Imitation Learning as Inferring Latent Expert *Values*

Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

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 ALGORITHM

$\pi_0 \leftarrow$ INITIALIZE POLICY WITH BEHAVIOR CLONING.

$D \leftarrow \{\ \}$ INITIALIZE EMPTY DATA BUFFER

FOR $i = 1 \dots N$

ROLLOUT π_i

$(s_1, a_1, s_2, a_2, \dots\dots)$

QUERY HUMAN π^* FOR CORRECT ACTIONS

$(s_1, \pi^*(s_1), s_2, \pi^*(s_2), \dots\dots)$

$$D \leftarrow D \cup \{(s_1, \pi^*(s_1), s_2, \pi^*(s_2), \dots\dots)\}$$

$\pi_i \leftarrow$ TRAIN (D)

DAGGER ALGORITHM

$\pi_0 \leftarrow$ INITIALIZE POLICY WITH BEHAVIOR CLONING.

$D \leftarrow \{\ \}$ INITALIZE EMPTY DATA BUFFER

FOR $i = 1 \ldots N$

ROLLOUT π_i

$(s_1, a_1, s_2, a_2, \ldots)$

QUERY HUMAN π^* FOR CORRECT ACTIONS

$(s_1, \pi^*(s_1), s_2, \pi^*(s_2), \ldots)$

$D \leftarrow D \cup \{(s_1, \pi^*(s_1), s_2, \pi^*(s_2), \ldots)\}$

$\pi_i \leftarrow$ TRAIN (D)

By training on *aggregated* data

π_i is playing

Follow the (Regularized) Leader!

$$l_i(\pi) = \mathbb{E}_{s \sim d_\pi} 1(\pi(s) \neq \pi^*(s))$$

$$\pi_{i+1} = \arg\min_{\pi} \sum_{j=0}^{i} l_i(\pi) + R(\pi)$$

DAGGER ALGORITHM

$\pi_0 \leftarrow$ INITIALIZE POLICY WITH BEHAVIOR CLONING.

$D \leftarrow \{\}$ INITIALIZE EMPTY DATA BUFFER

FOR $i = 1 \ldots N$

ROLLOUT π_i

$(s_1, a_1, s_2, a_2, \ldots)$

QUERY HUMAN π^* FOR CORRECT ACTIONS

$(s_1, \pi^*(s_1), s_2, \pi^*(s_2), \ldots)$

$D \leftarrow D \cup \{(s_1, \pi^*(s_1), s_2, \pi^*(s_2), \ldots)\}$

$\pi_i \leftarrow$ TRAIN (D)

DAGGER results in an imitation gap of $O(\epsilon T)$

Assume the best policy in our policy class can drive down average loss to ϵ

Then DAGGER finds a policy π_i

$$J(\pi_i) - J(\pi^*) \leq T\, l_i(\pi_i)$$
$$\leq T\, \epsilon$$

Original results from DAGGER!

DAGGER is a foundation

Imitation under uncertainty

SAIL

ExPLORE STROLL

Counterfactual Teaching

DPI LOLS

NRPI

Reinforcement Learning

Agnostic SysID DaaD

Model learning

DAEQUIL

AGGREVATE(D)

Imitation learning

EIL

HG-DAGGER

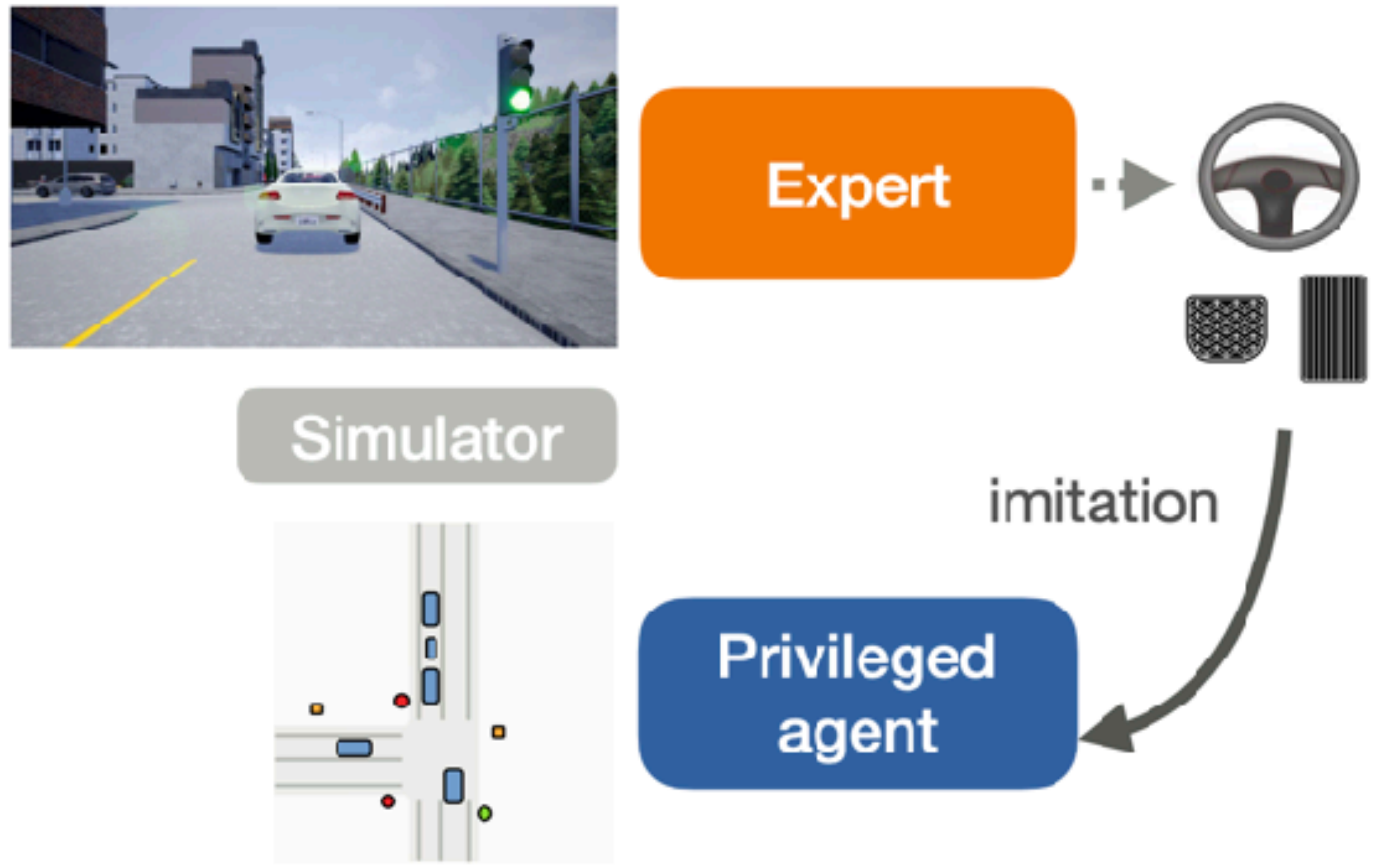
SHIV

Query efficient imitation learning

DAGGER

Many cool applications of DAGGER in robotics

Lee et al, Learning quadrupedal locomotion over challenging terrain (2020)

