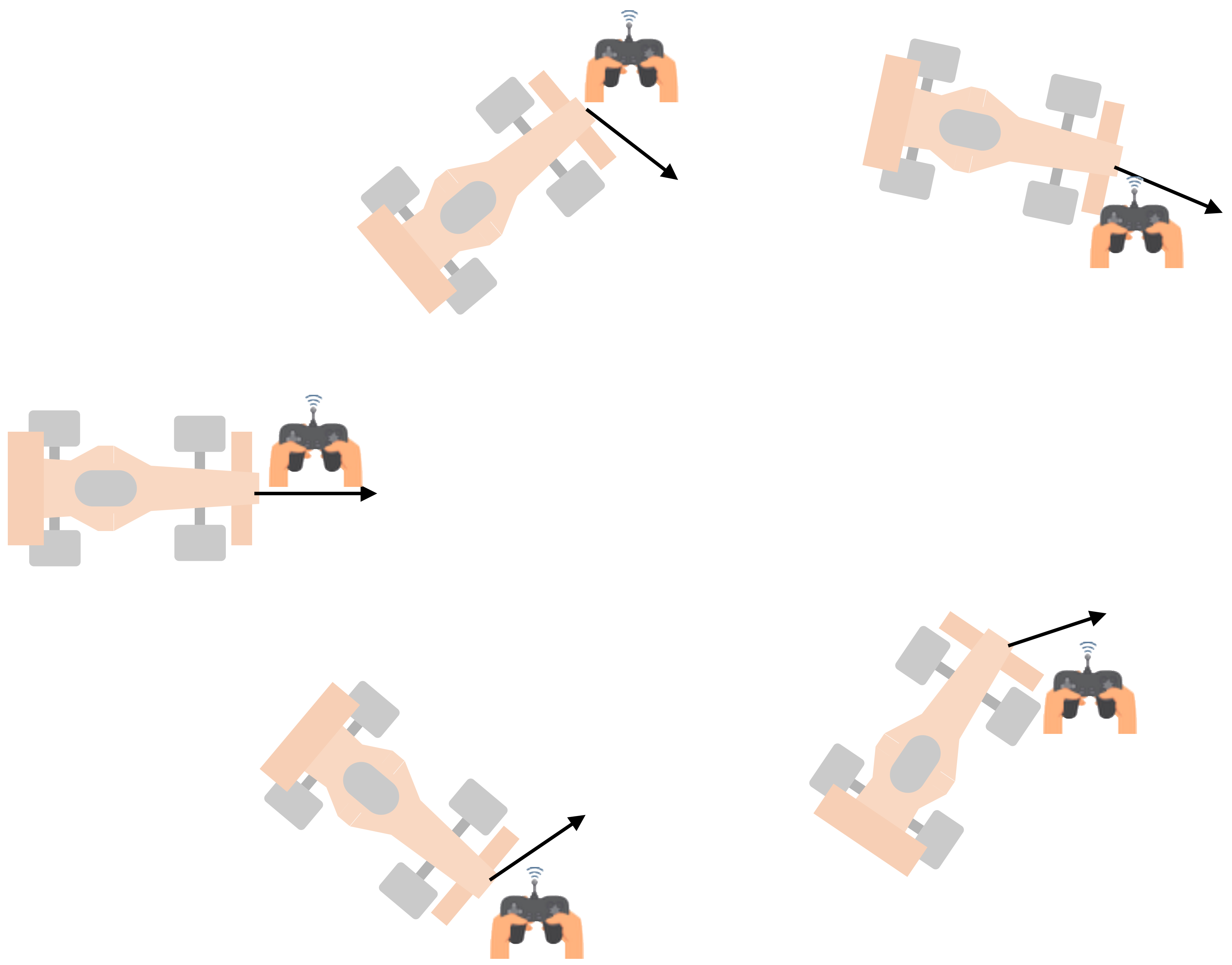
Data

"What is the distribution of states?"

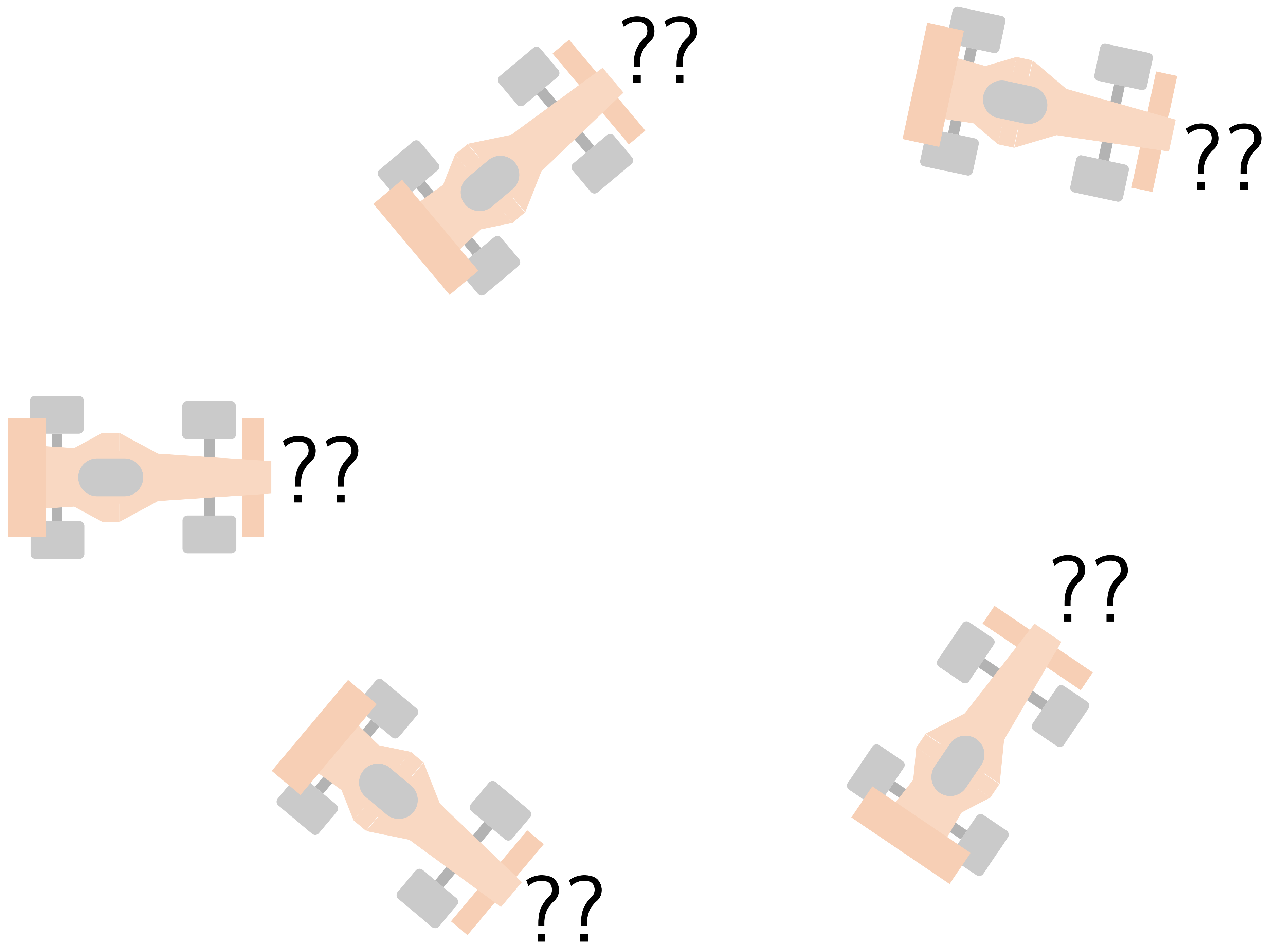
Loss

"What is the metric to match to human?"

DAGGER queries the human at every state!



Impractical: Too much human effort!



Can we learn from minimal human interaction?

Today's topic: Can we learn from minimal human feedback?

Think of the most minimal feedback:
An E-STOP!

How can we learn from this
1 bit feedback?

Recap: DAGGER

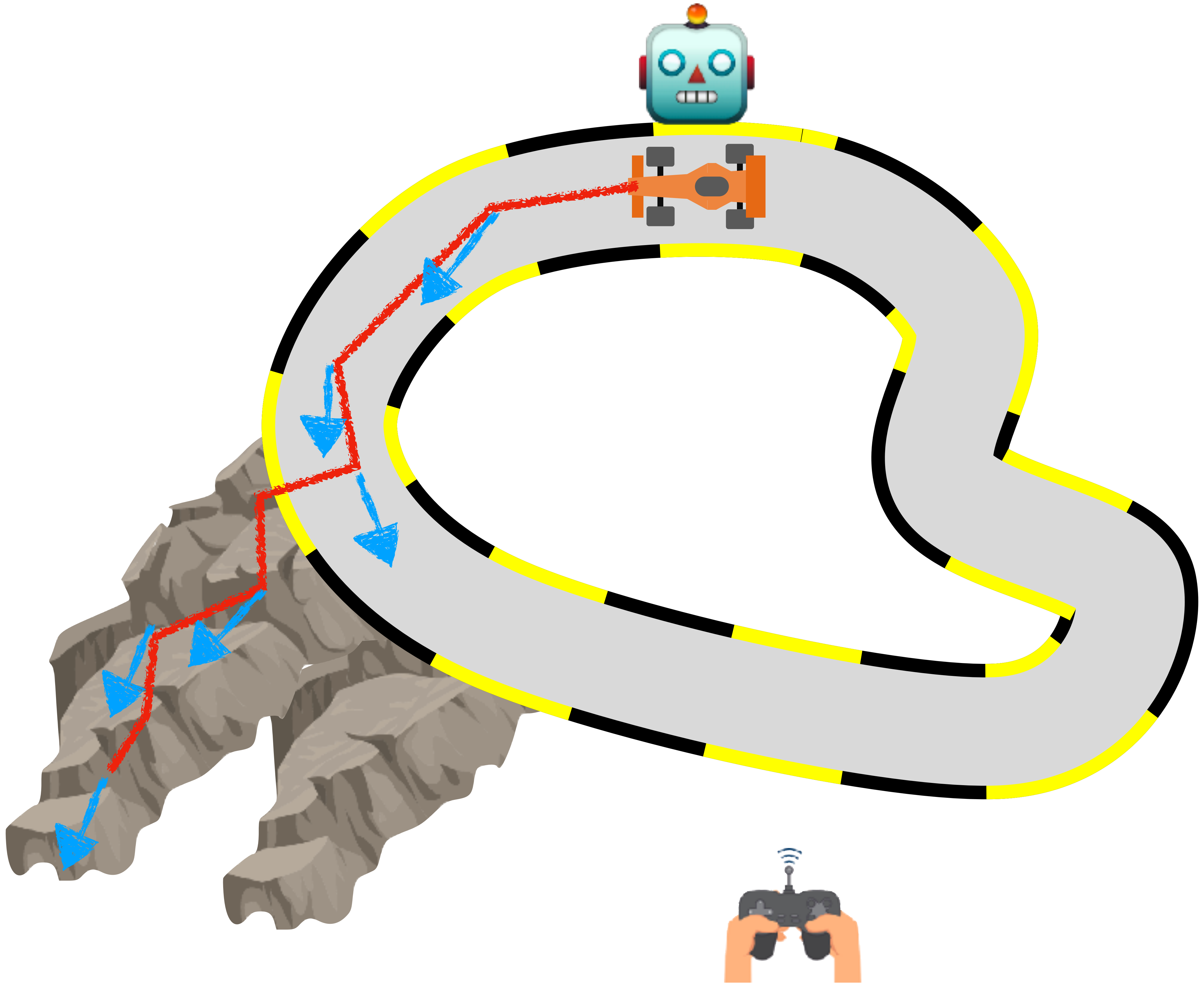
Roll out a learner policy

Collect expert actions

Aggregate data

Update policy

$$\min_{\pi} \mathbb{E}_{s,a^* \sim \mathcal{D}} 1(\pi(s) \neq a^*)$$



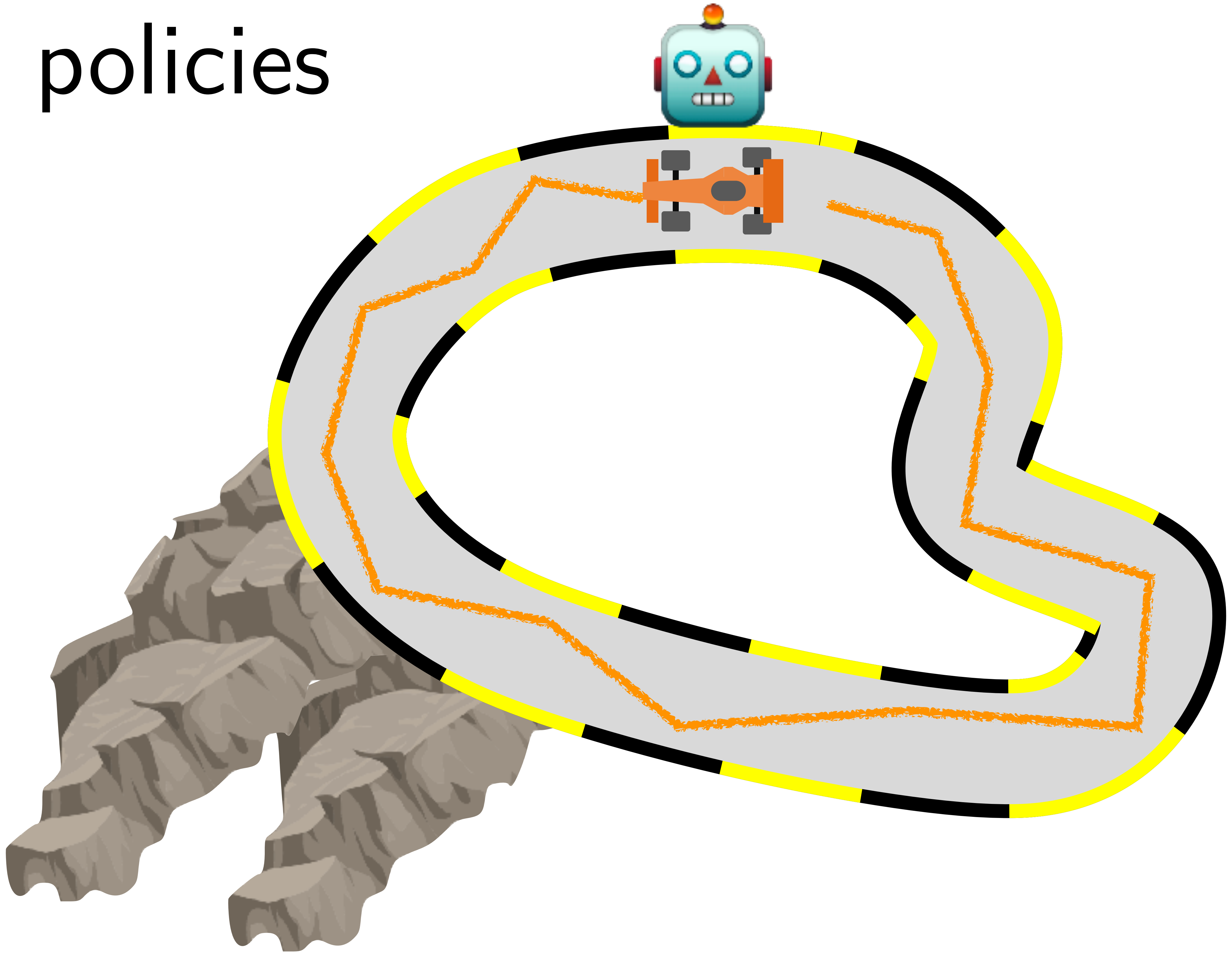
Not all errors are equal

What does DAGGER guarantee?