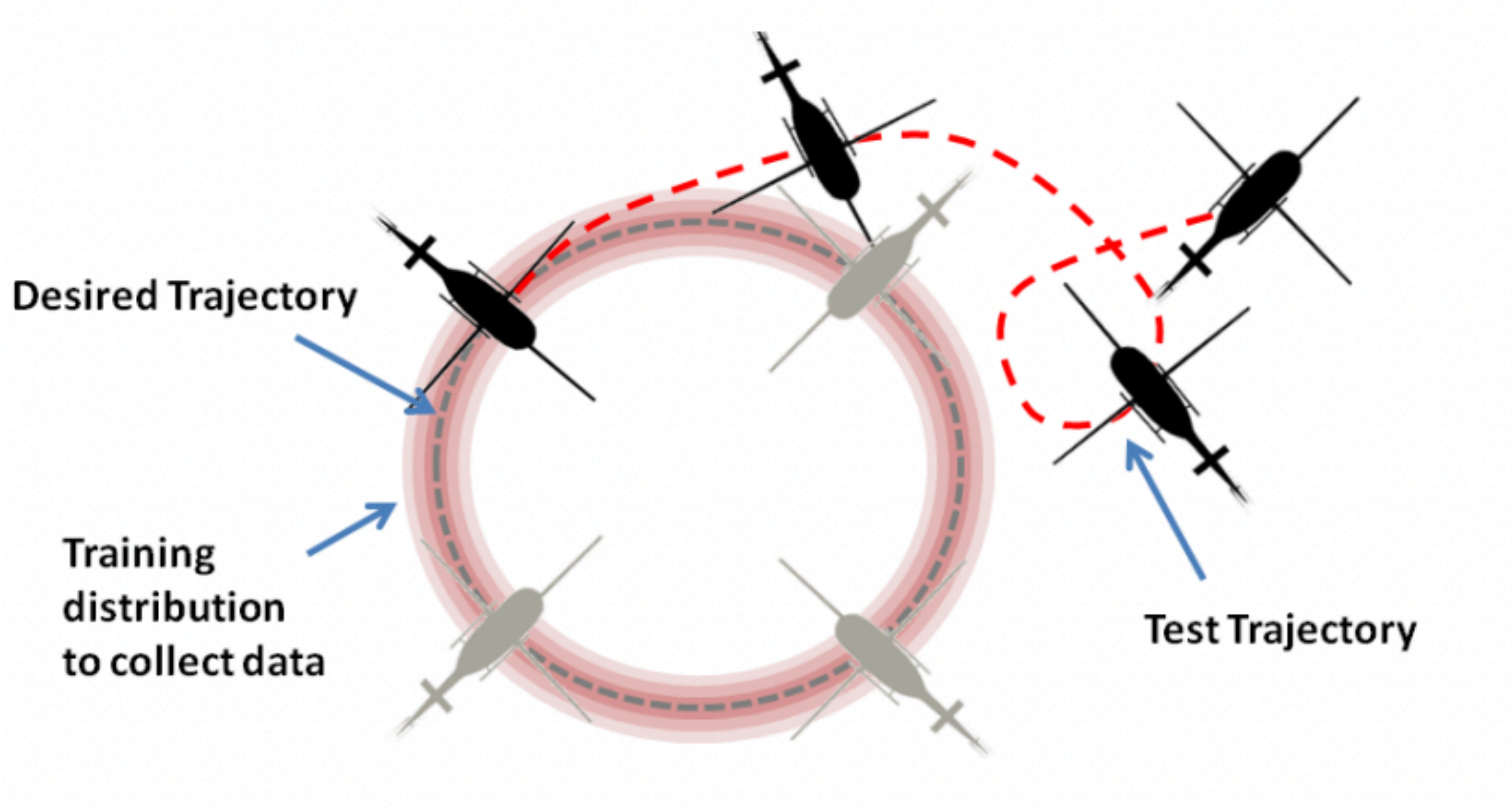


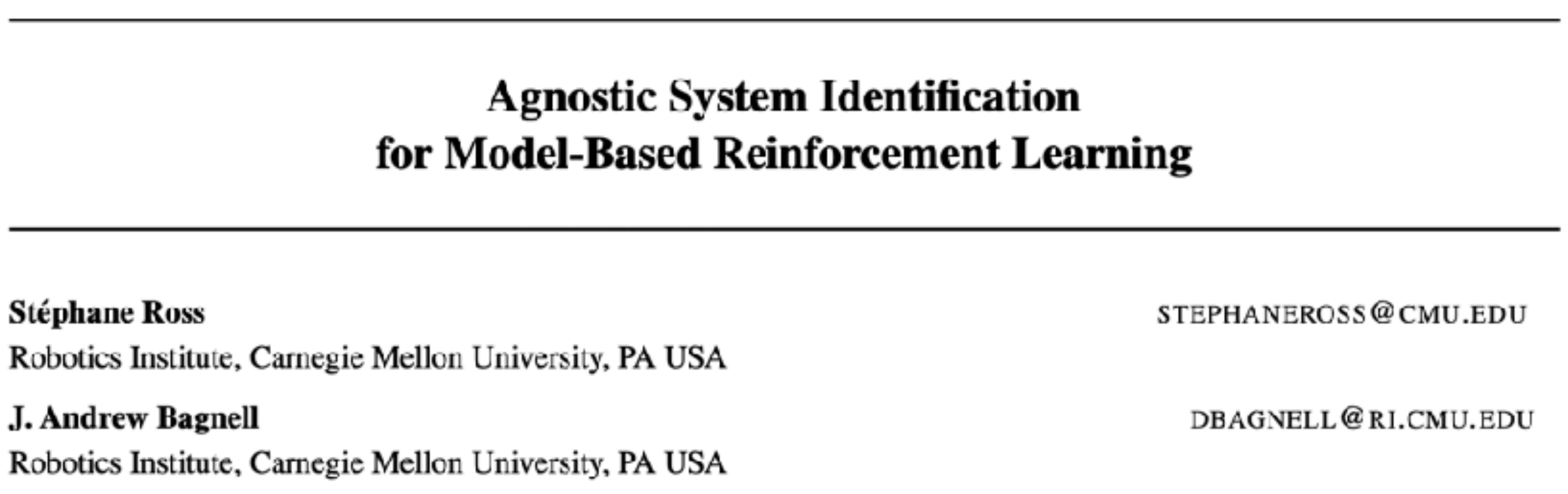
Chen et al Learning by Cheating(2020)

Choudhury et al, Data Driven Planning via Imitation Learning (2018)

Pan et al Imitation learning for agile autonomous driving (2019)

DAGGER is not *just* for imitation learning!





Agnostic System Identification for Model-Based Reinforcement Learning

Stéphane Ross STEPHANEROSS@CMU.EDU
Robotics Institute, Carnegie Mellon University, PA USA

J. Andrew Bagnell DBAGNELL@RI.CMU.EDU
Robotics Institute, Carnegie Mellon University, PA USA

Model-based Reinforcement Learning

Hidden charges from DAGGER

Hidden Charge #1:

Not all errors are equal

Recap: DAGGER

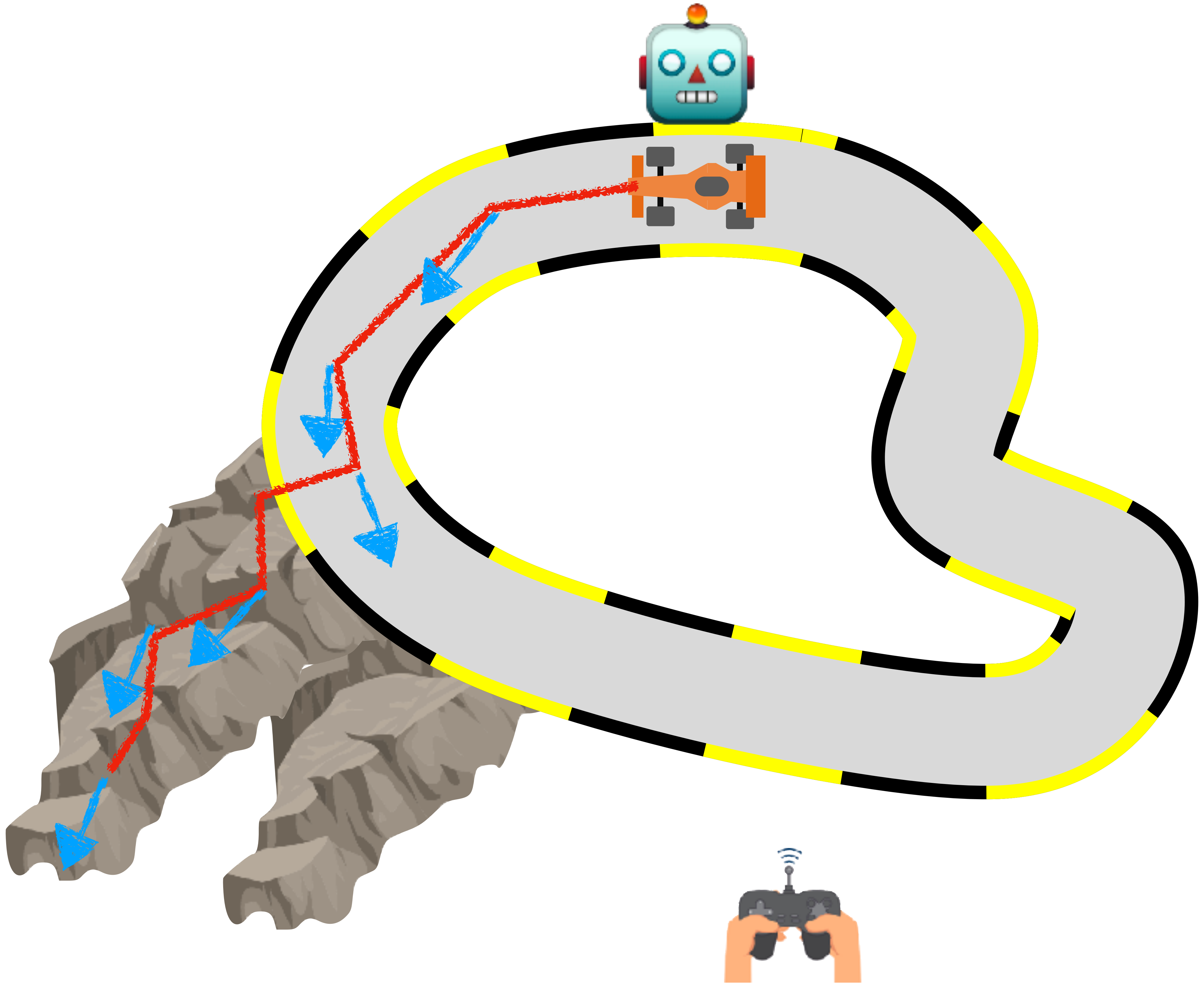
Roll out a learner policy

Collect expert actions

Aggregate data

Update policy

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

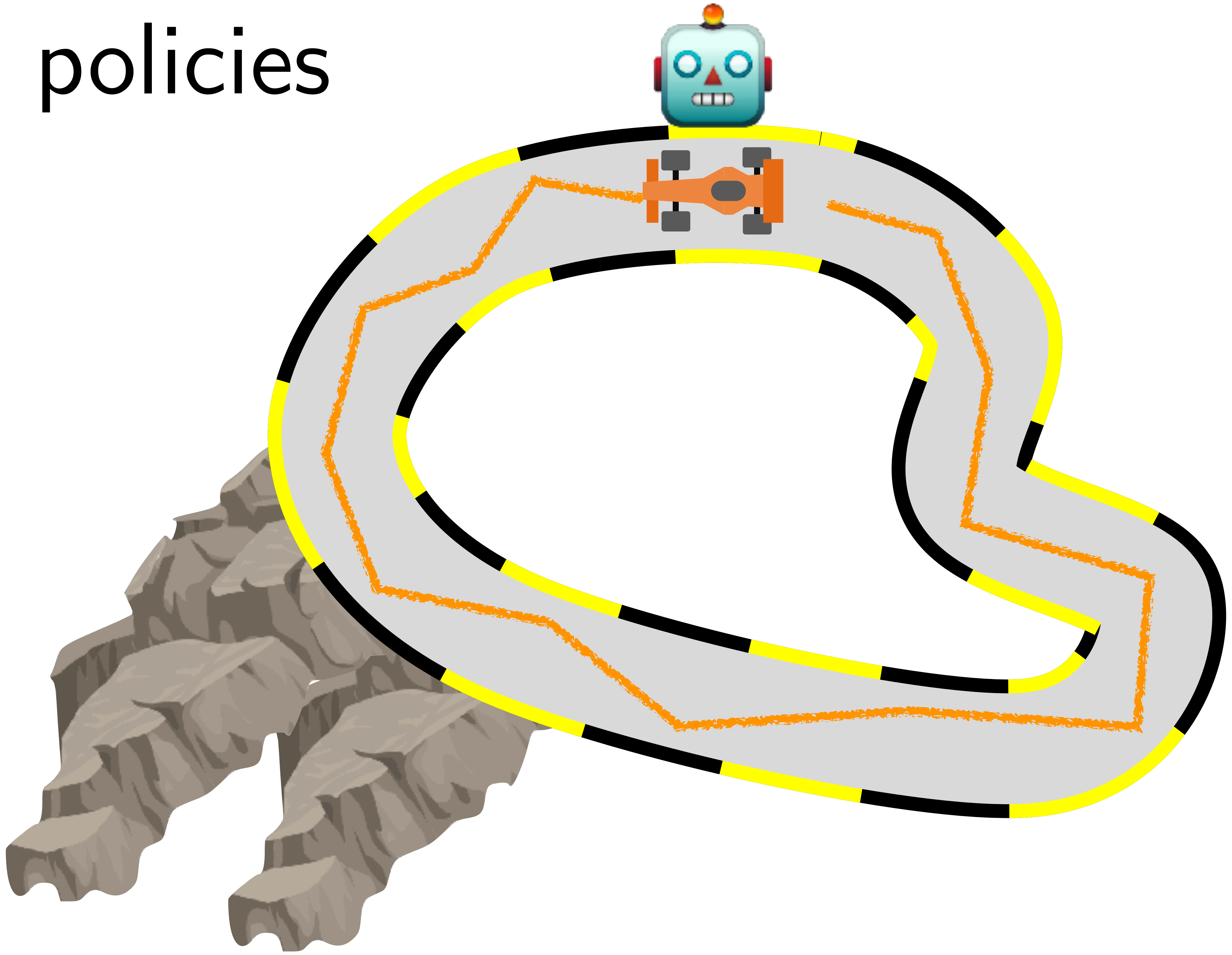


What does DAGGER guarantee?

Let's say your policy class Π has 2 policies

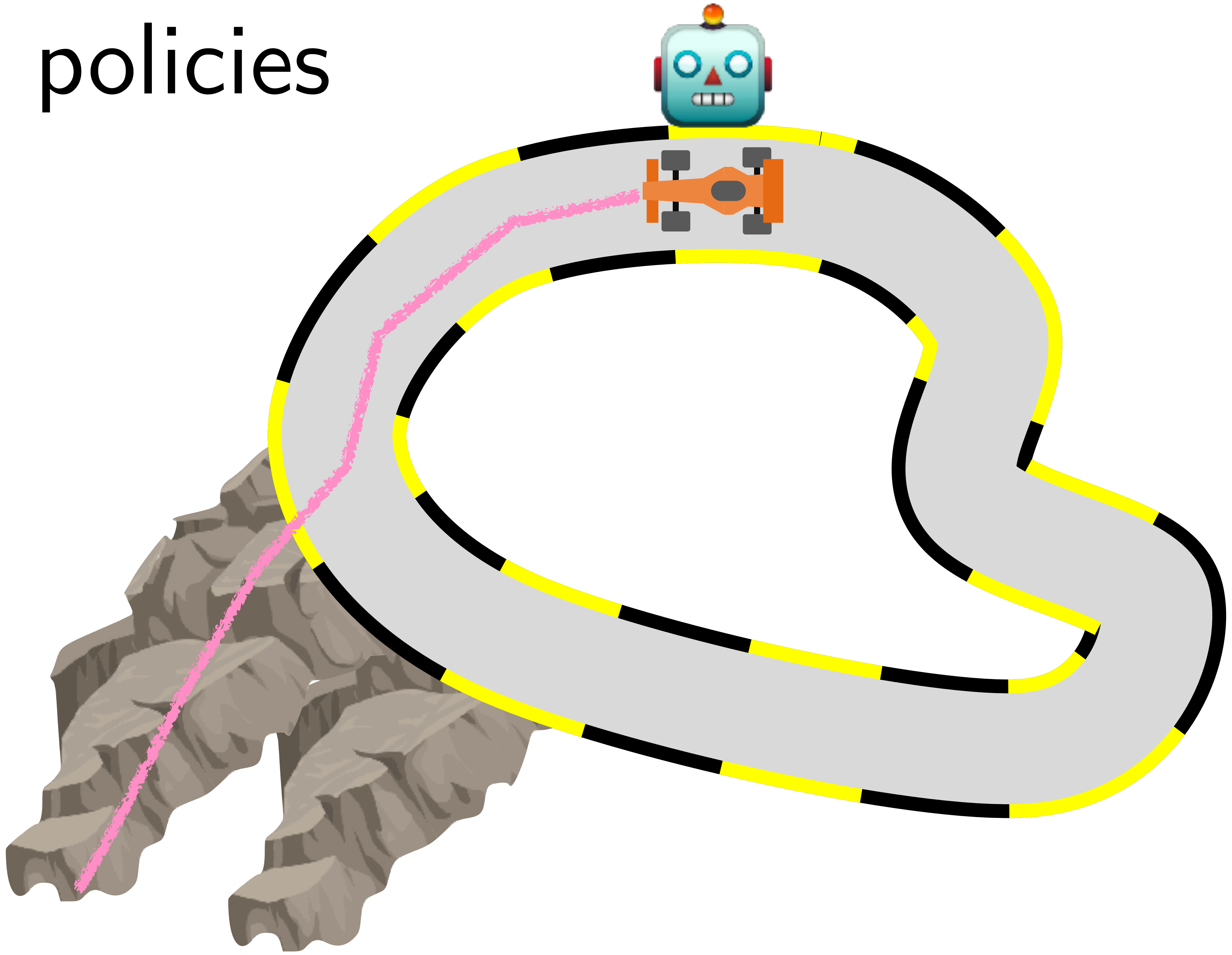
Policy π_1:

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



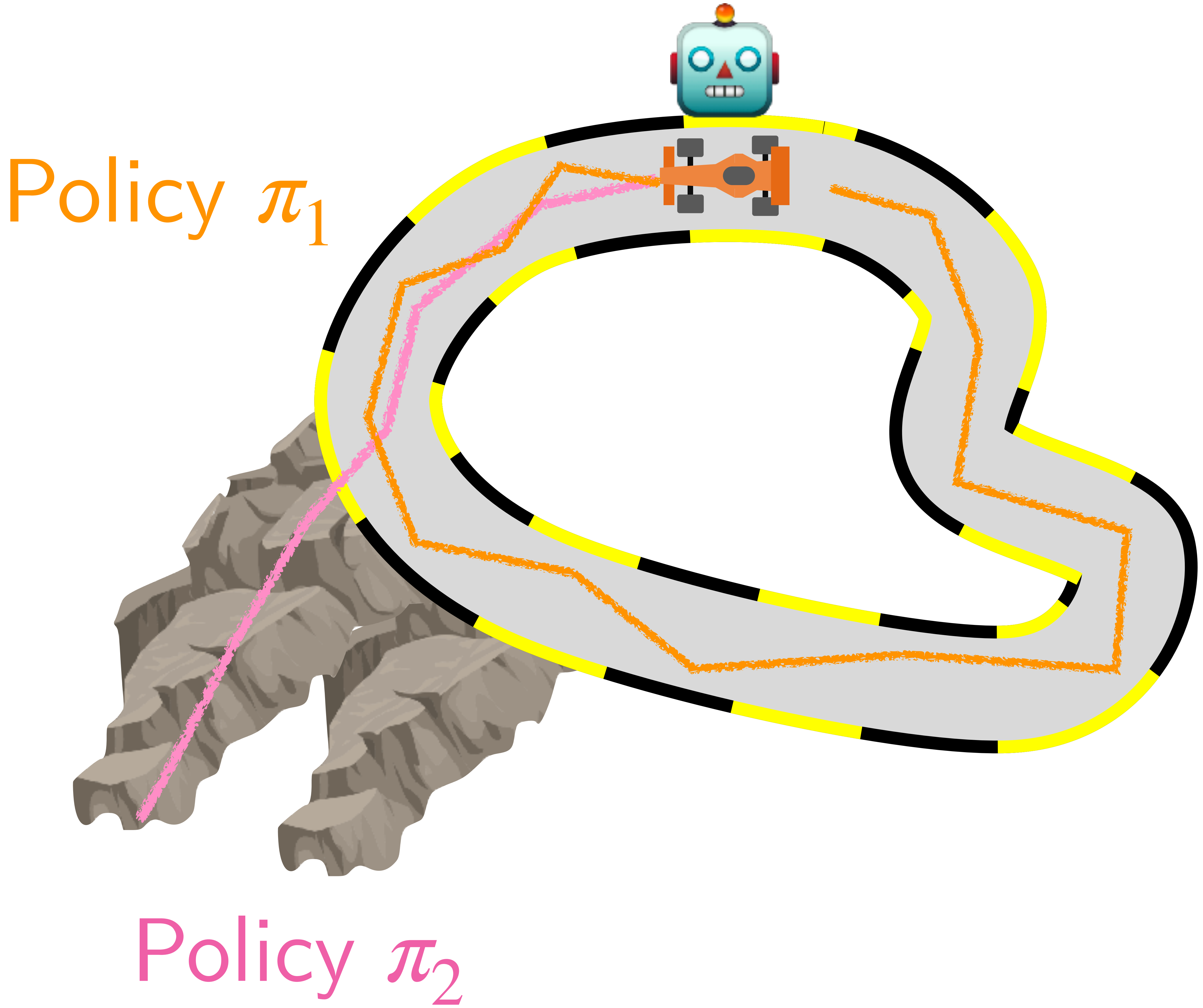
Policy π_2:

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?



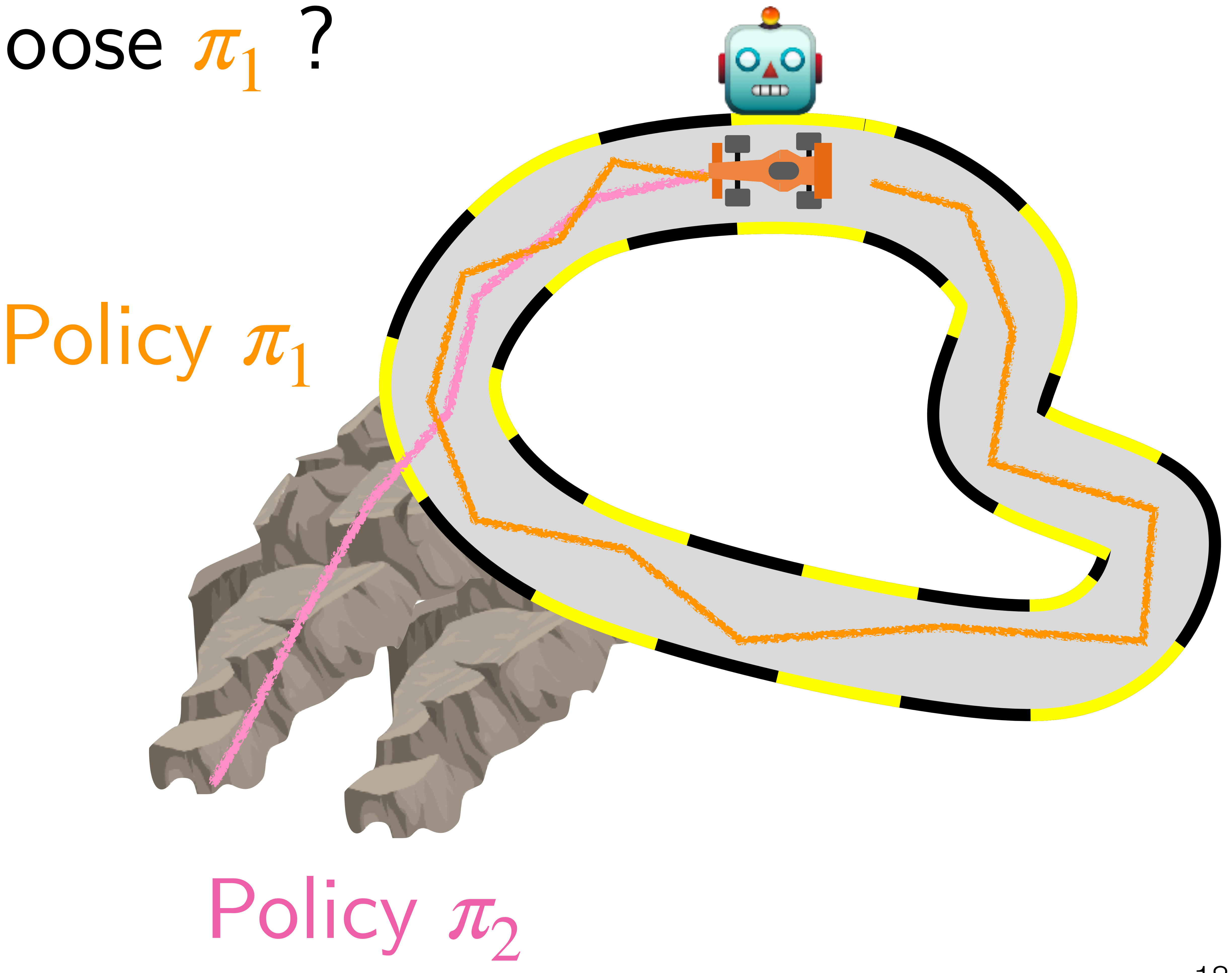
Activity!

Think-Pair-Share!

Think (30 sec): Which policy would DAGGER return? How would you get it to choose π_1 ?
Is DAGGER really $O(\epsilon T)$?