Let's say your policy class Π has 2 policies

Policy π_1:

Shaky hands,
never goes out of racetrack,
but can't recover if it did

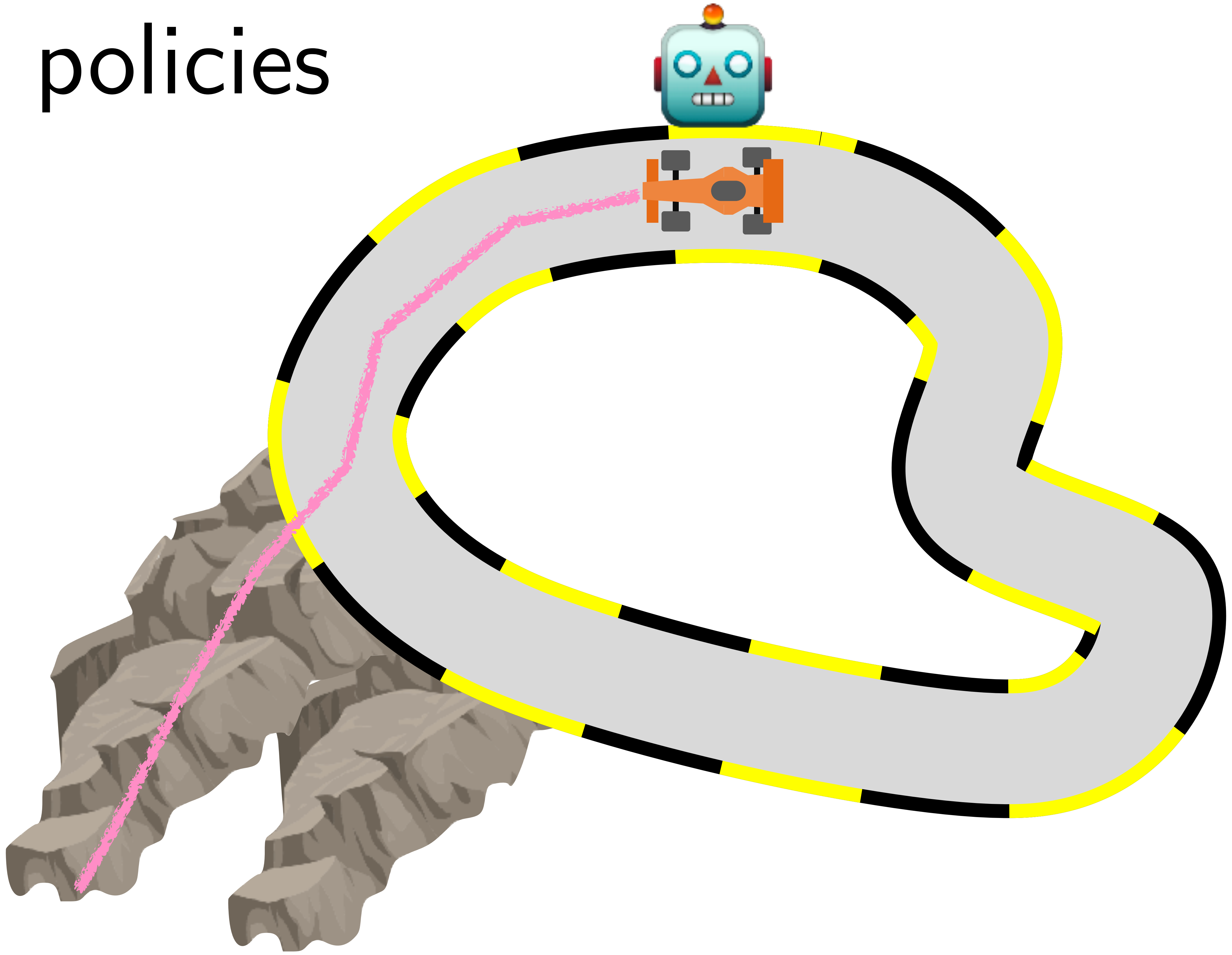


What does DAGGER guarantee?

Let's say your policy class Π has 2 policies

Policy π_2:

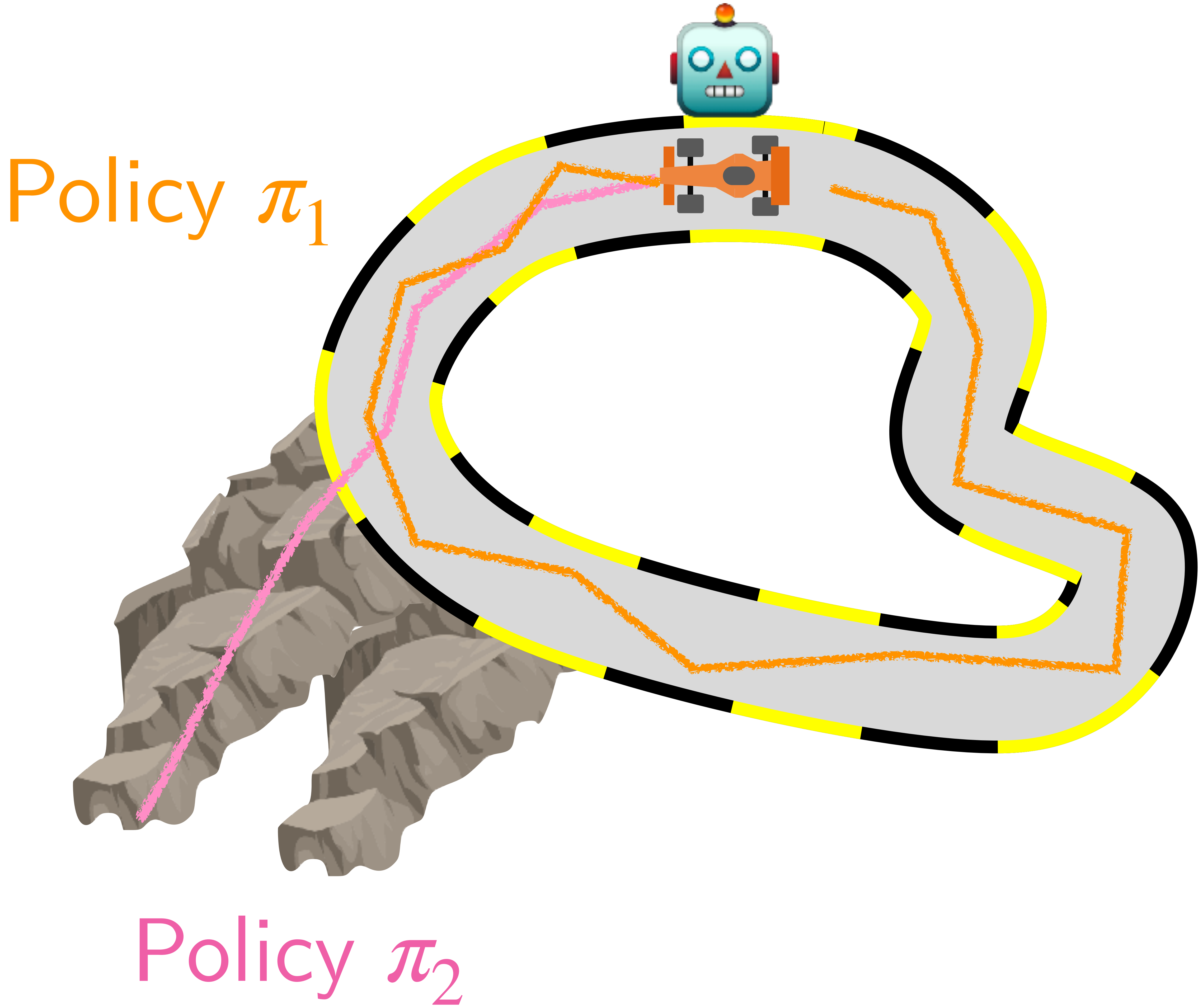
Perfect on straight turns,
Perfect when falling off the cliff,
But makes mistake on the curve



What does DAGGER guarantee?

Which policy would you like to learn?

Which policy might DAGGER return?



Activity!

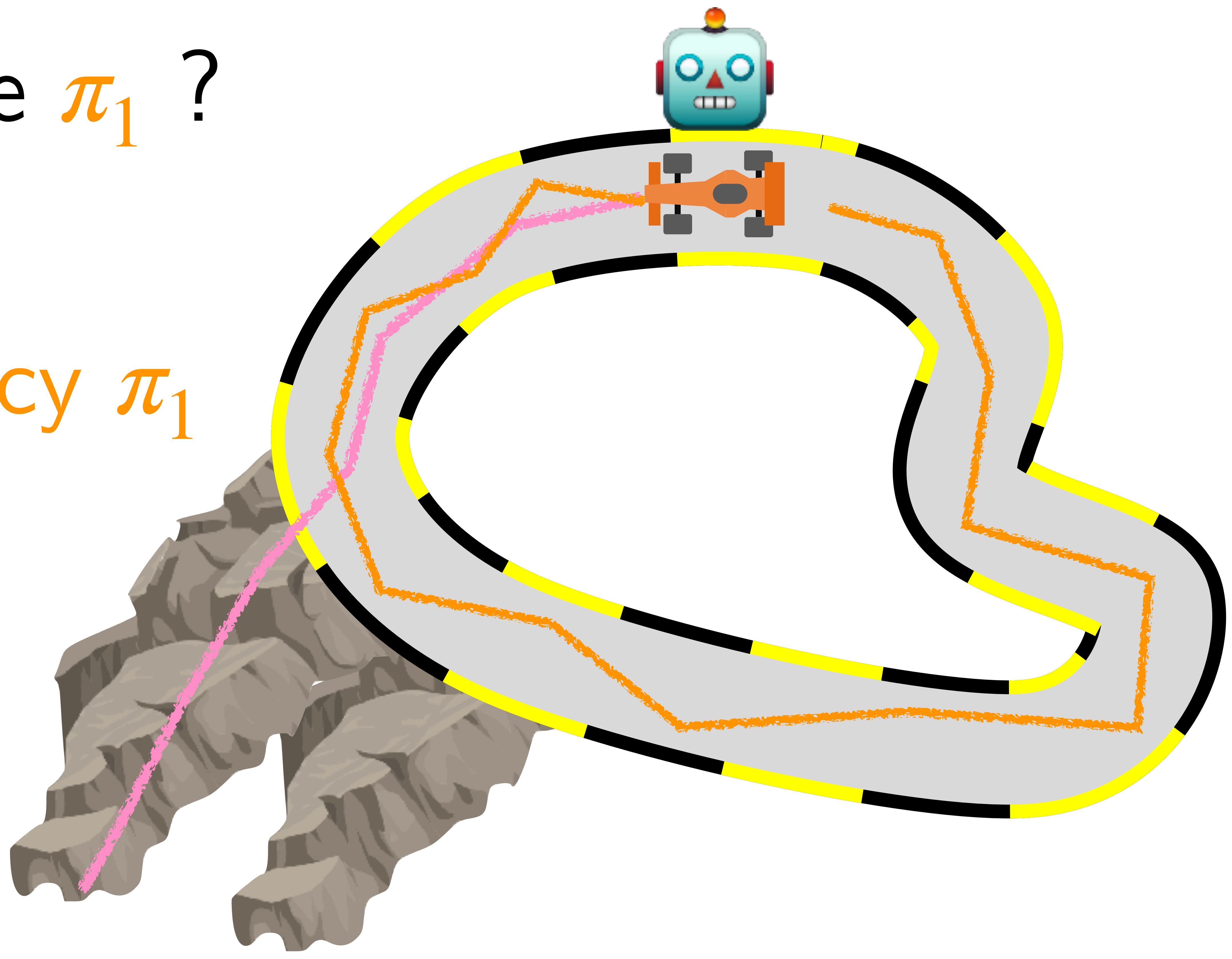
Think-Pair-Share!

Think (30 sec): Which policy would DAGGER return? How would you get it to choose π_1 ?

Pair: Find a partner

Share (45 sec): Partners exchange ideas

Policy π_1



Policy π_2

What is
theoretically the best
we can do in
imitation learning?

Performance Difference Lemma

Is there a theoretically best imitation learning algorithm?

AGGREVATE

Reinforcement and Imitation Learning via Interactive No-Regret Learning

Stéphane Ross **J. Andrew Bagnell**
stephaneross@cmu.edu dbagnell@ri.cmu.edu
The Robotics Institute
Carnegie Mellon University,
Pittsburgh, PA, USA

AGGREVATE(D)

Deeply AggreVaTeD:
Differentiable Imitation Learning for Sequential Prediction

Wen Sun† WENSUN@CS.CMU.EDU
Arun Venkatraman† ARUNVENK@CS.CMU.EDU
Geoffrey J. Gordon† GGORDON@CS.CMU.EDU
Byron Boots* BBOOTS@CC.GATECH.EDU
J. Andrew Bagnell† DBAGNELL@RI.CMU.EDU
†School of Computer Science, Carnegie Mellon University, USA
*College of Computing, Georgia Institute of Technology, USA

AGGREVATE: Expert provides values

Just like DAGGER

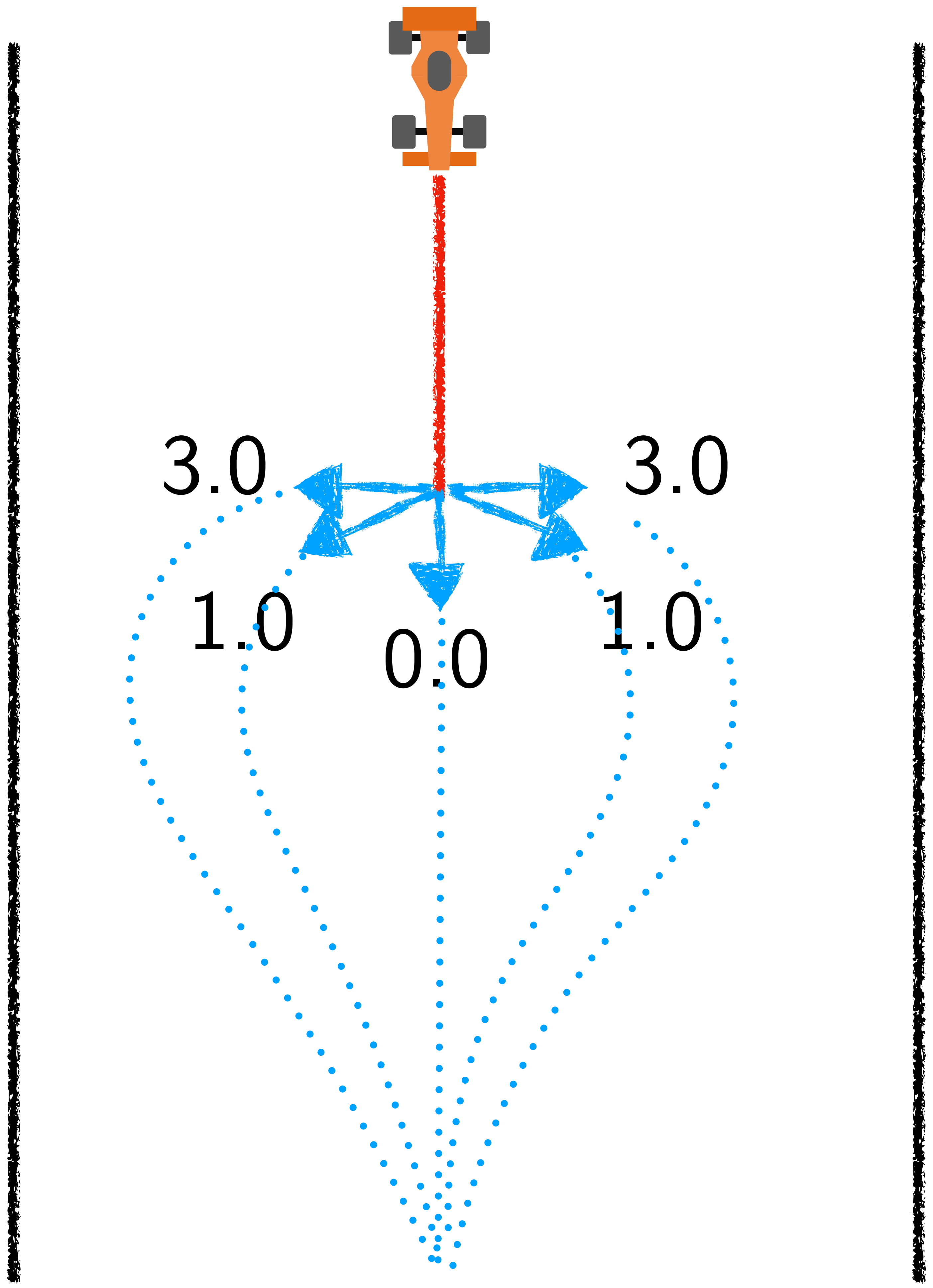
For i = 0 … N-1

Roll-in learner π_i to get $\{s \sim d_{\pi_i}\}$

Query expert for advantage vector $A^*(s, .)$

Aggregate data $\mathcal{D} \leftarrow \mathcal{D} \cup \{s, A^*(s, .)\}$

Train policy $\pi_{i+1} = \mathbb{E}_{s, A^* \sim \mathcal{D}}(A^*(s, \pi(s)))$



AGGREVATE: Expert provides values

Just like DAGGER

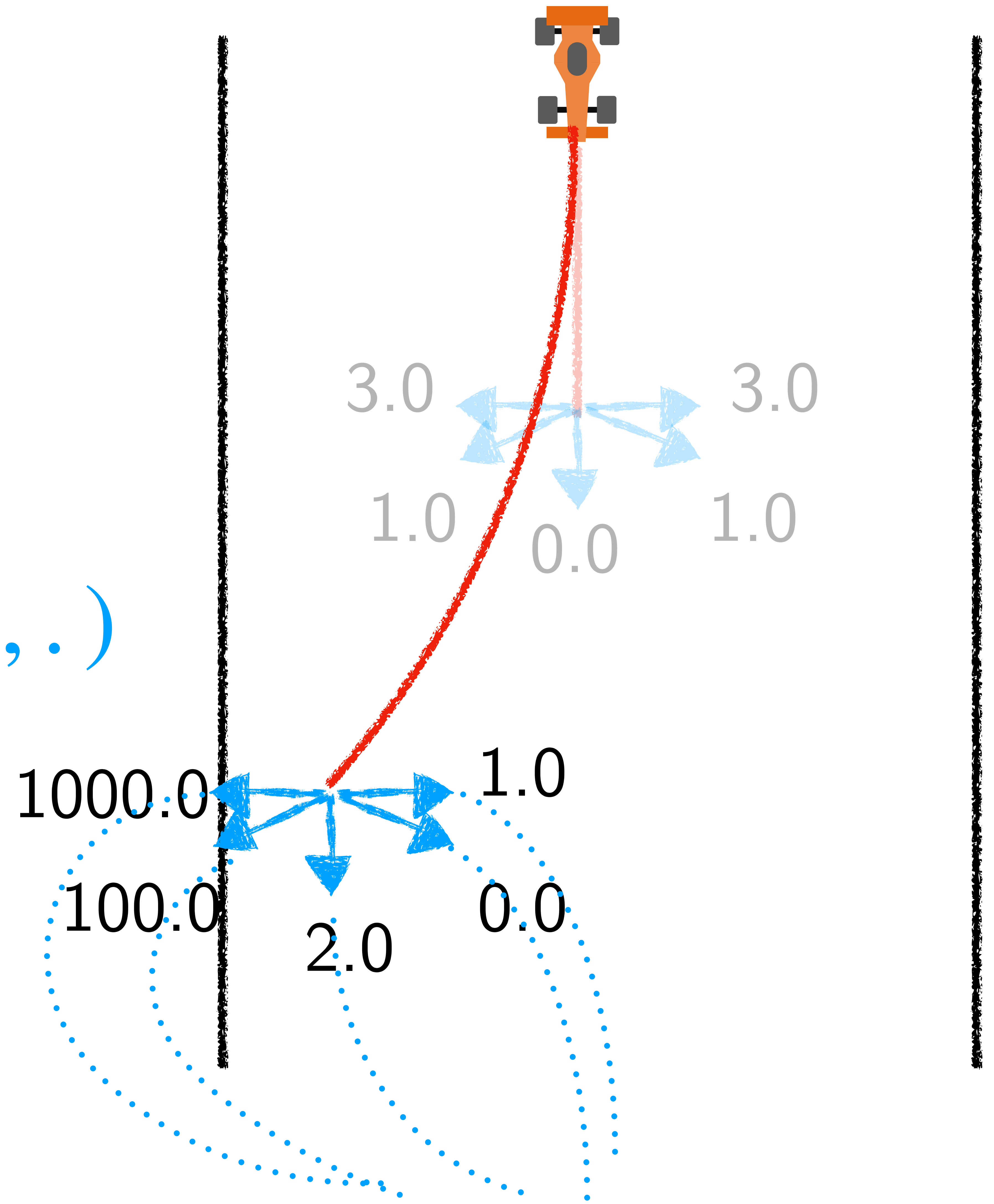
For i = 0 … N-1

Roll-in learner π_i to get $\{s \sim d_{\pi_i}\}$

Query expert for advantage vector $A^*(s, .)$

Aggregate data $\mathscr{D} \leftarrow \mathscr{D} \cup \{s, A^*(s, .)\}$

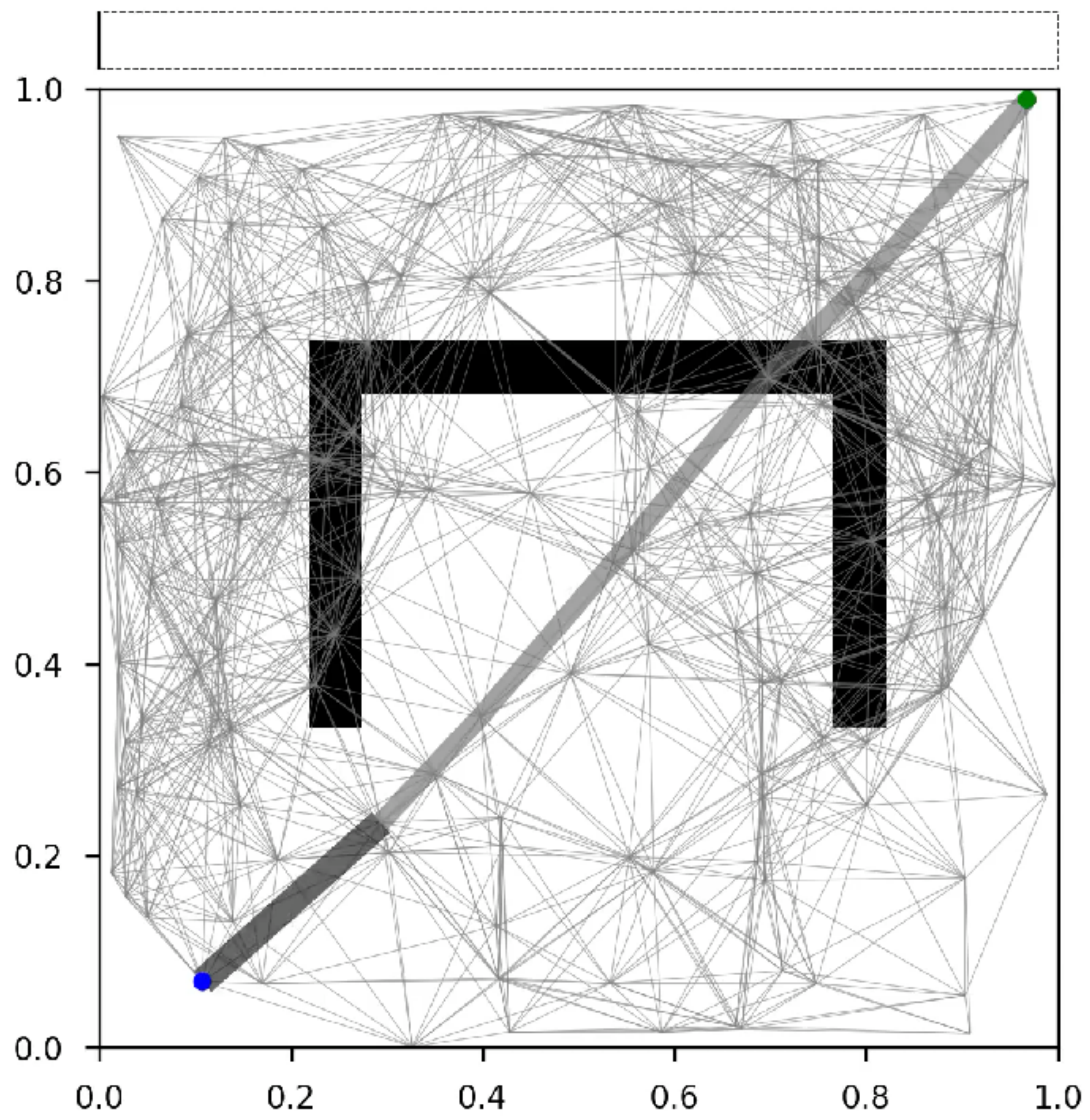
Train policy $\pi_{i+1} = \mathbb{E}_{s,A^* \sim \mathscr{D}}(A^*(s, \pi(s)))$

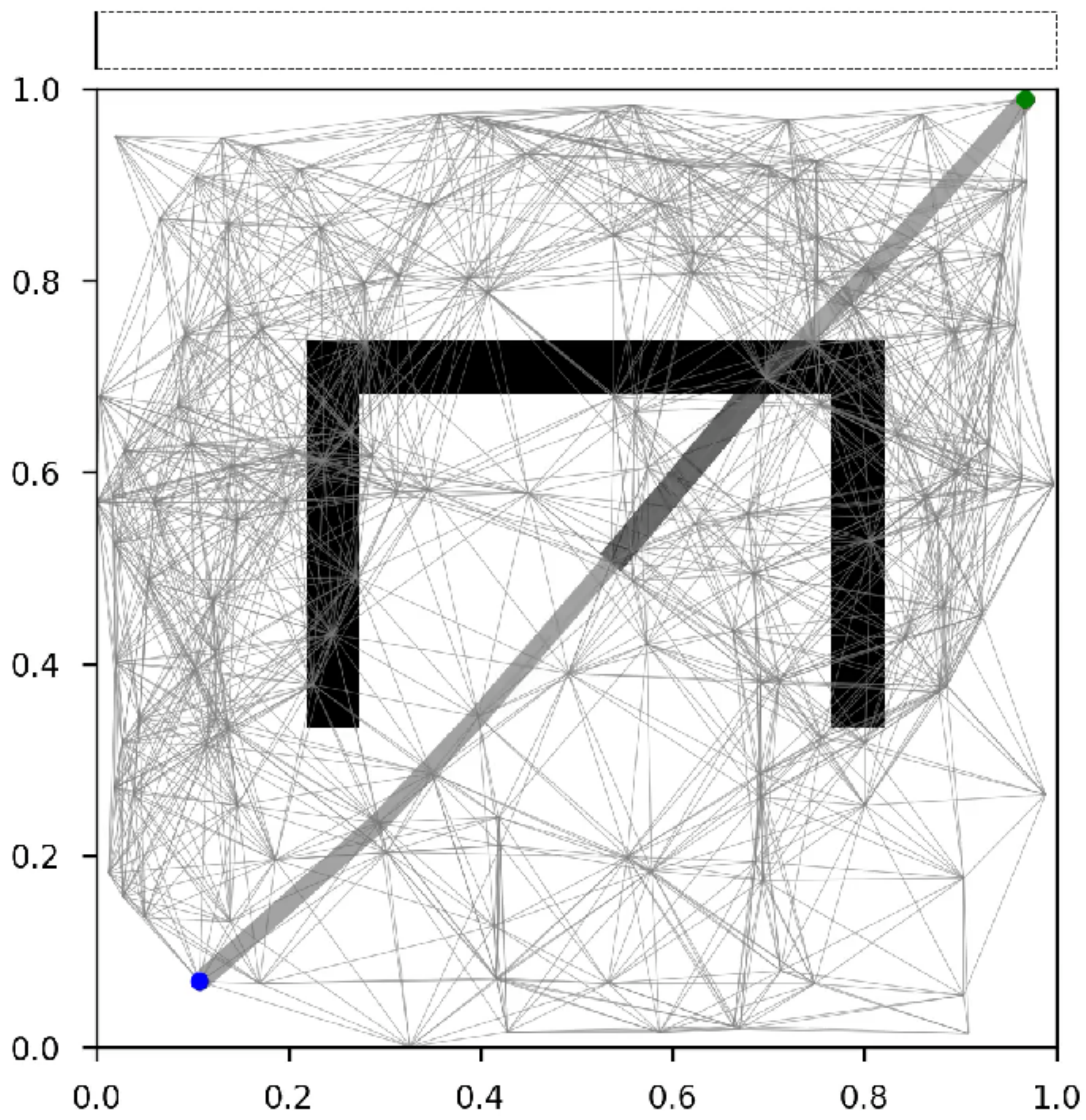


Is Aggrevate even practical?

Yes! AGGREVATE useful for imitating oracles

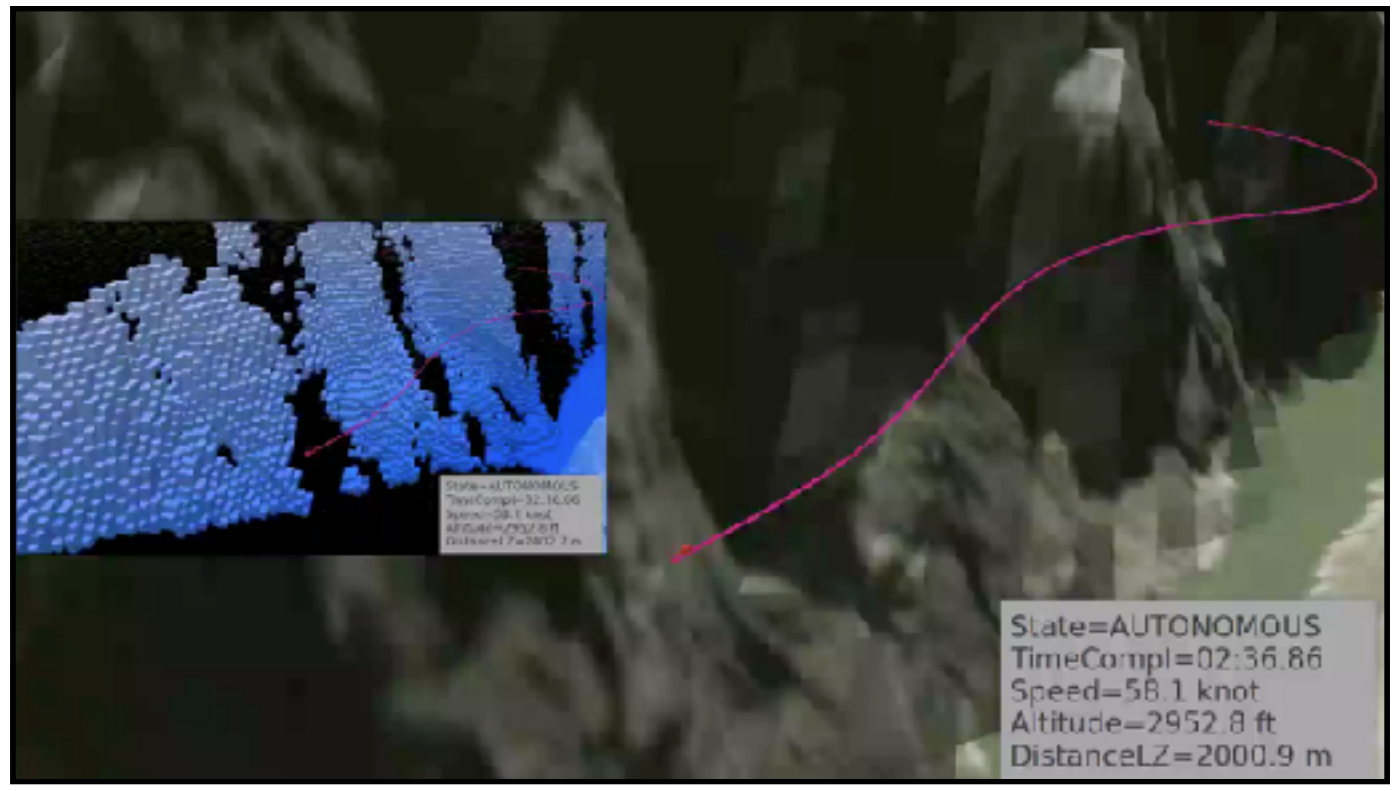
Train search heuristics by imitating oracular planners



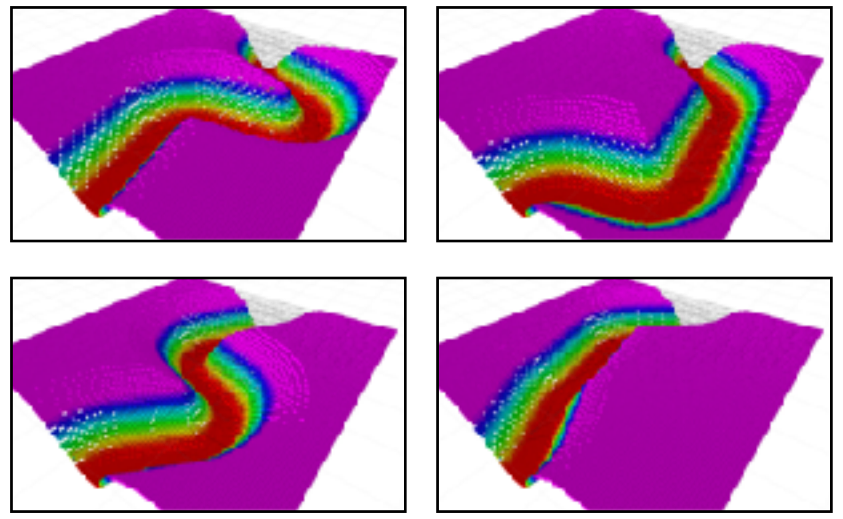


Choudhury, S., Bhardwaj, M., Arora, S., Kapoor, A., Ranade, G., Scherer, S., & Dey, D. Data-driven planning via imitation learning. *IJRR'18*