Pair: Find a partner

Share (45 sec): Partners exchange ideas



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: Expert provides values

Just like DAGGER

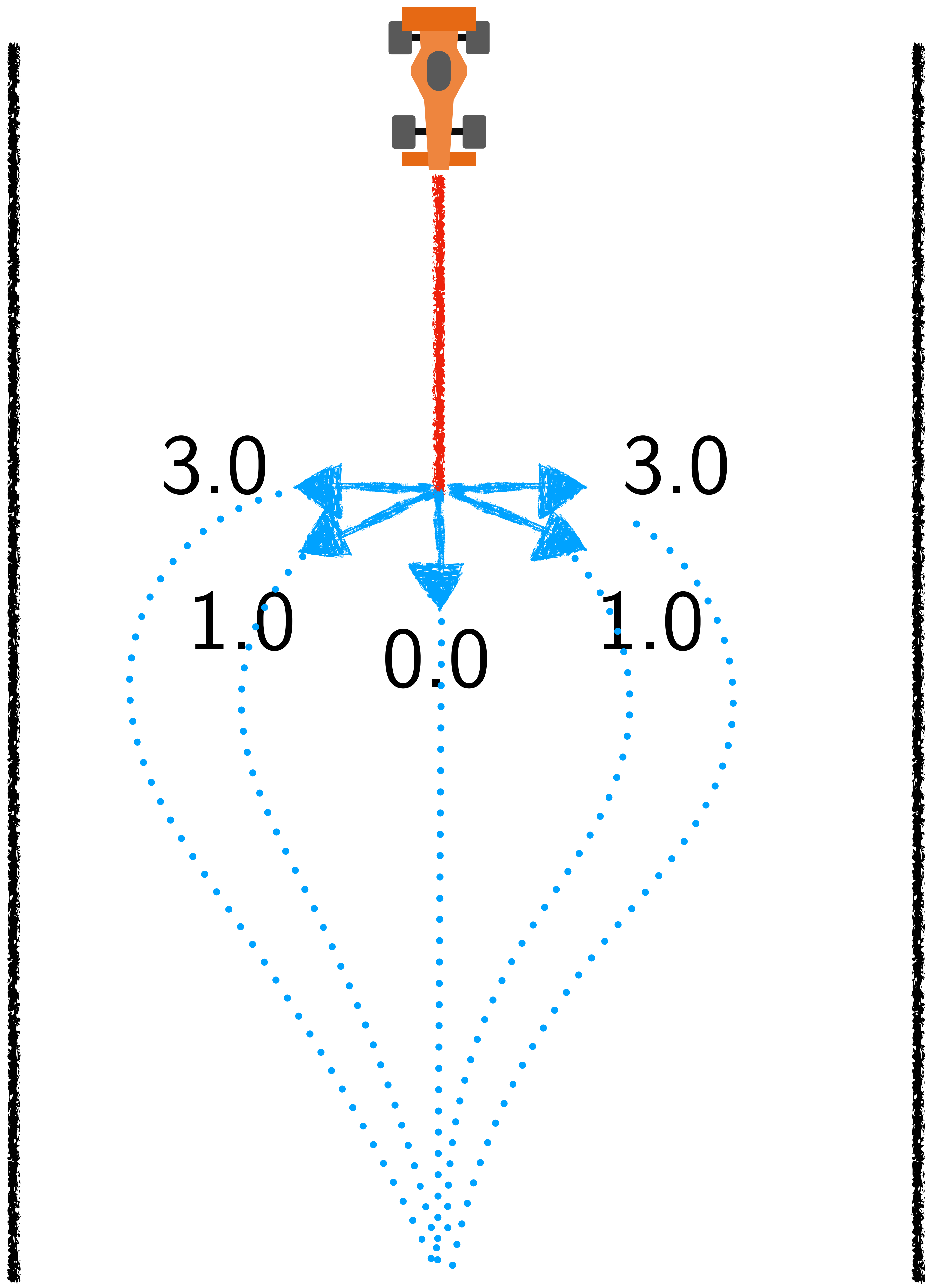
For i = 0 … N-1

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



AGGREVATE: Expert provides values

Just like DAGGER

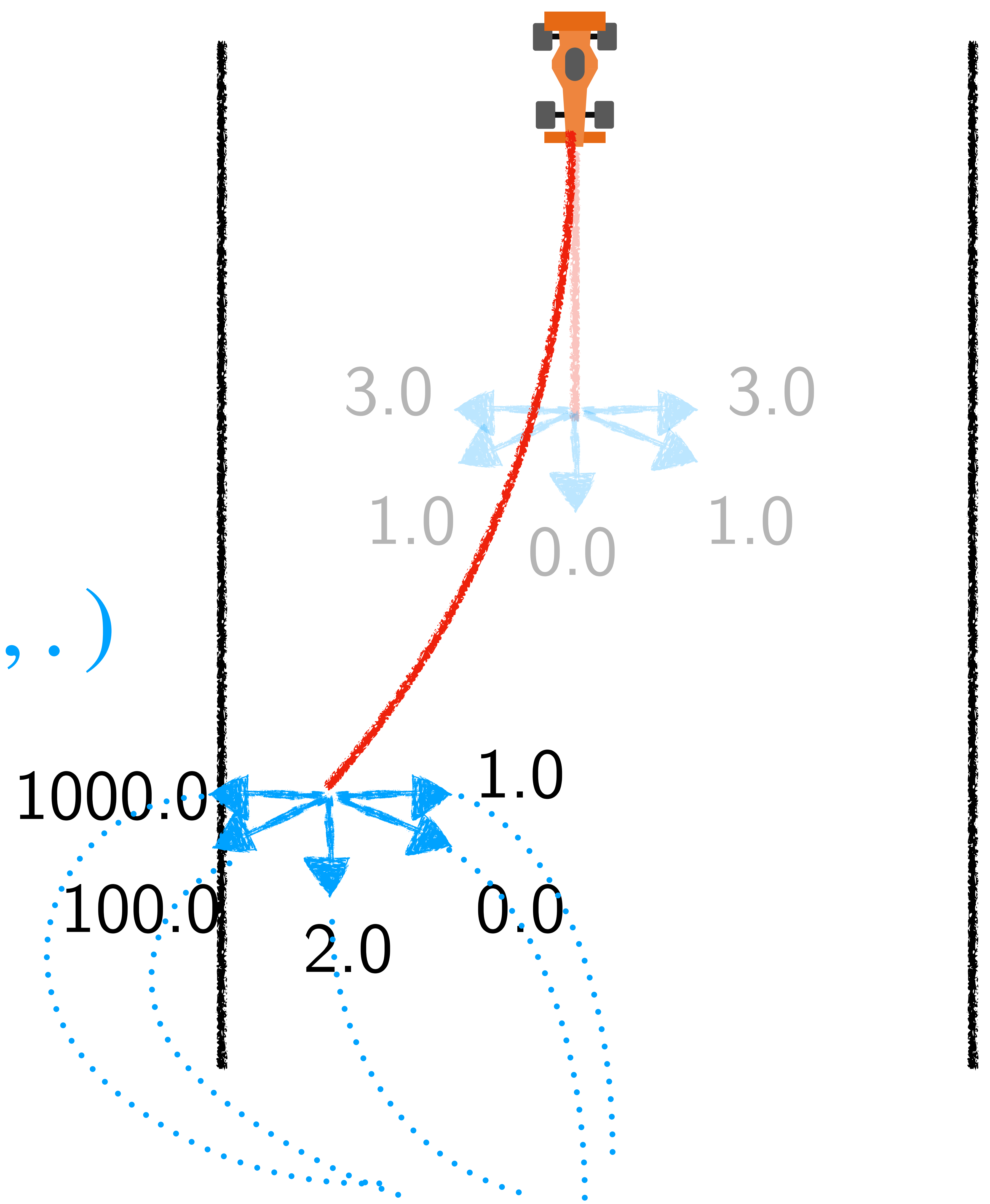
For i = 0 … N-1

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

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

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

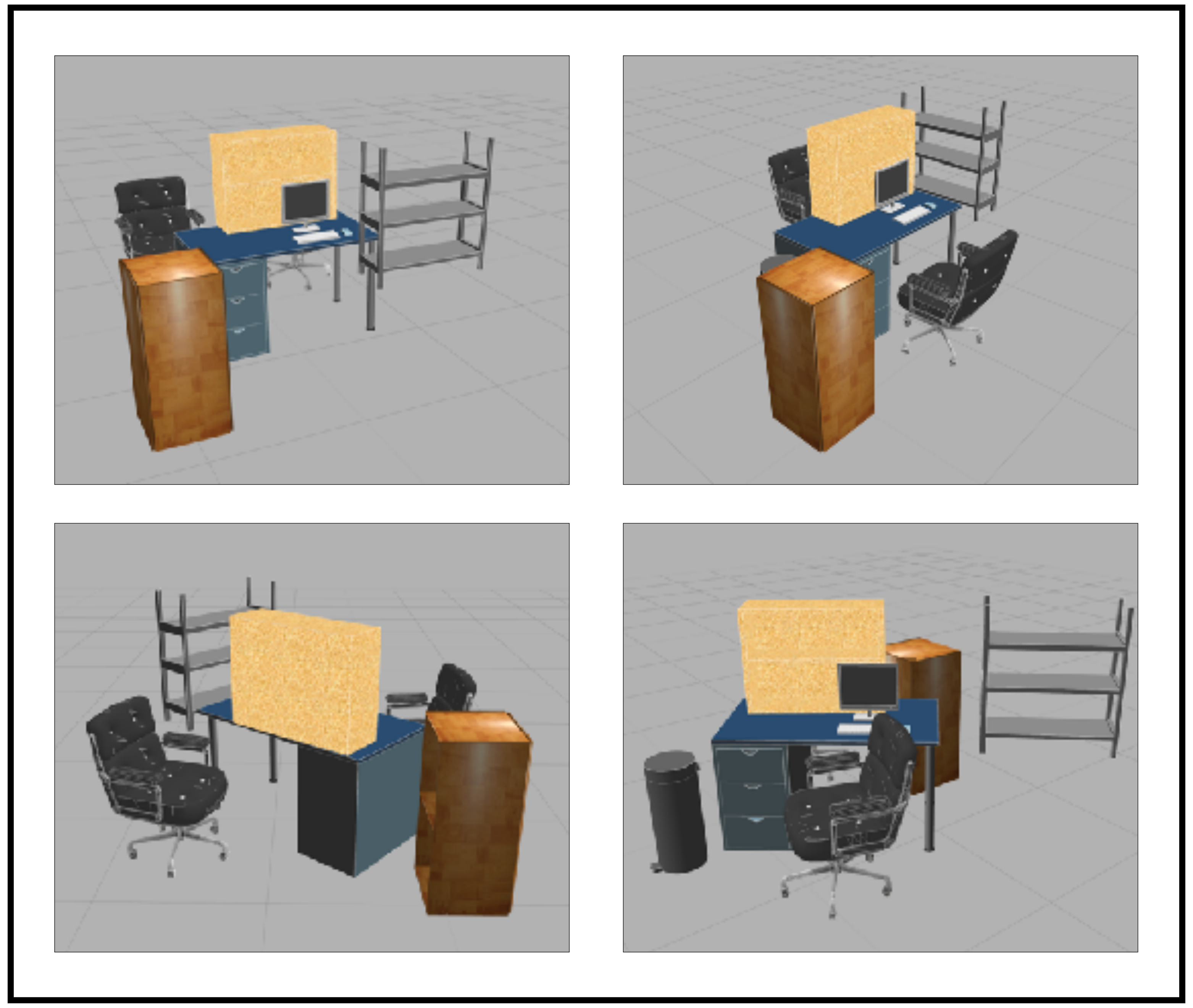
Train policy $\pi_{i+1} = \mathbb{E}_{s, A^* \sim \mathscr{D}}(A^*(s, \pi(s)))$



Is Aggrevate even practical?