AGGREVATE for helicopter planning

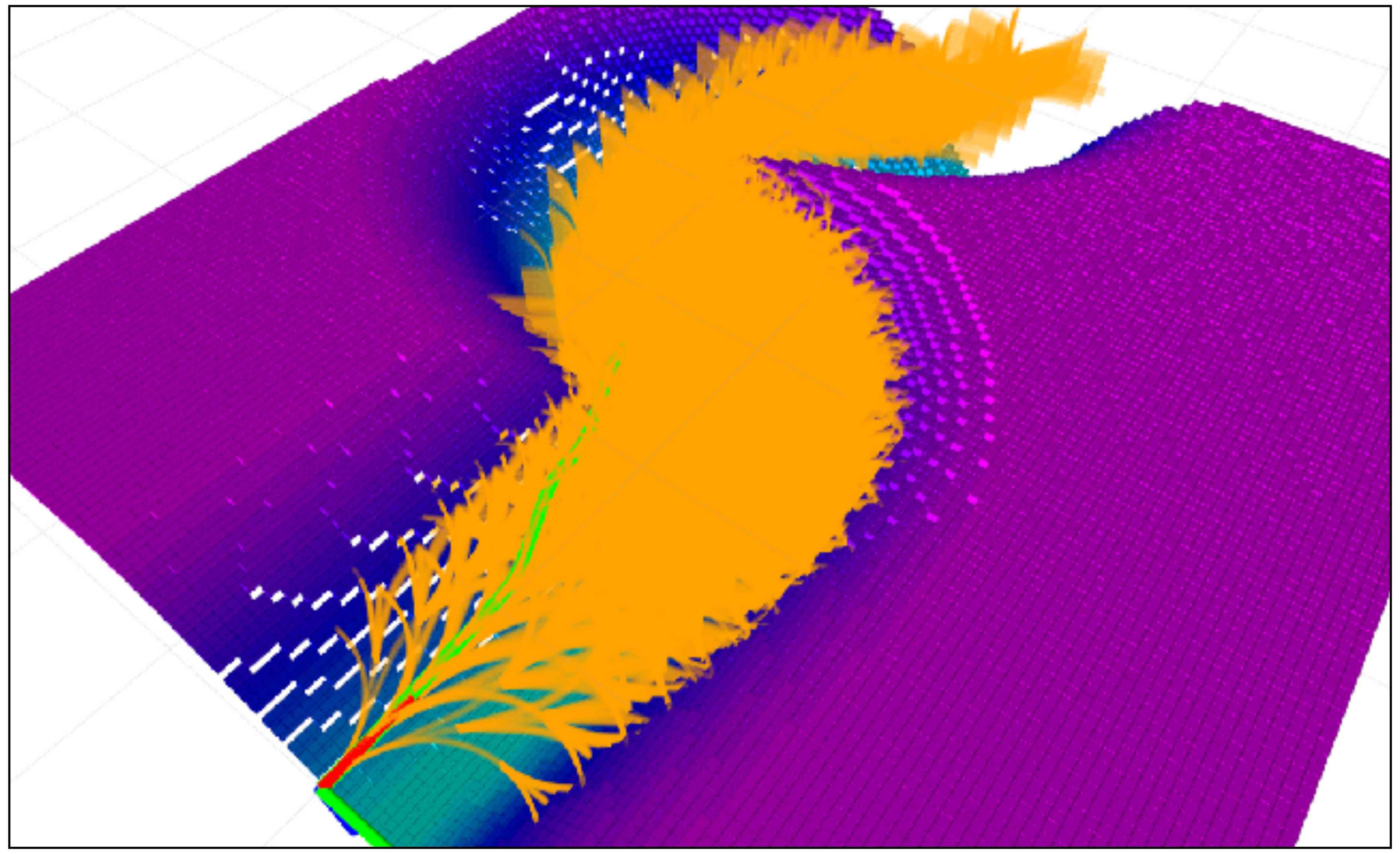


An autonomous helicopter navigating in a canyon

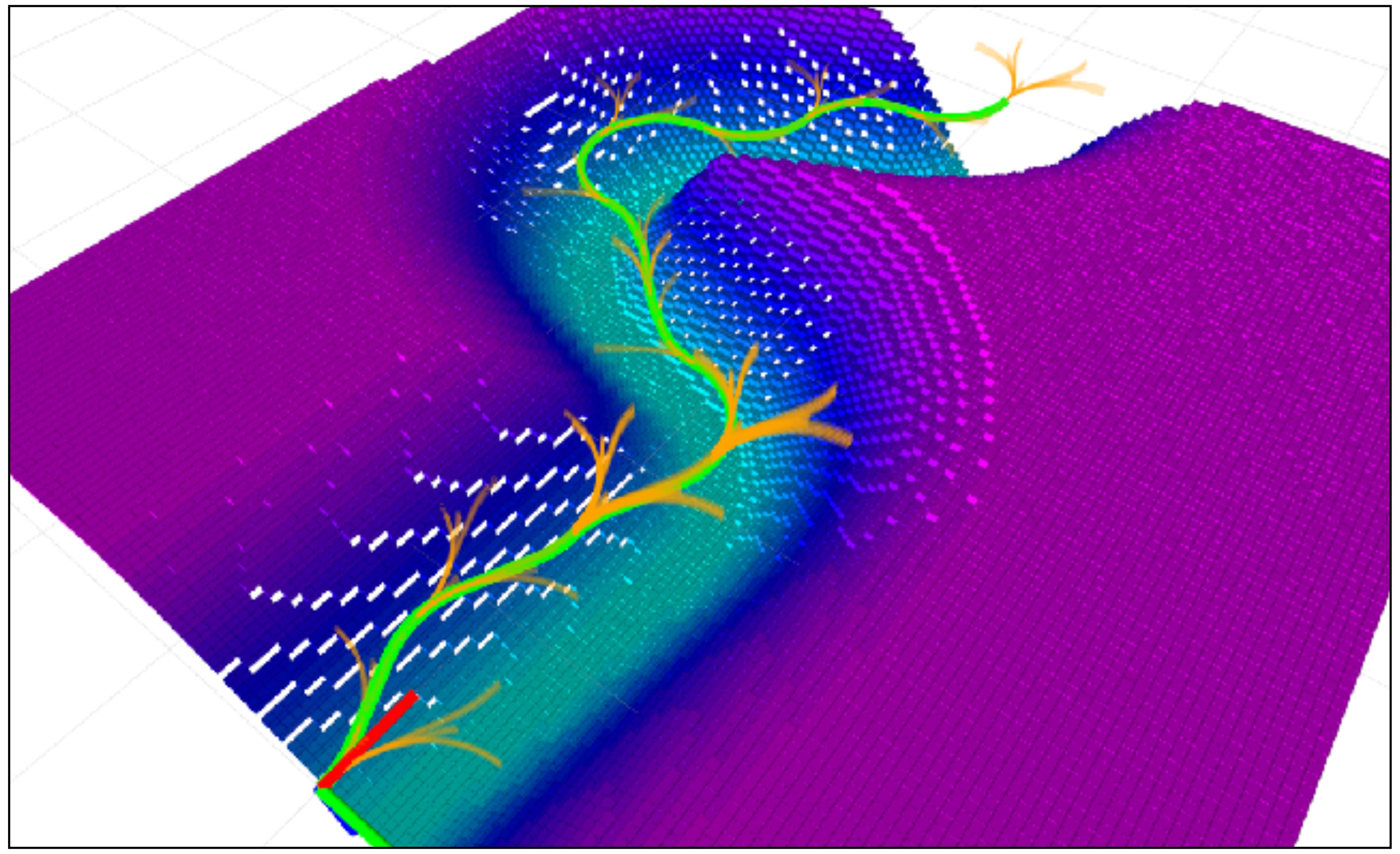


Dataset of canyons

Learning a heuristic for 4D search (x,y,z,heading)



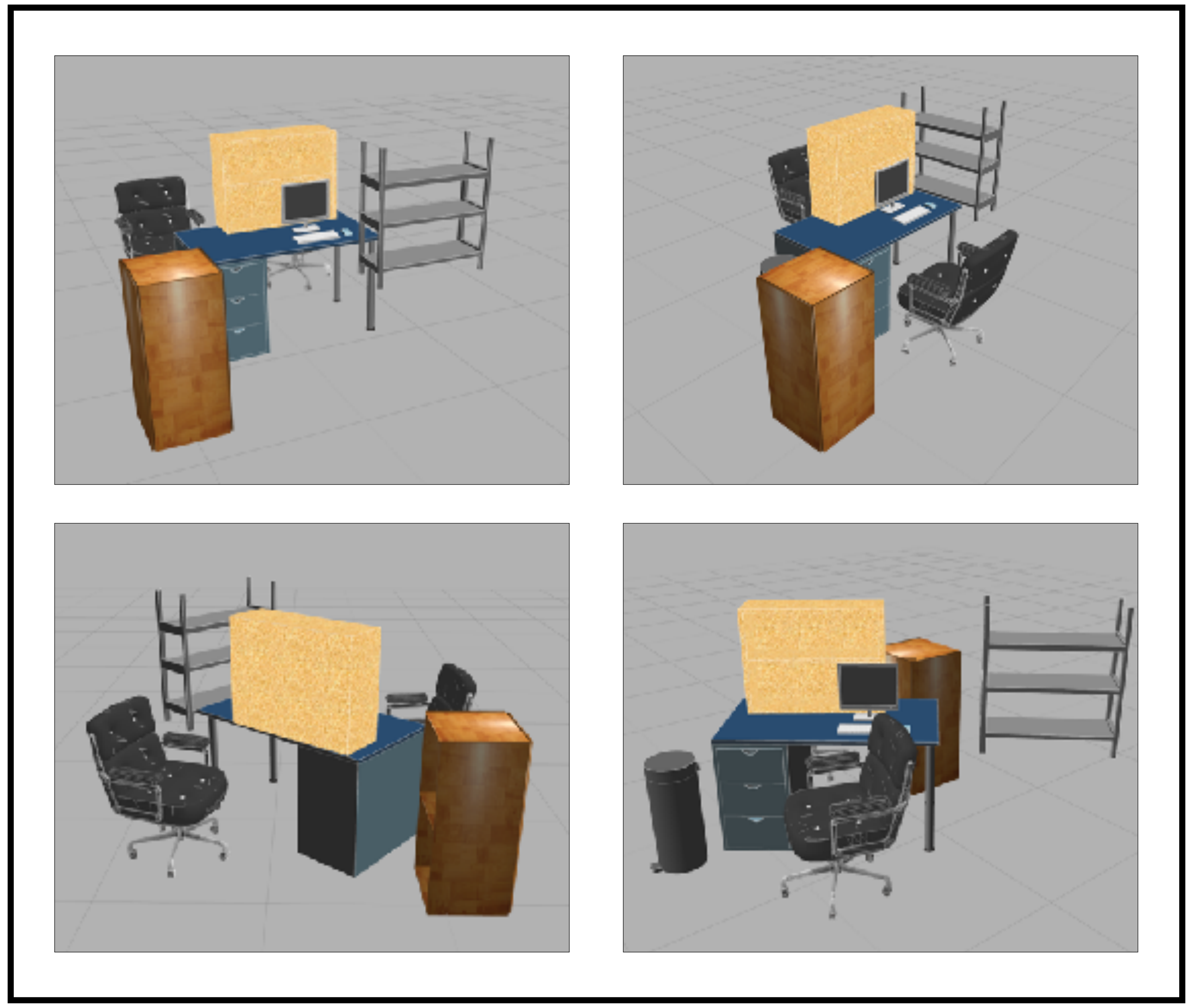
A* using dubins distance heuristic
times out (2531 states, 7000ms)



SAIL expands 18 states in 100 ms

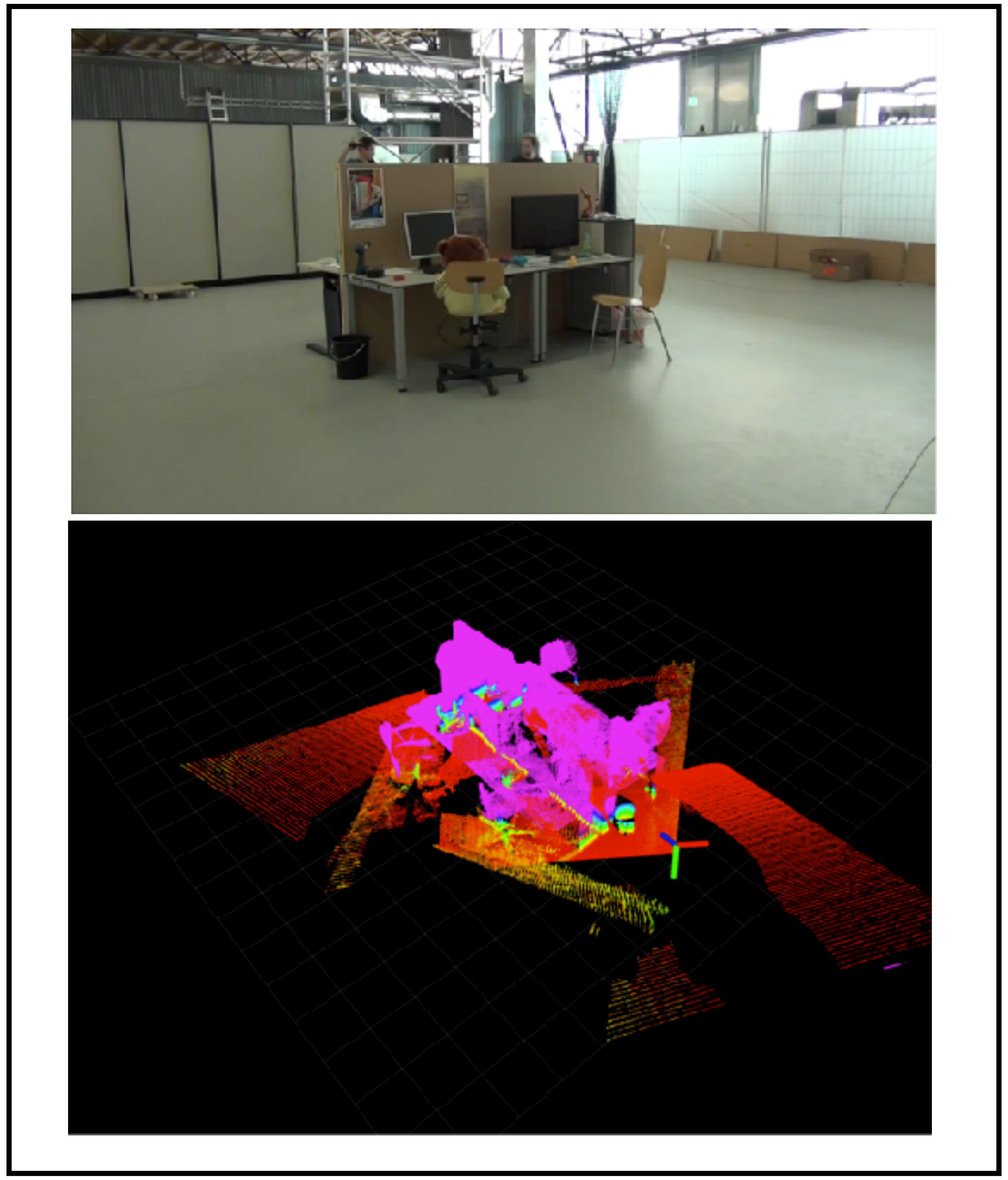
AGGREVATE for mapping unknown environments

Train Data: Office desks created in Gazebo



Learn Policy

Test Data: RGBD data (Sturm et al.)



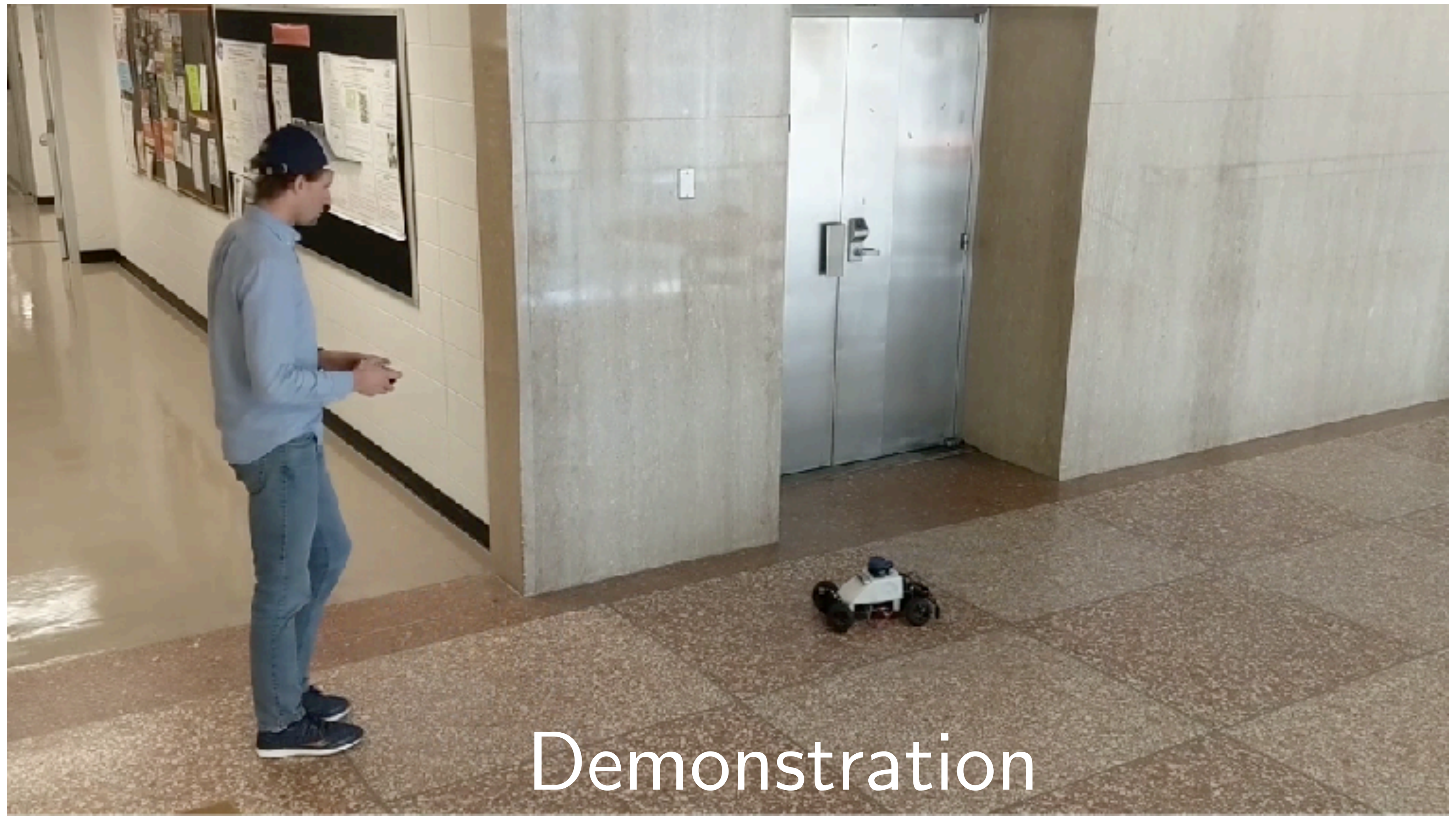
Okay ...
But how do we learn from natural human feedback?