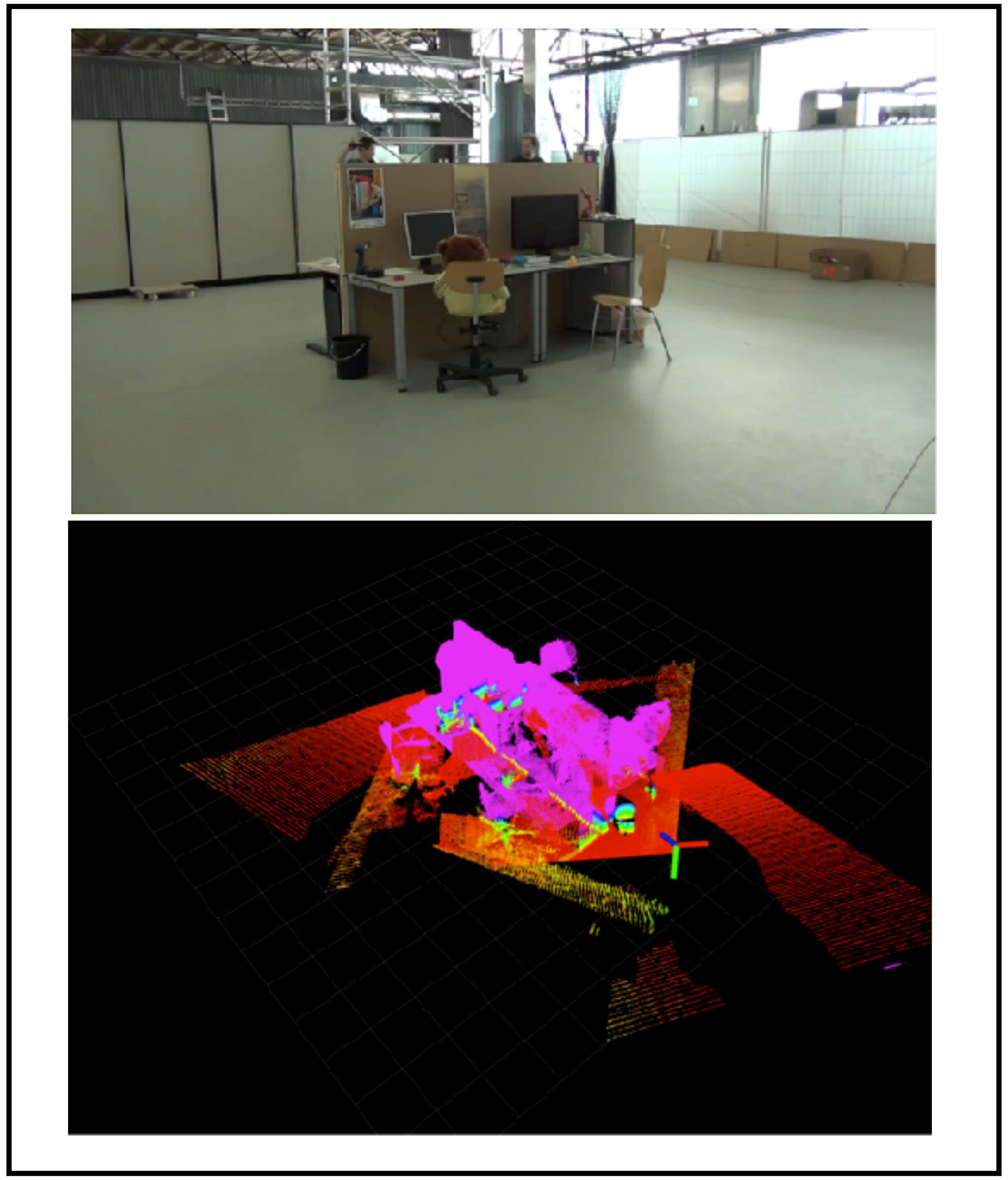
Yes*! When you are imitating algorithmic oracles

Train in Simulation



Learn Mapping Policy

Test in the real world



Choudhury, S. et al Data-driven planning via imitation learning. *IJRR'18*

Okay …

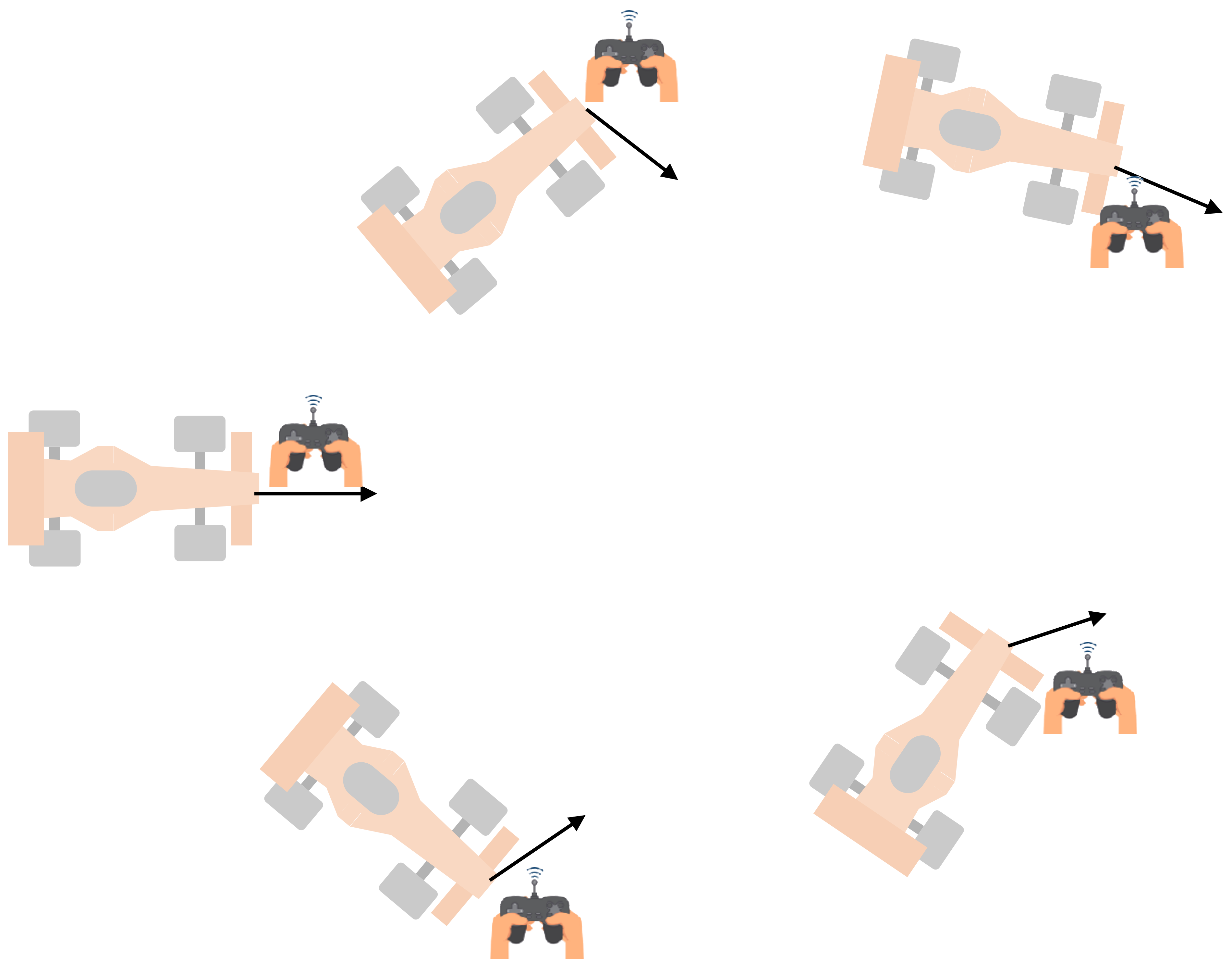
But how do we learn from natural human feedback?

Hidden Charge #2:

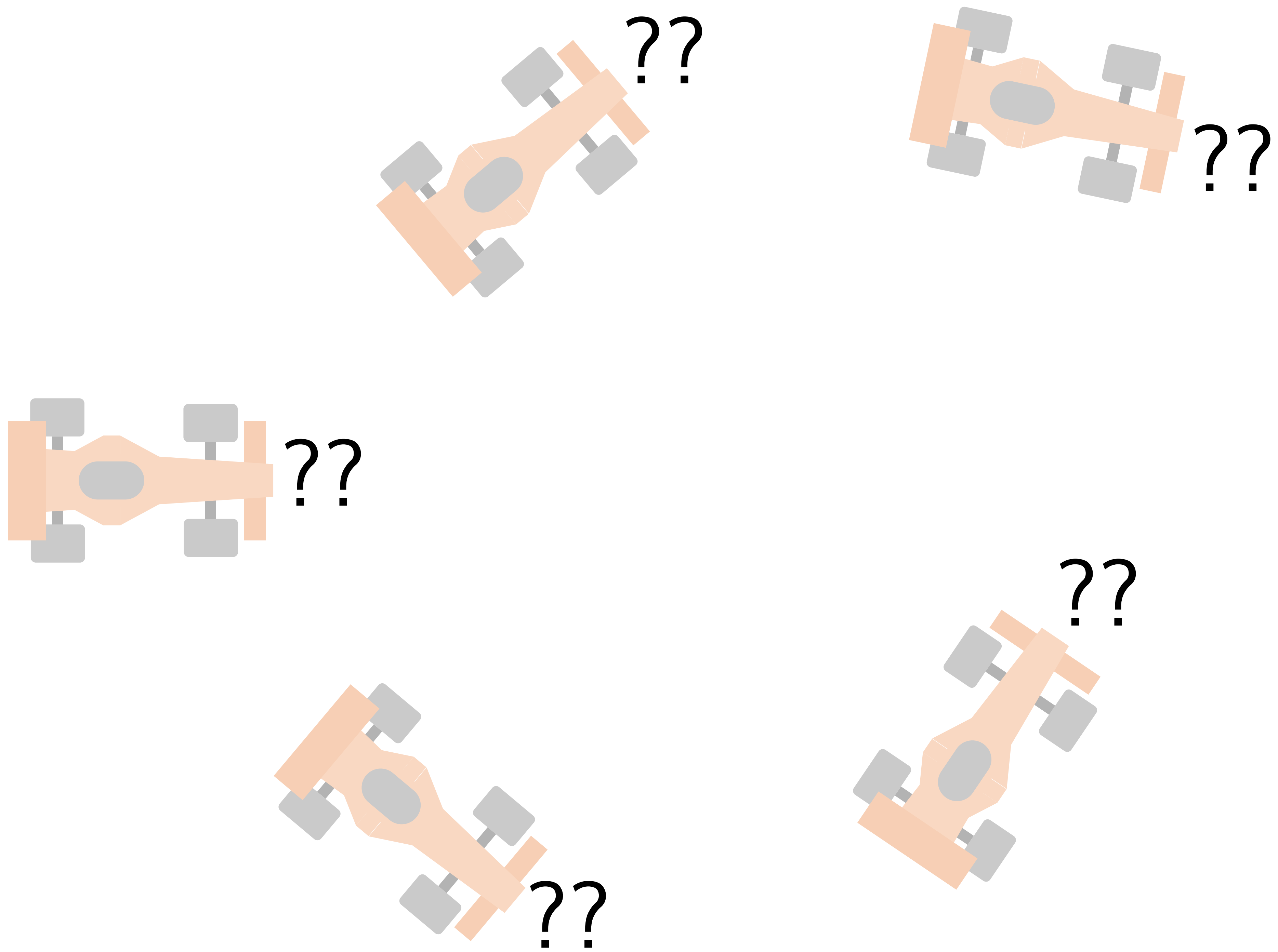
DAGGER queries the human at *every* state

Behavior Cloning crashes into a wall

DAGGER queries the human at every state!