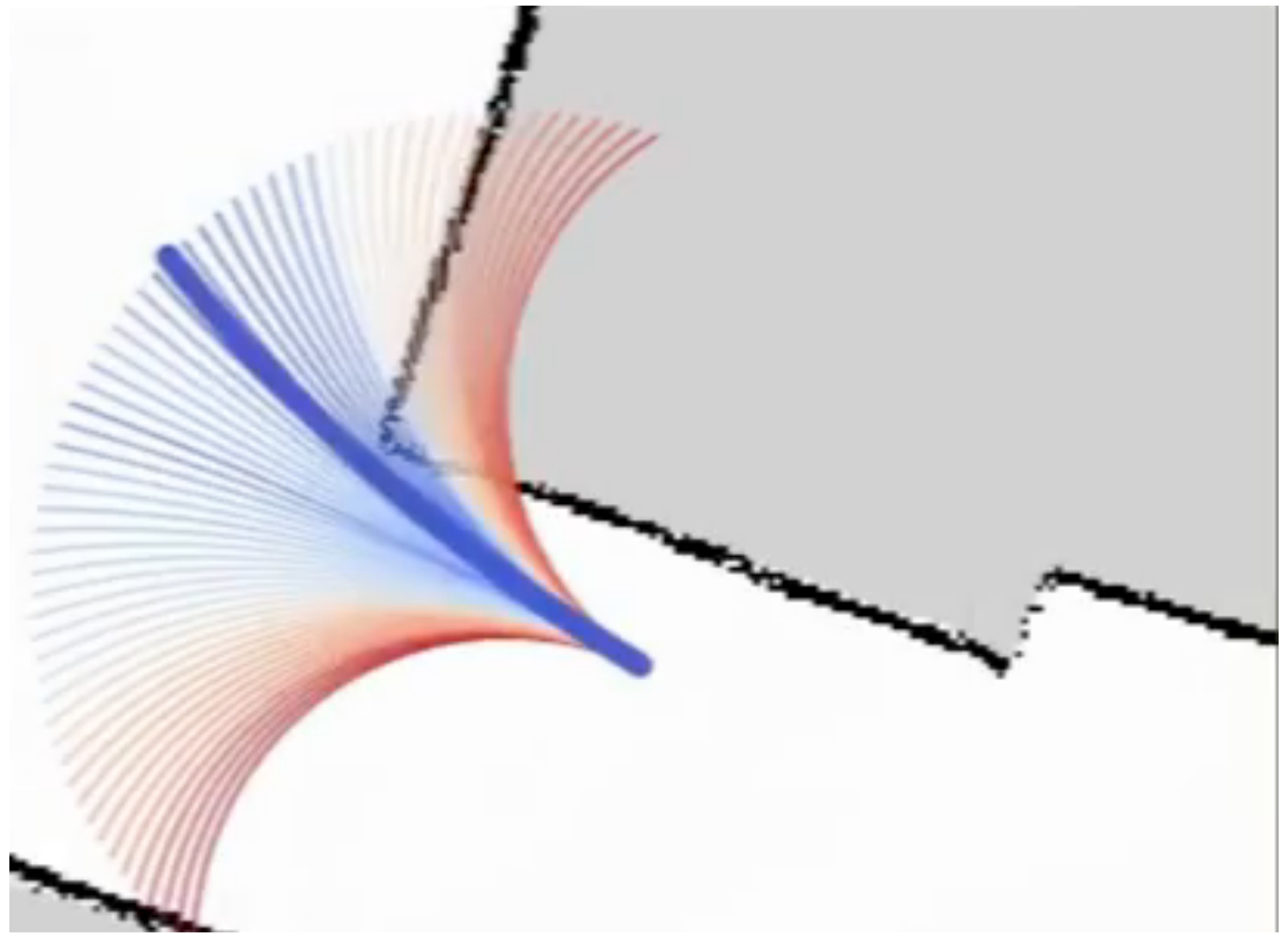
Impedance

Learning

Recap: Learning to drive

[SCB+ RSS'20]

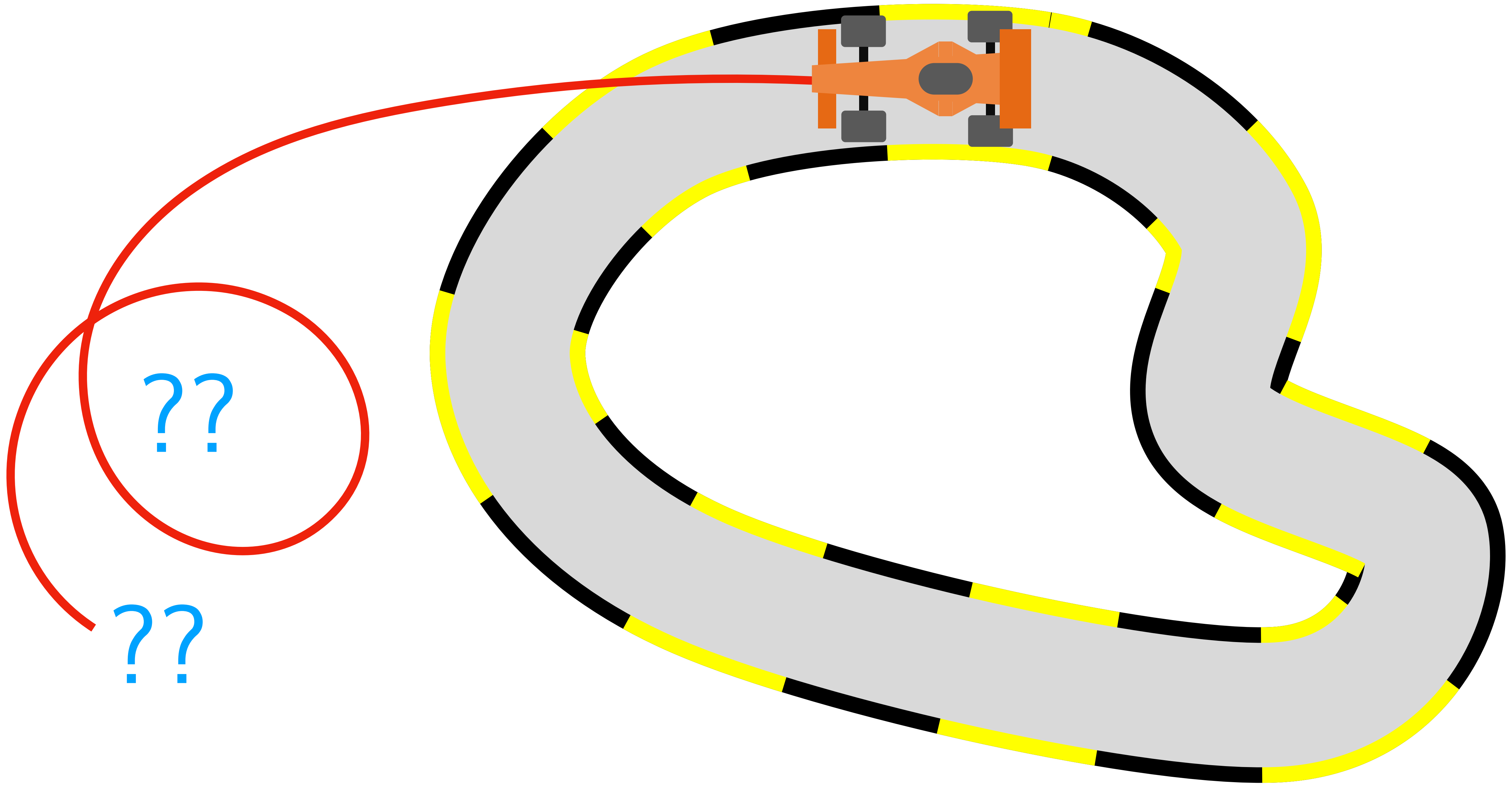




Learnt policy

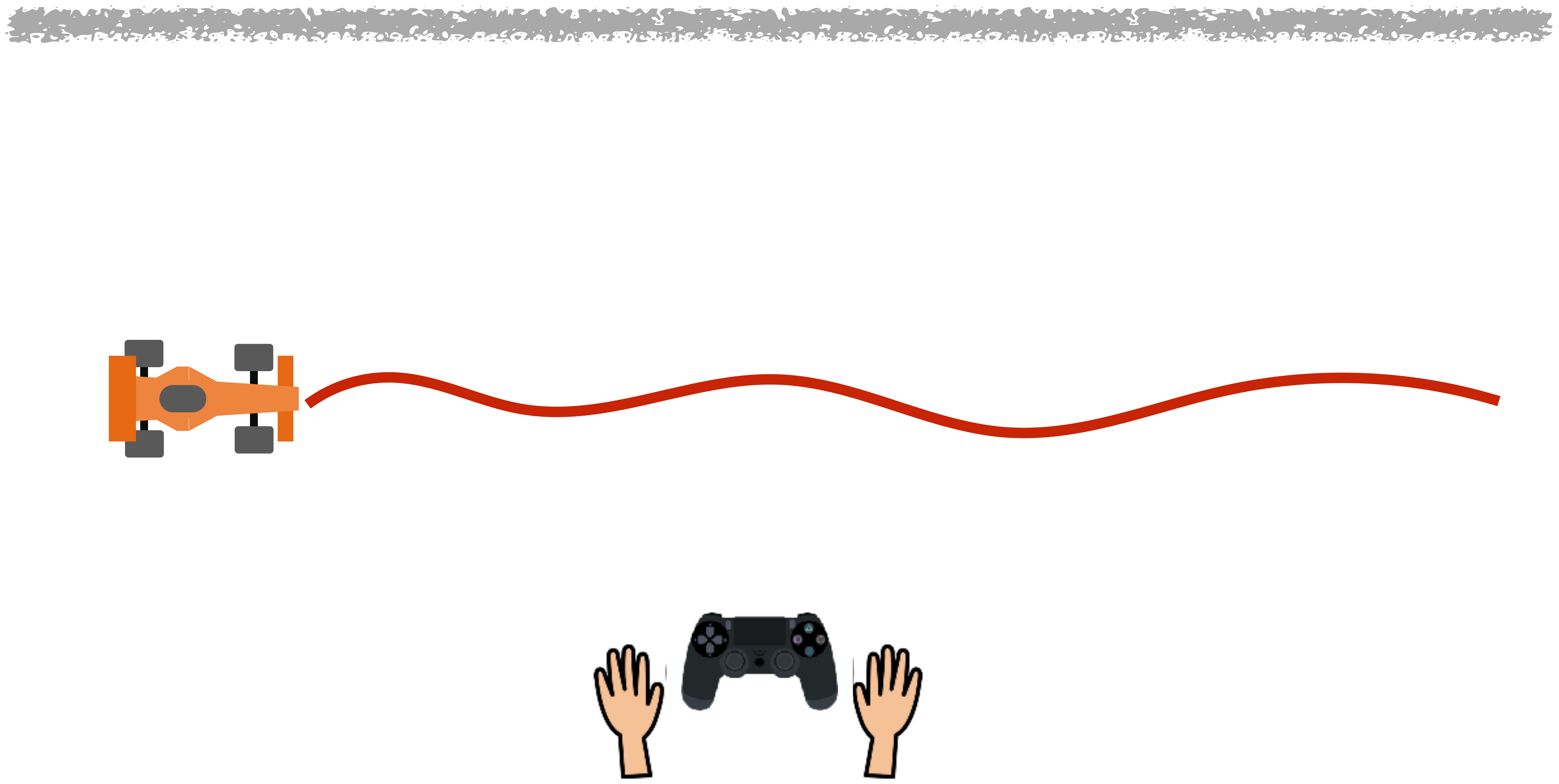
Behavior Cloning crashes into a wall

Problem: Impractical to query expert everywhere



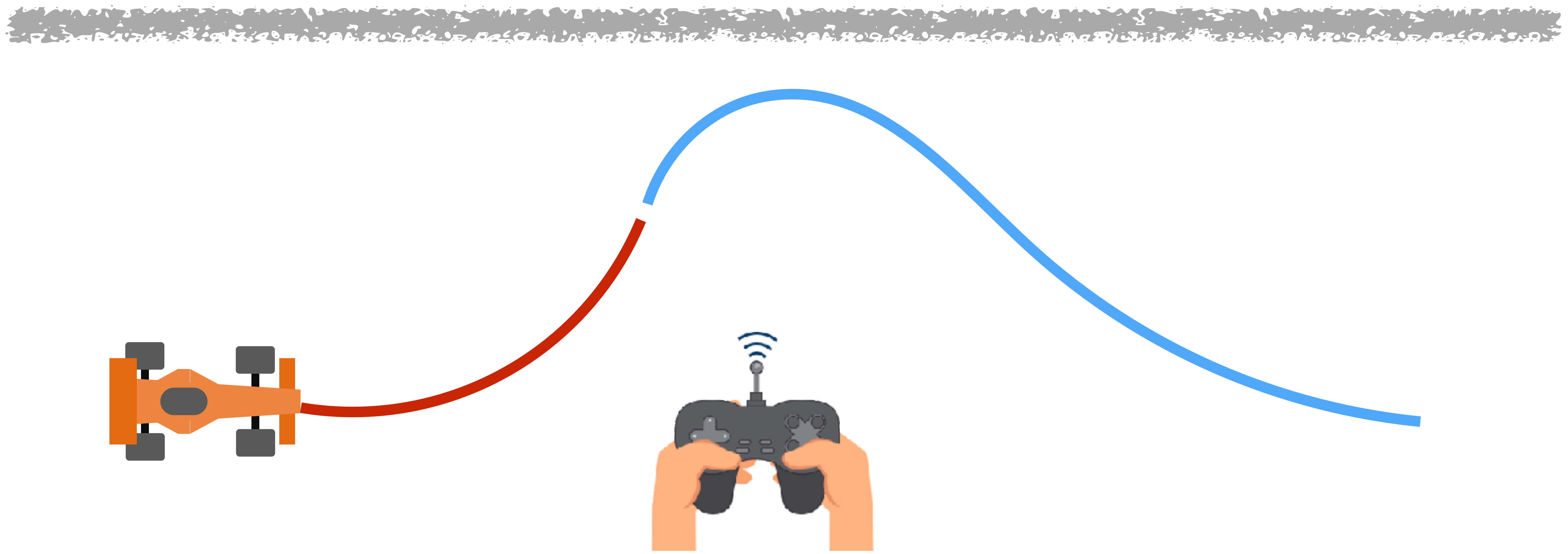
Can we learn from natural human interaction, e.g., interventions?

Learn from natural human interventions?



Hands free, no corrections!

Learn from natural human interventions?



Take over and drive back!