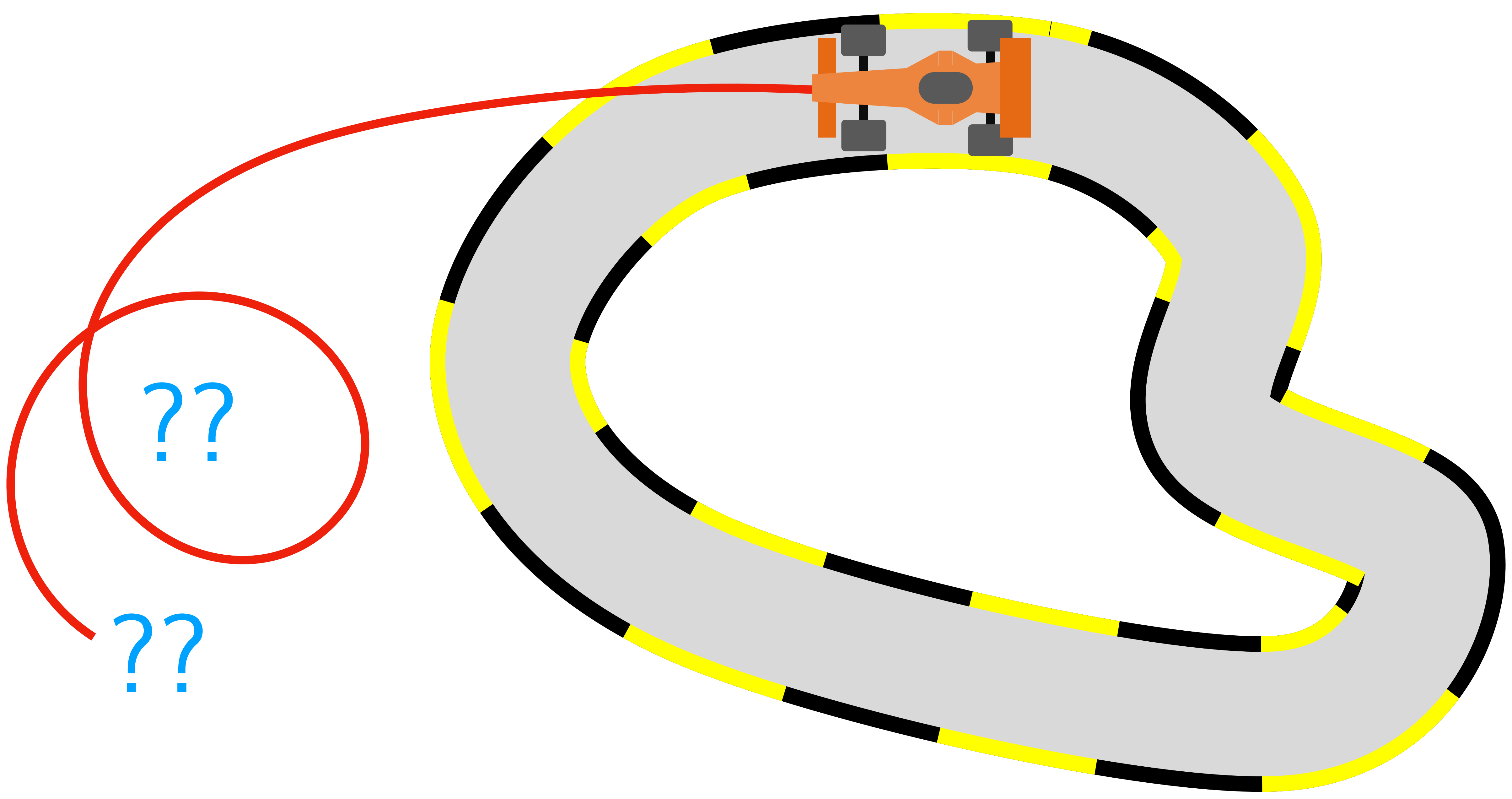


Impractical: Too much human effort!



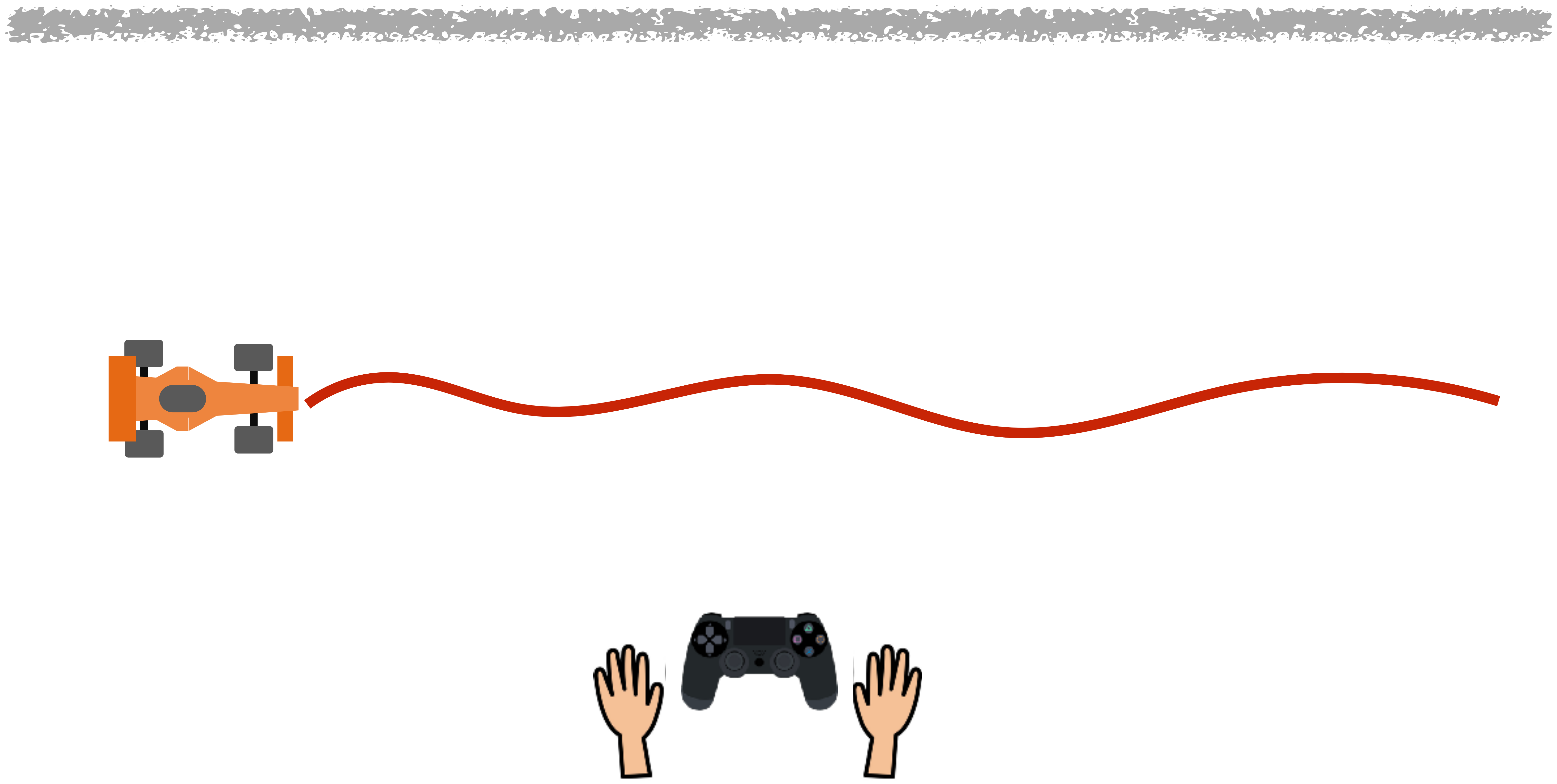
Can we learn from minimal human interaction?

Problem: Impractical to query expert everywhere



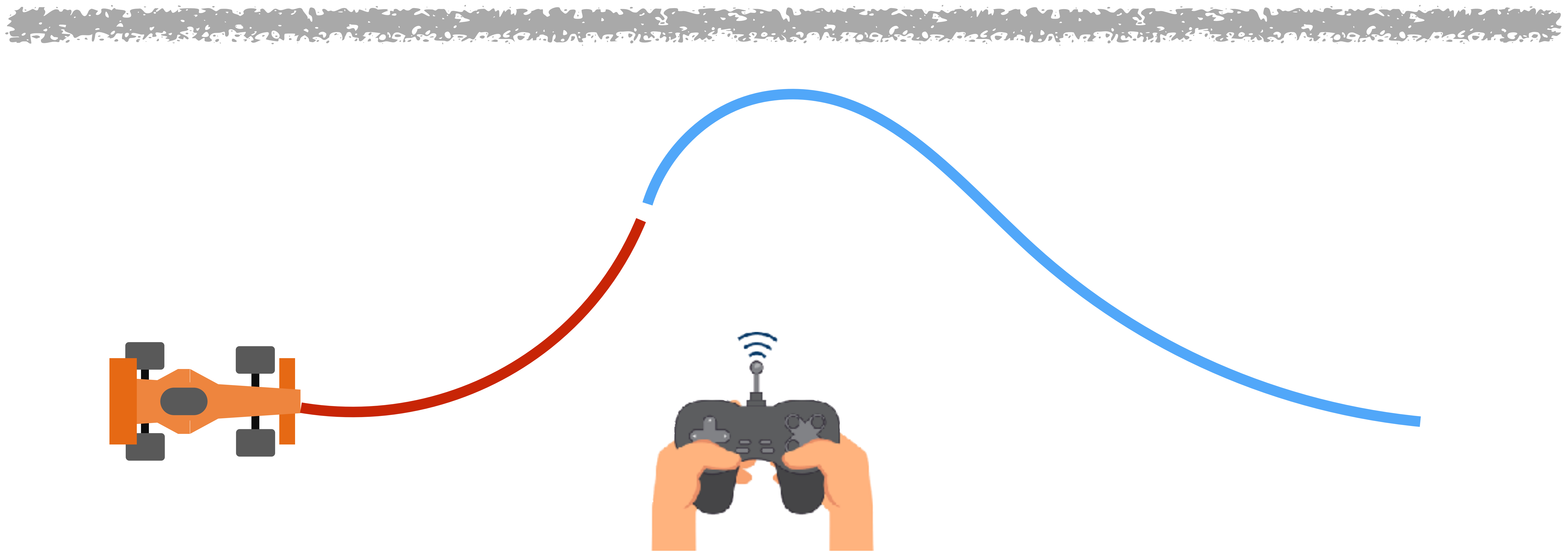
Can we learn from natural human interaction, e.g., interventions?

Learn from natural human interventions?



Hands free, no corrections!

Learn from natural human interventions?



Take over and drive back!