HG-DAGGER: Learning from interventions

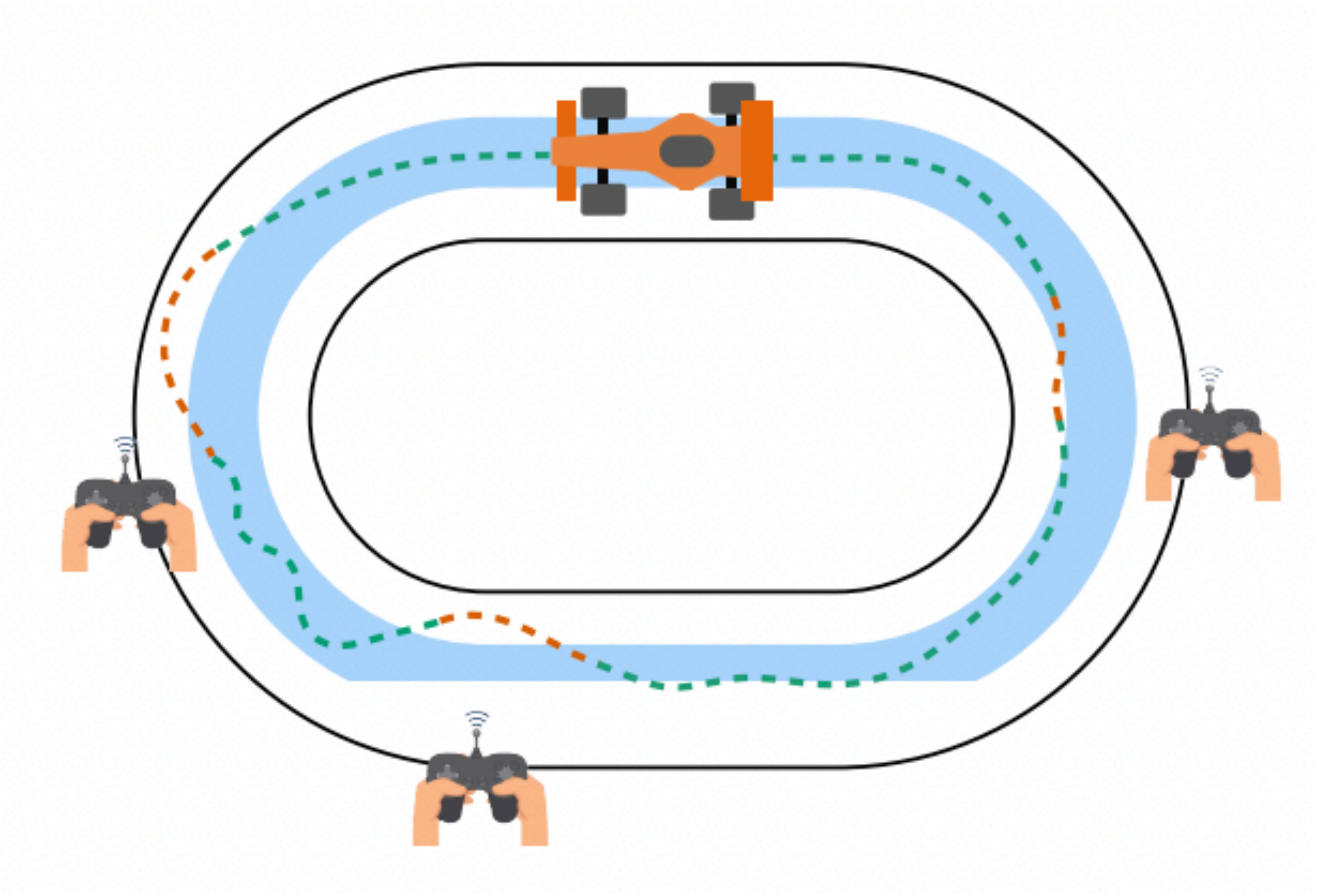
Roll out a learner policy

Collect expert actions on states where expert intervened

Aggregate data

Update policy

$$\min_{\pi} \mathbb{E}_{s,a^* \sim \mathcal{D}} 1(\pi(s) \neq a^*)$$



HG-DAgger: Interactive Imitation Learning with Human Experts

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer

Does this work?

Interventions are tell us something about the expert's latent value function

Expert Intervention Learning (EIL)

[SCB+ RSS'20]

The expert action-value function is latent ...

$$Q^*(s, a)$$

... and must be inferred from human interventions

Expert Intervention Learning (EIL)

[SCB+ RSS'20]

Interventions are just constraints on latent action-value function

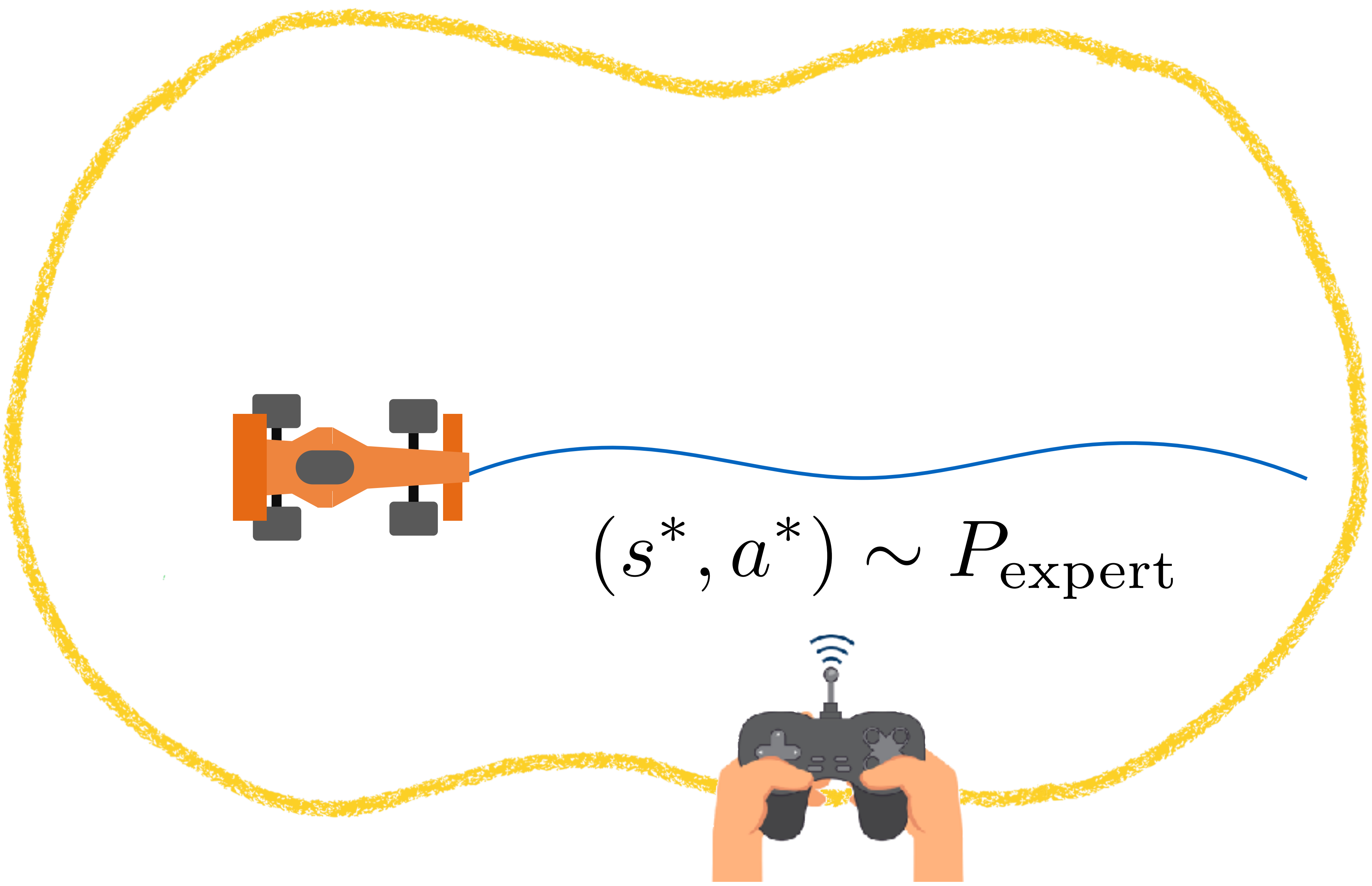
Expert Intervention Learning (EIL)

[SCB+ RSS'20]

Interventions are just constraints on latent action-value function

$$\min_{Q \in \mathcal{Q}} \mathbb{E}_{(s^*, a^*) \sim P_{\text{expert}}} \ell(Q(s^*, .), a^*)$$

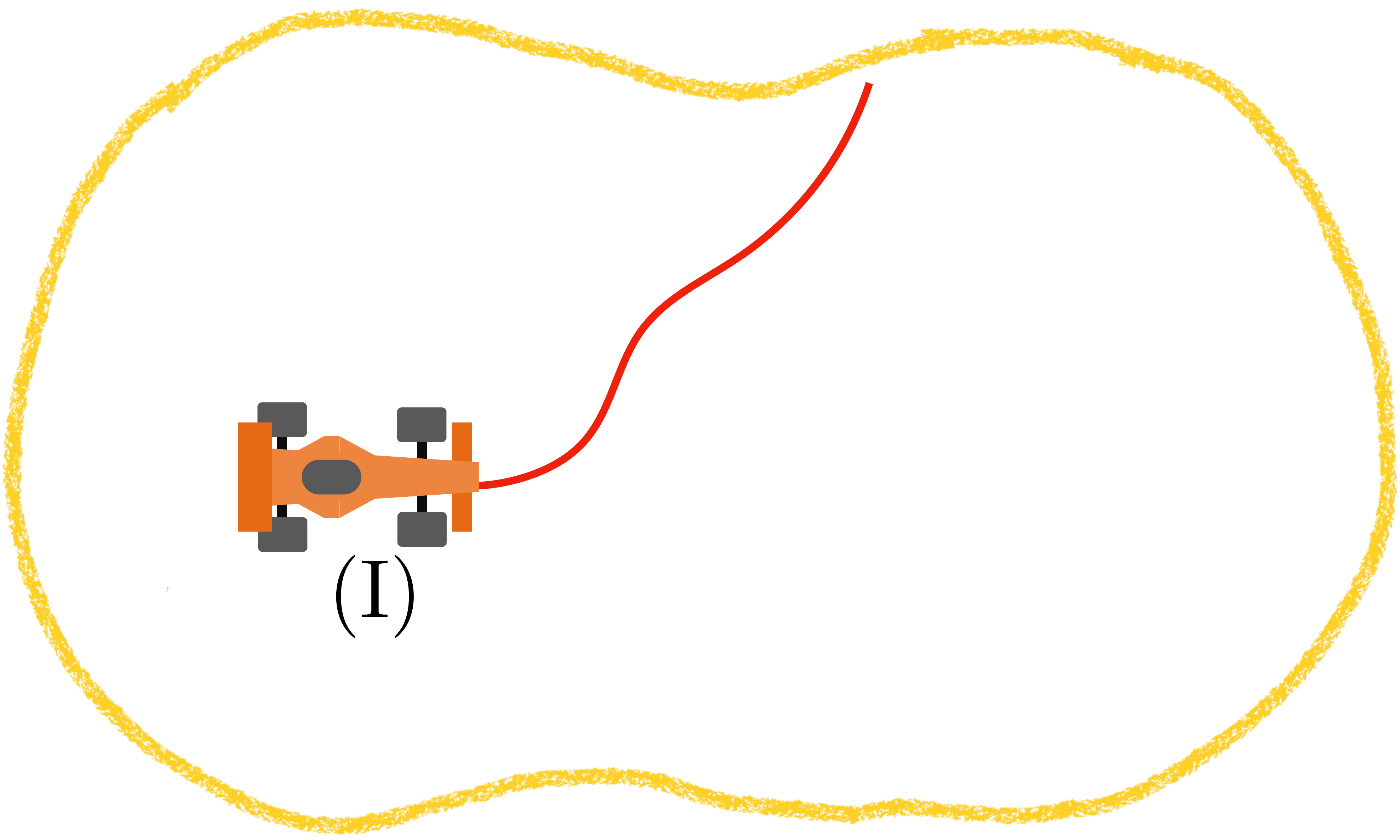
classify demonstrations



Expert Intervention Learning (EIL)

[SCB+ RSS'20]

Interventions are just constraints on latent action-value function



$$\min_{Q \in \mathcal{Q}} \mathbb{E}_{(s^*, a^*) \sim P_{\text{expert}}} \ell(Q(s^*, .), a^*)$$

classify demonstrations

s.t.

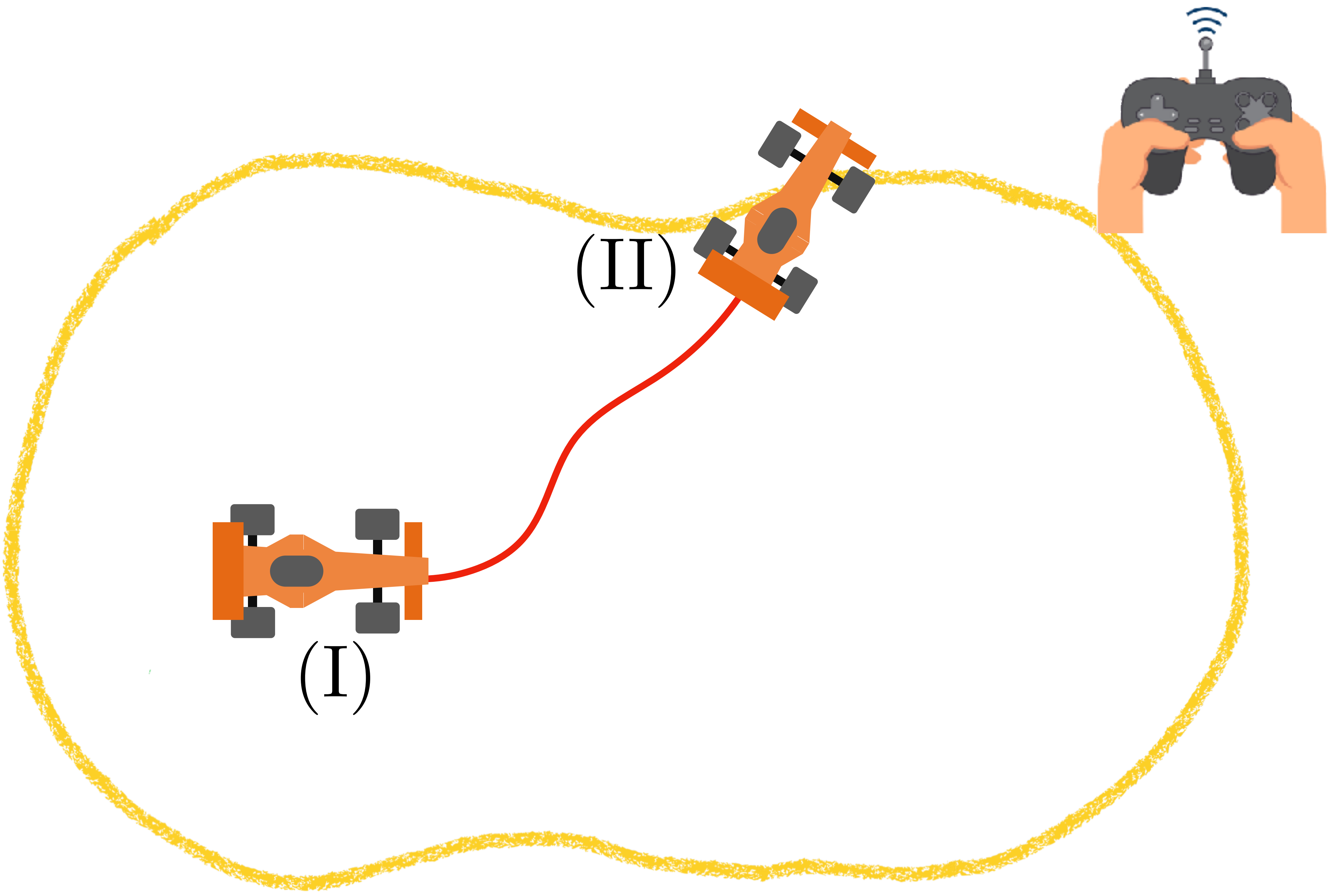
$$Q(s,a) \leq \delta_{\text{good}} \qquad \forall (s,a) \in \text{(I)}$$

before expert intervenes

Expert Intervention Learning (EIL)

[SCB+ RSS'20]

Interventions are just constraints on latent action-value function



$$\min_{Q \in \mathcal{Q}} \mathbb{E}_{(s^*, a^*) \sim P_{\text{expert}}} \ell(Q(s^*, .), a^*)$$

classify demonstrations

s.t.

$$Q(s, a) \leq \delta_{\text{good}} \qquad \forall (s, a) \in (\text{I})$$

before expert intervenes

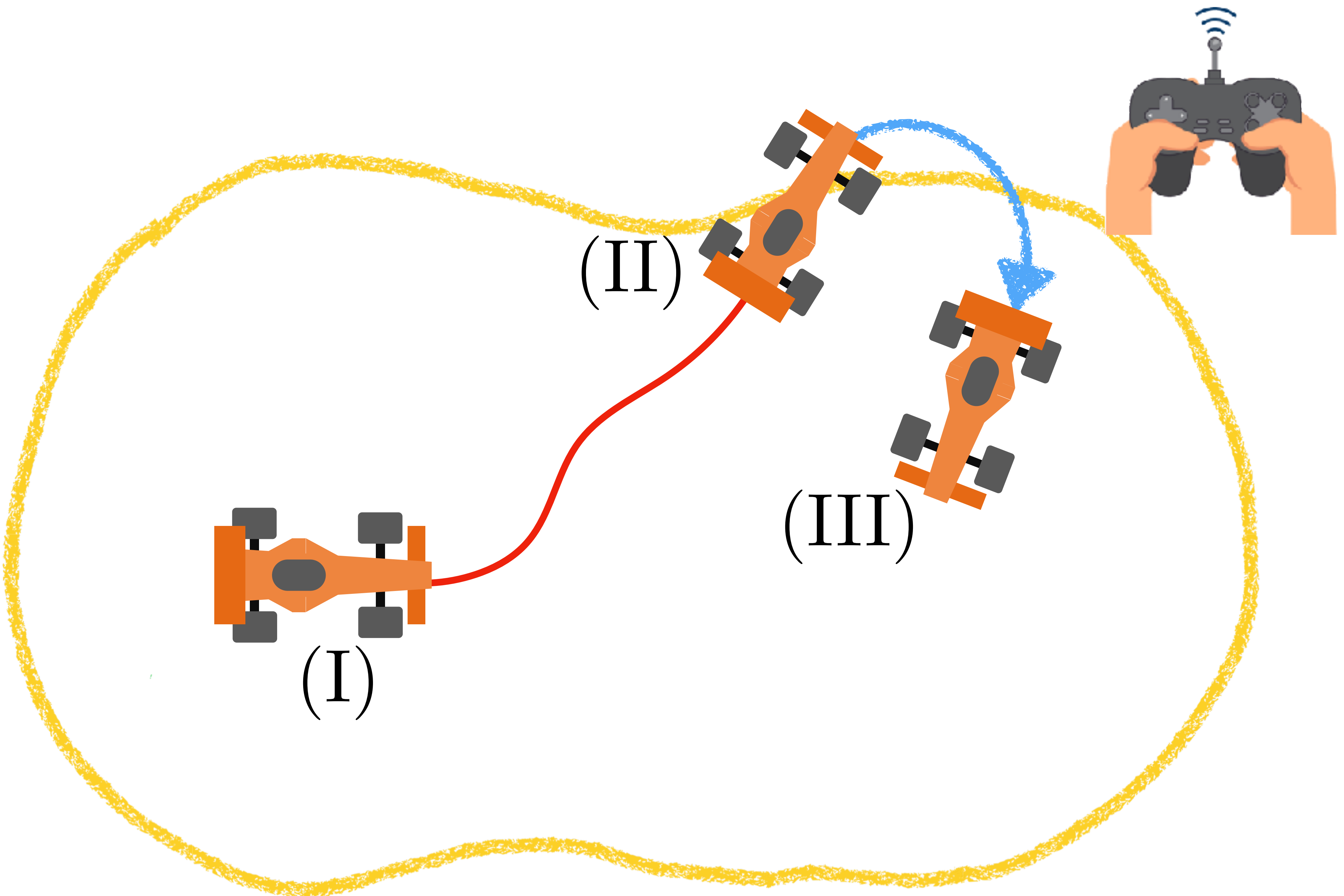
$$Q(s, a) \geq \delta_{\text{good}} \qquad \forall (s, a) \in (\text{II})$$

after expert intervenes

Expert Intervention Learning (EIL)

[SCB+ RSS'20]

Interventions are just constraints on latent action-value function



$$\min_{Q \in \mathcal{Q}} \mathbb{E}_{(s^*,a^*) \sim P_{\text{expert}}} \ell(Q(s^*,.), a^*)$$

classify demonstrations

s.t.

$$Q(s,a) \leq \delta_{\text{good}}$$

$\forall (s,a) \in (\mathrm{I})$
before expert intervenes

$$Q(s,a) \geq \delta_{\text{good}}$$

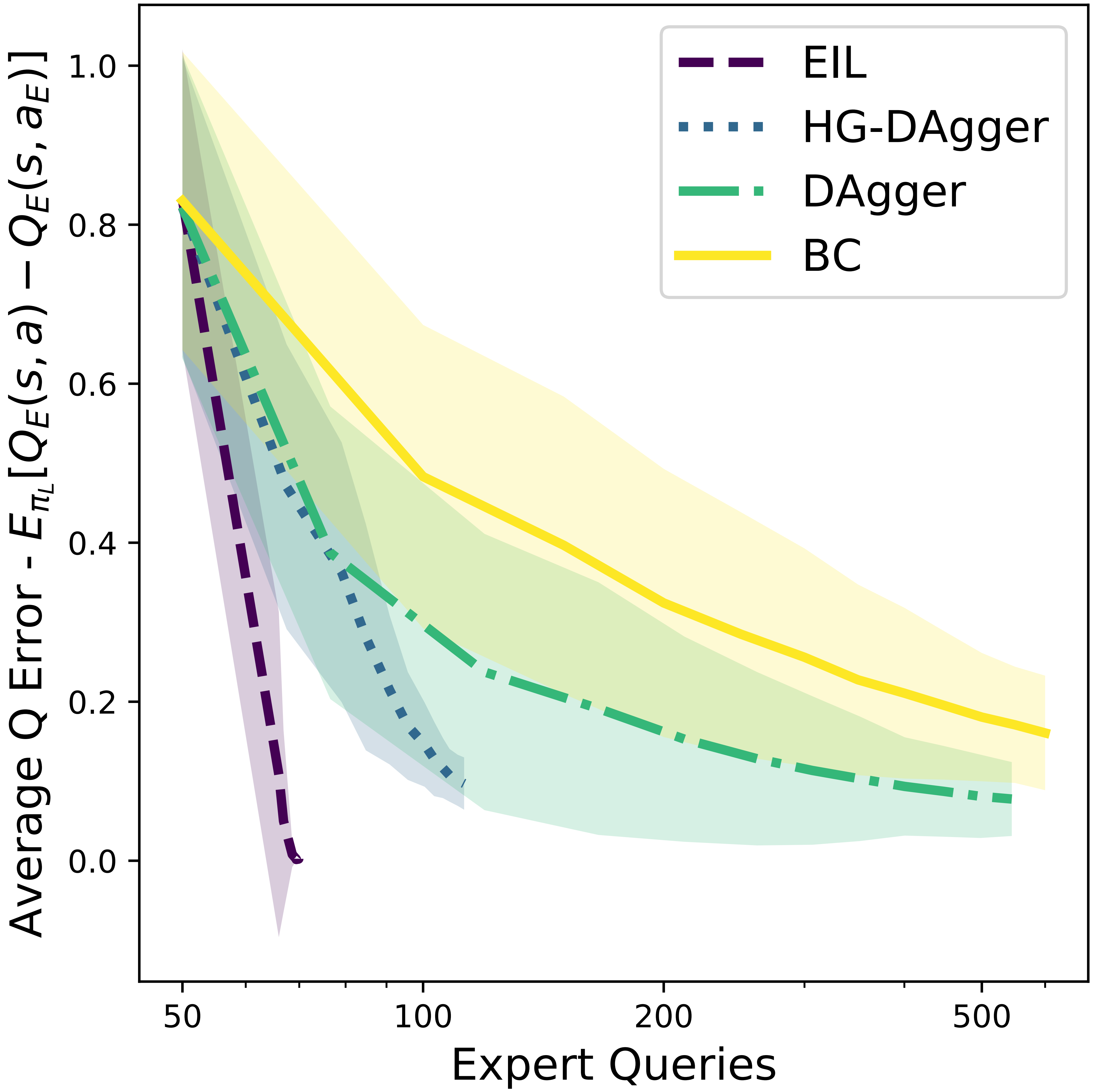
$\forall (s,a) \in (\mathrm{II})$
after expert intervenes

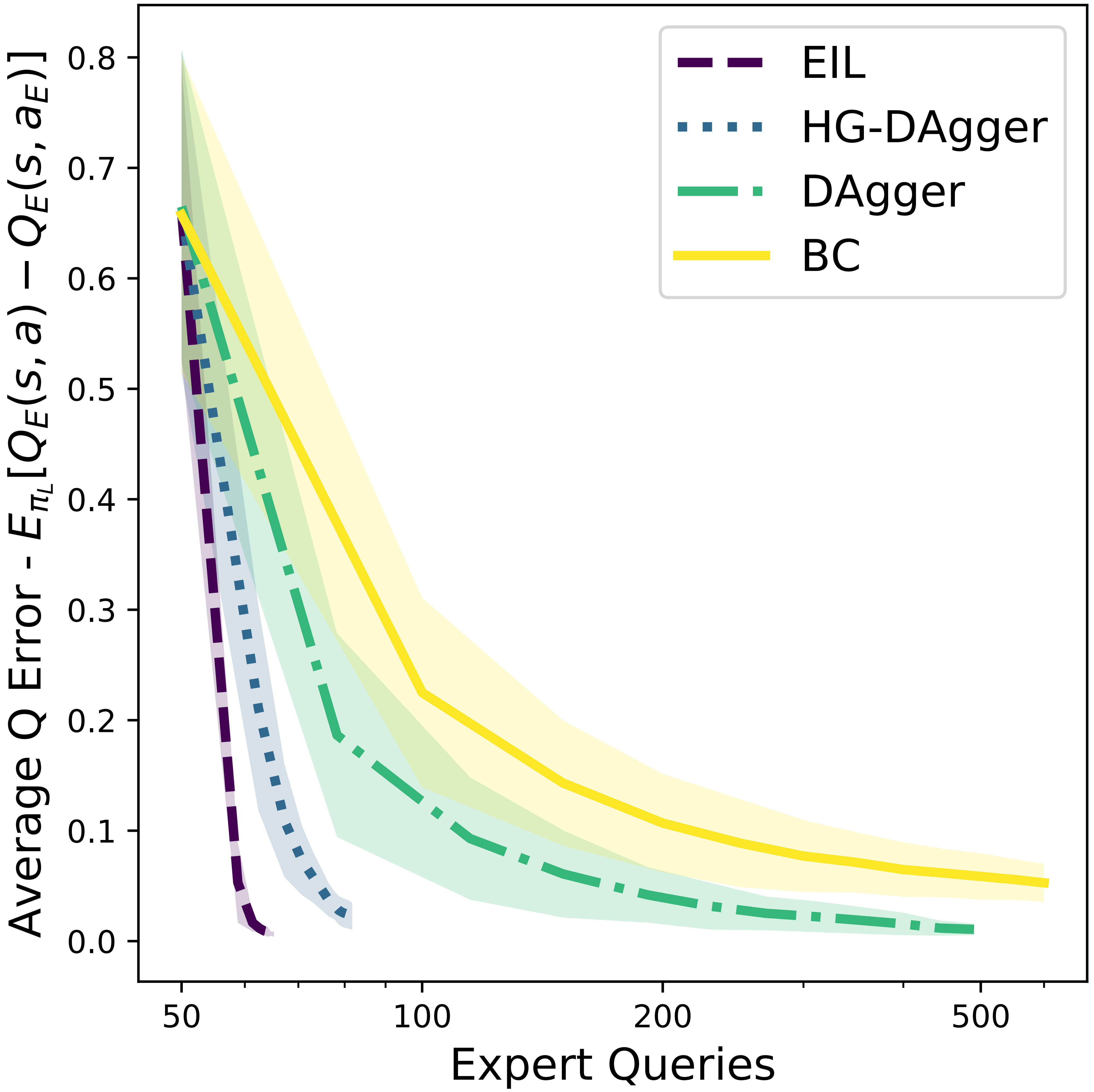
$$Q(s,a) \leq \min_{a'} Q(s,a)$$

$\forall (s,a) \in (\mathrm{III})$
during expert intervention

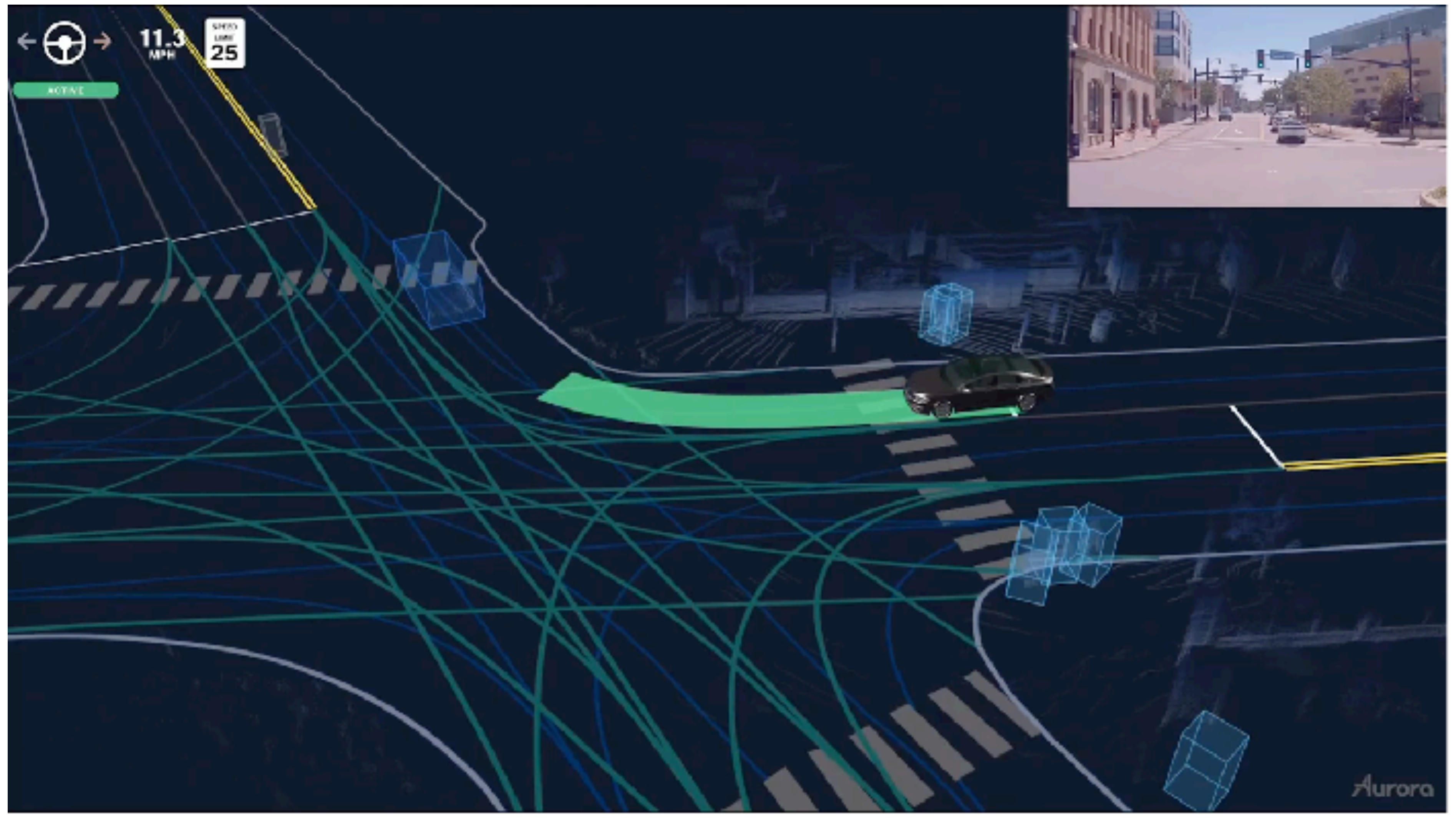
Reduce to online, convex optimization! $O(\epsilon T)$

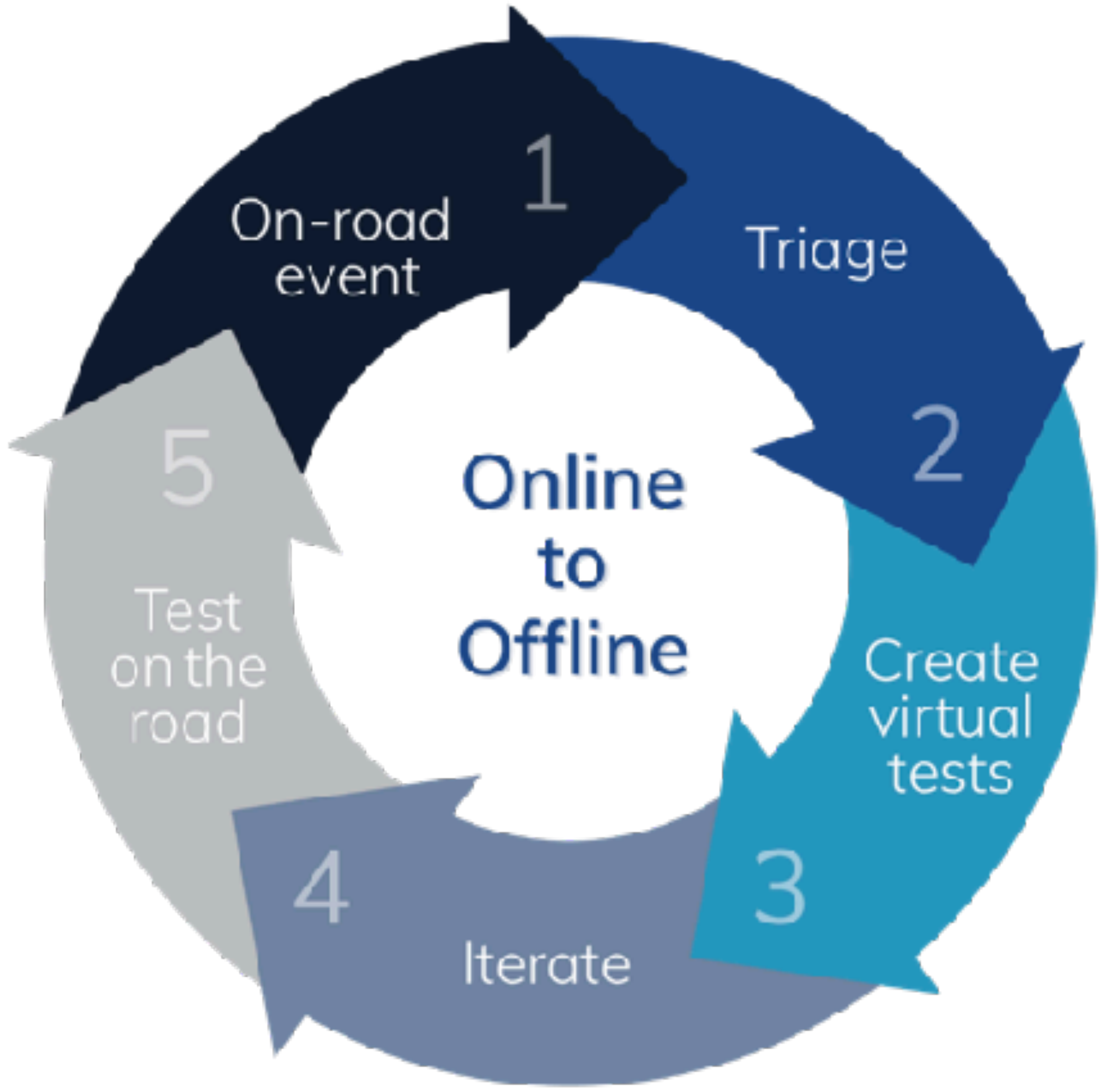
EIL drives down error with less expert query