HG-DAGGER: Learning from interventions

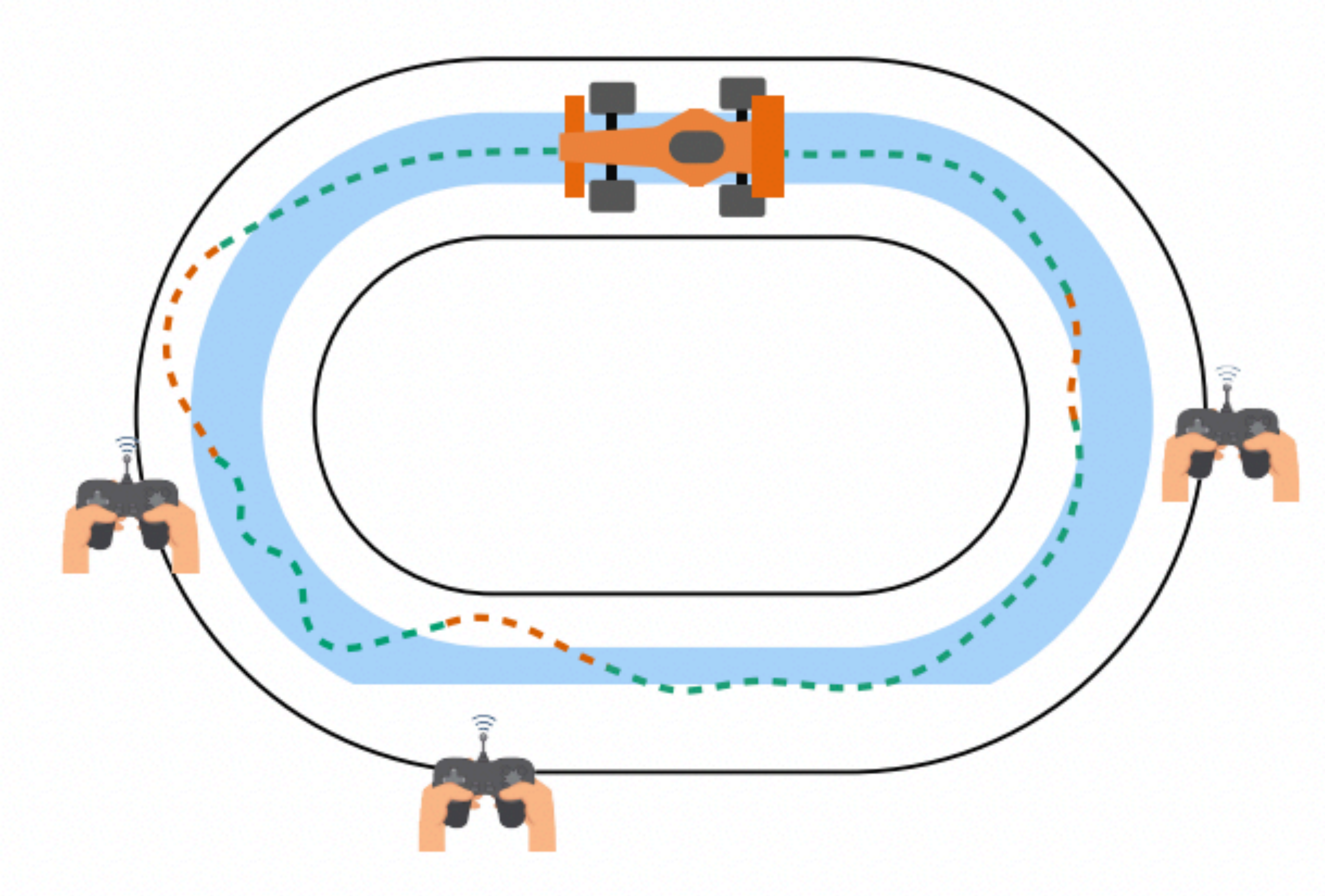
Roll out a learner policy

Collect expert actions on states where expert intervened

Aggregate data

Update policy

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



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 ...

... and must be inferred from human interventions

Expert Intervention Learning (EIL)

[SCB+ RSS'20]

Interventions are just constraints on latent action-value function

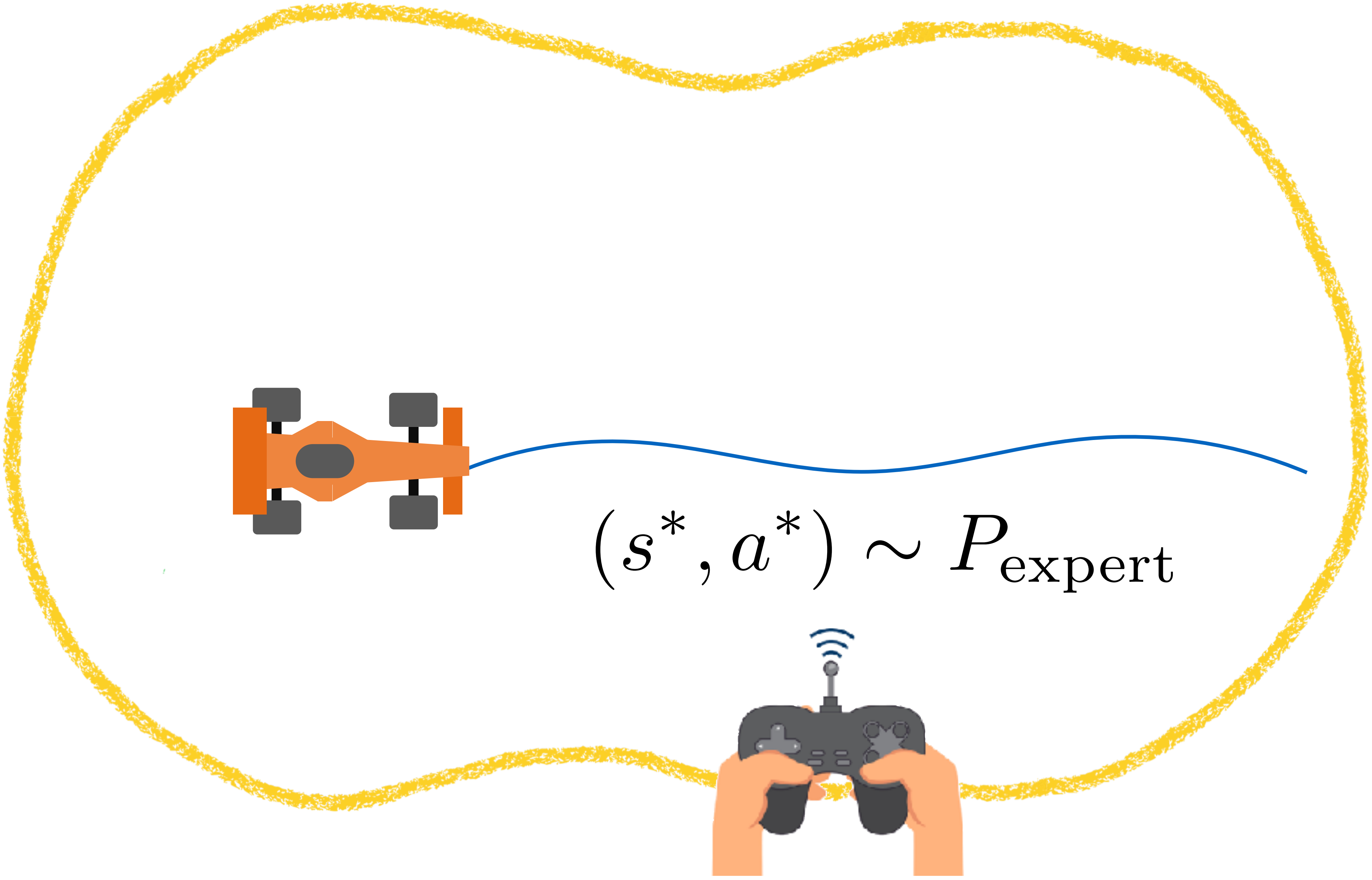
Expert Intervention Learning (EIL)

[SCB+ RSS'20]

Interventions are just constraints on latent action-value function

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

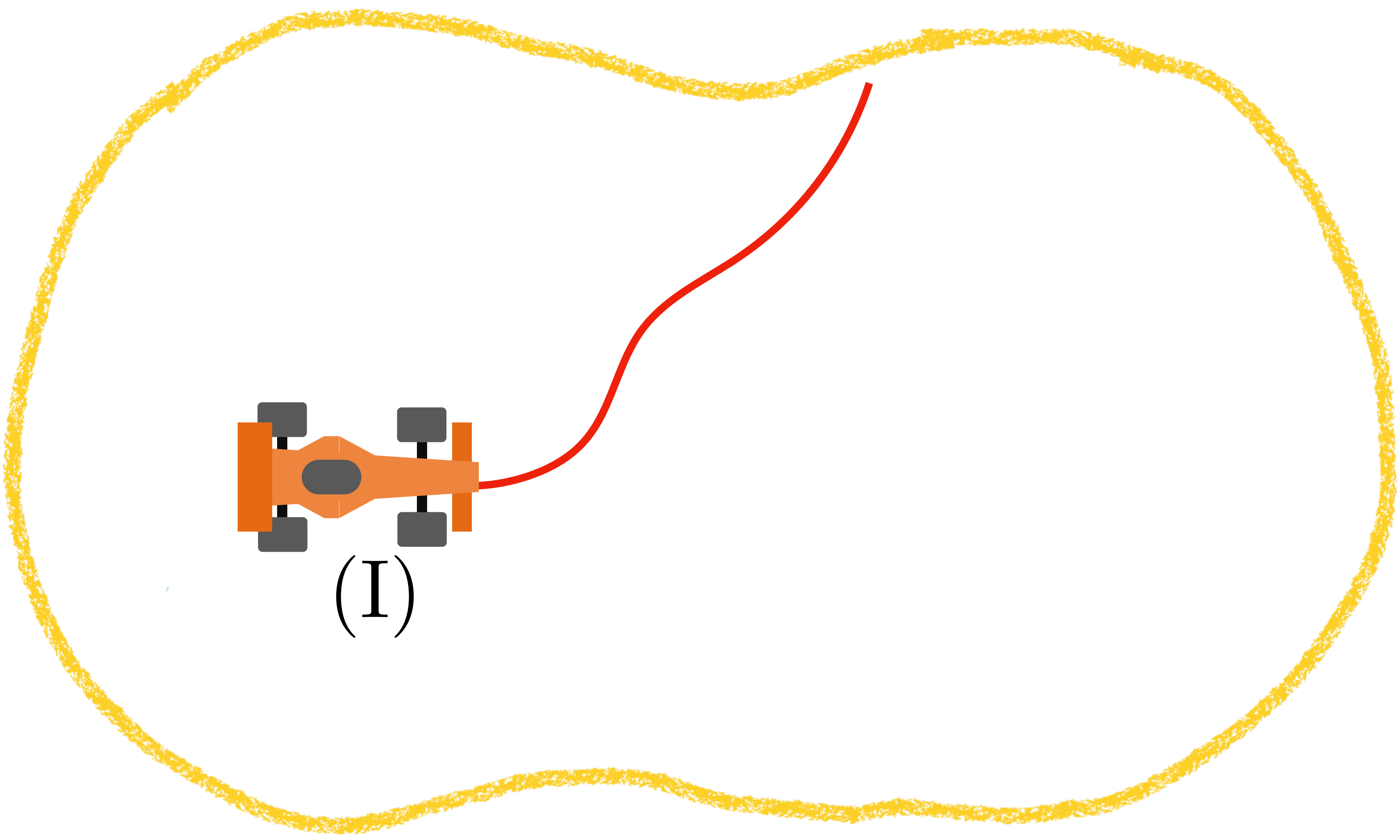
classify demonstrations



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.