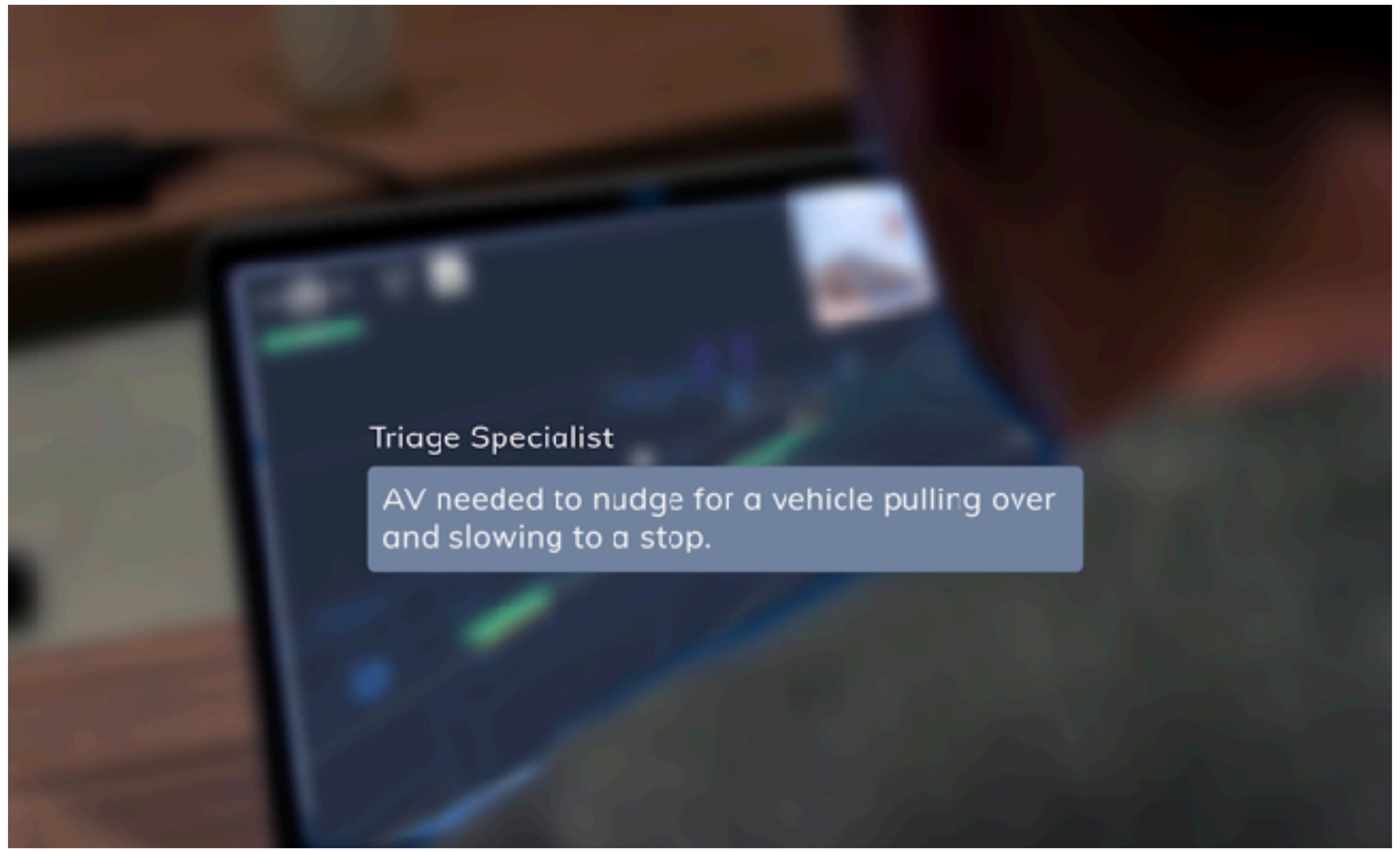


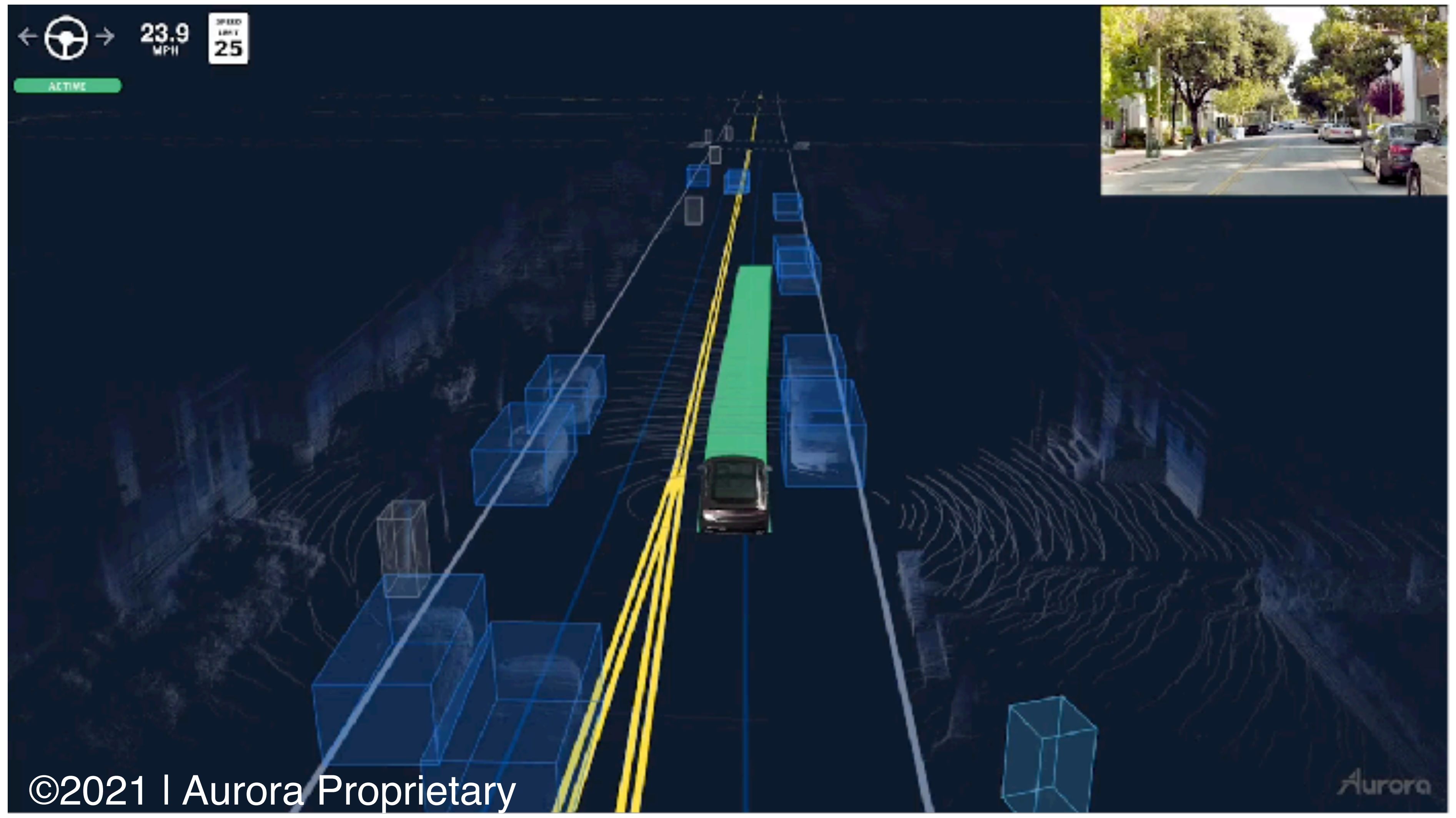


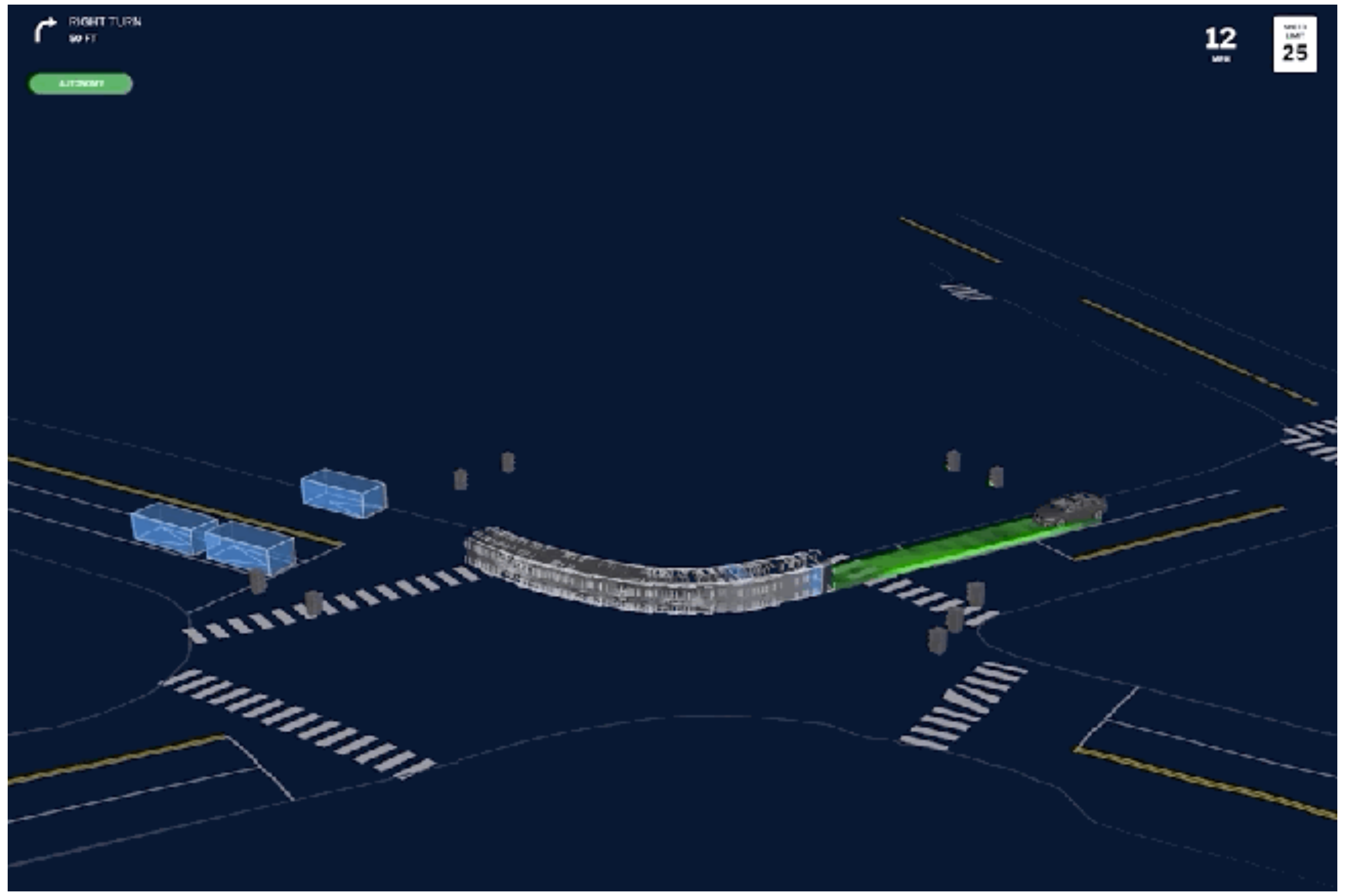
Turning interventions to simulations for learner











https://medium.com/aurora-blog/online-to-offline-turning-real-world-experience-into-virtual-tests-231c1cf8cbcd

The Big Picture

What we really want to solve is:

$$\min_{\pi} \mathbb{E}_{s \sim d_\pi}[Q^*(s, \pi(s)) - Q^*(s, \pi^*(s))]$$

Data

✓ *"What is the distribution of states?"*

Use interactive online learning!

Loss

✓ *"What is the metric to match to human?"*

Difference in Q values!

The Big Picture

What we really want to solve is:

$$\min_{\pi} \mathbb{E}_{s \sim d_\pi}[Q^*(s, \pi(s)) - Q^*(s, \pi^*(s))]$$

$$Q^*(s, a)$$

Loss

"What is the metric to match to human?"

Difference in Q values!

But Q* is latent!

The Big Picture

Estimate Q* from demonstrations, interventions, preferences, ..
and even E-stops!

Demonstrations

Interventions

Preferences

E-stops

$\mathscr{L}(Q^*_\theta)$

Loss

tl;dr

The Big Picture

What we really want to solve is:

$$\min_{\pi} \mathbb{E}_{s \sim d_\pi}[Q^*(s, \pi(s)) - Q^*(s, \pi^*(s))]$$

Data

✓ *"What is the distribution of states?"*

Use interactive online learning!

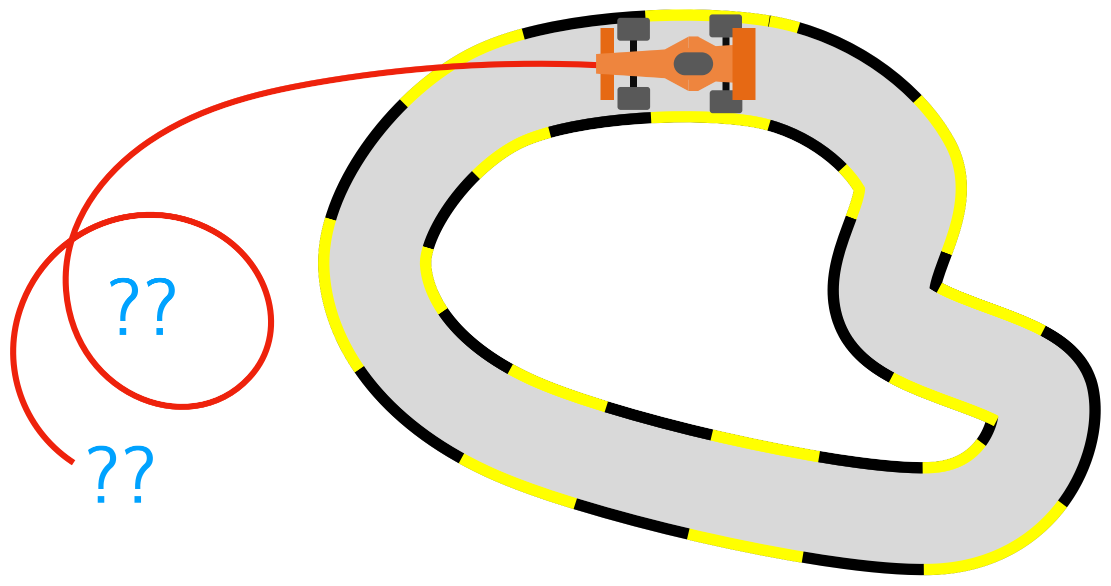
Loss

✓ *"What is the metric to match to human?"*

Difference in Q values!

x

Problem: Impractical to query expert everywhere



Can we learn from natural human interaction, e.g., interventions?

x

Expert Intervention Learning (EIL)

[SCB+ RSS'20]

The expert action-value function is latent ...

... and must be inferred from human interventions

x