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

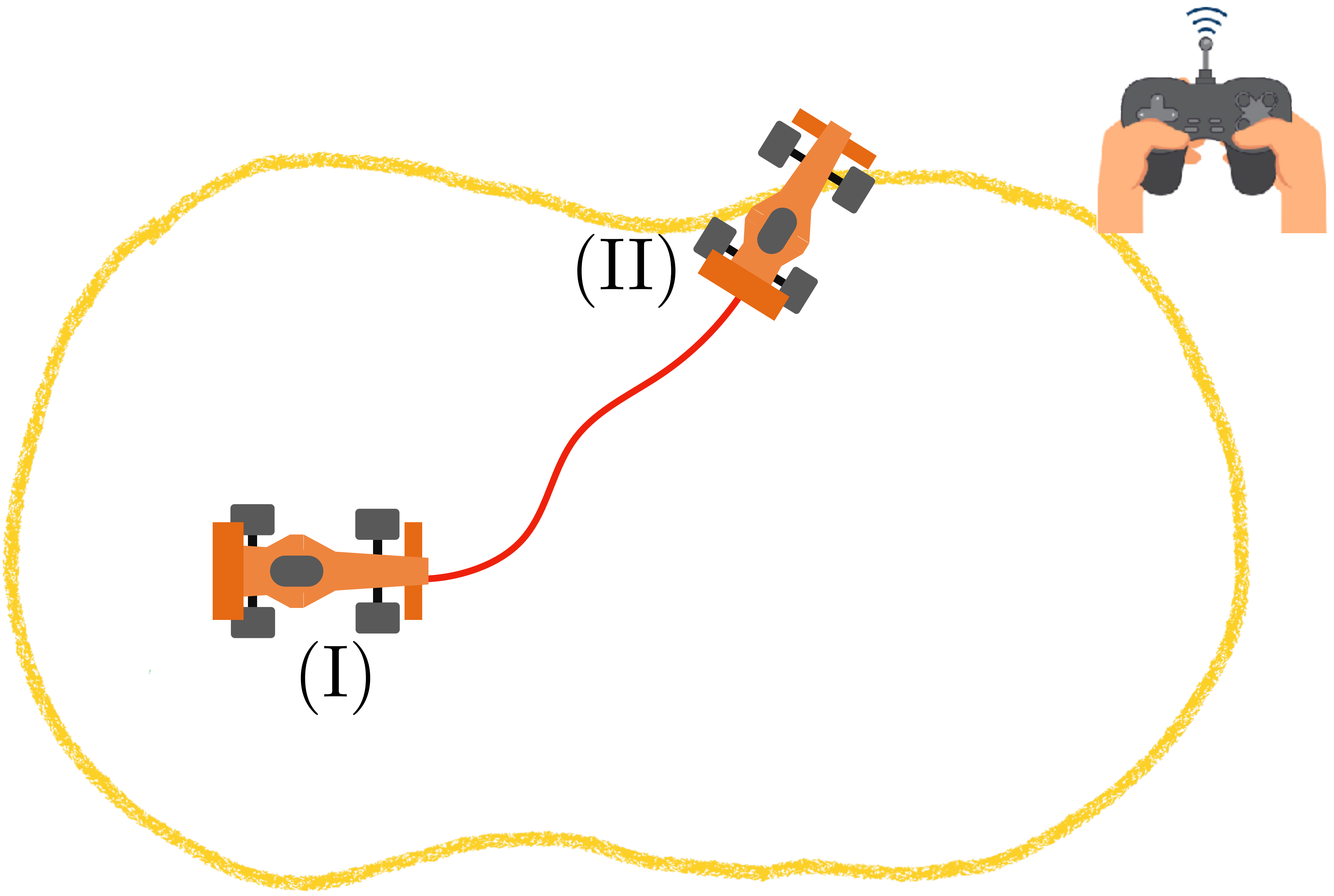
$\forall (s,a) \in (\mathrm{I})$

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.

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

before expert intervenes

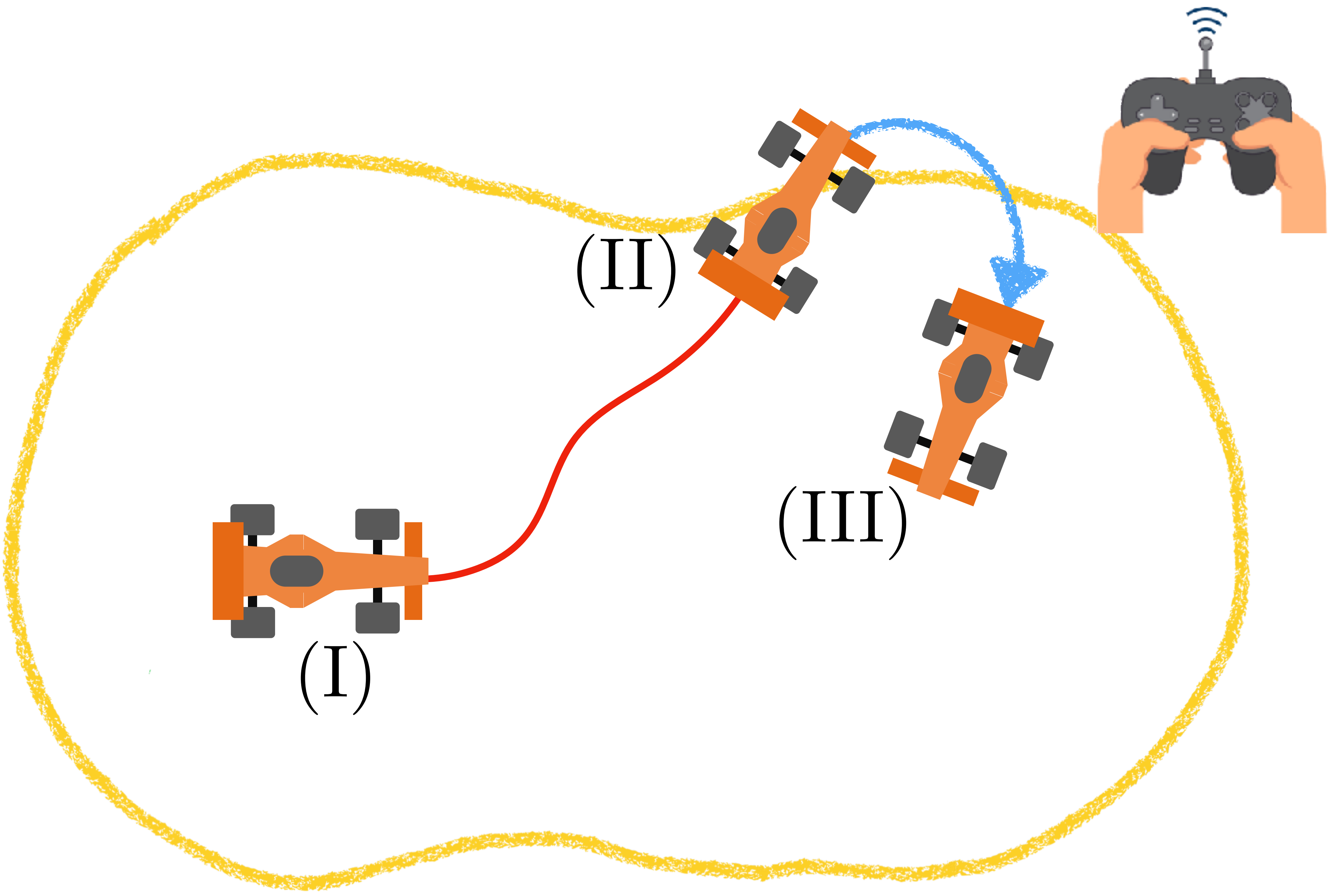
$Q(s, a) \geq \delta_{\text{good}}$ $\quad \forall (s, a) \in (\mathrm{II})$

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.

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

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

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

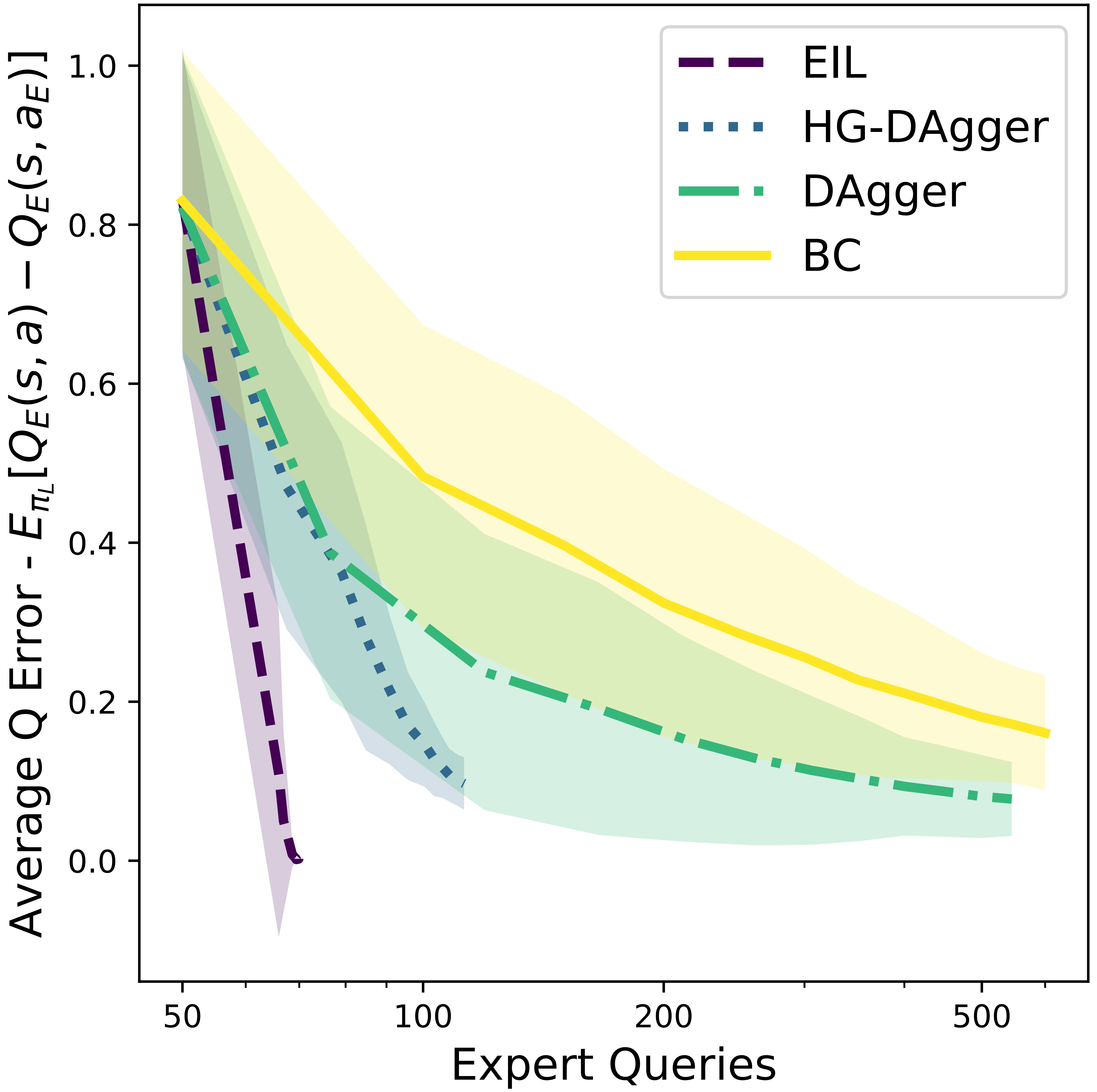
$\forall (s,a) \in (\text{II})$
after expert intervenes

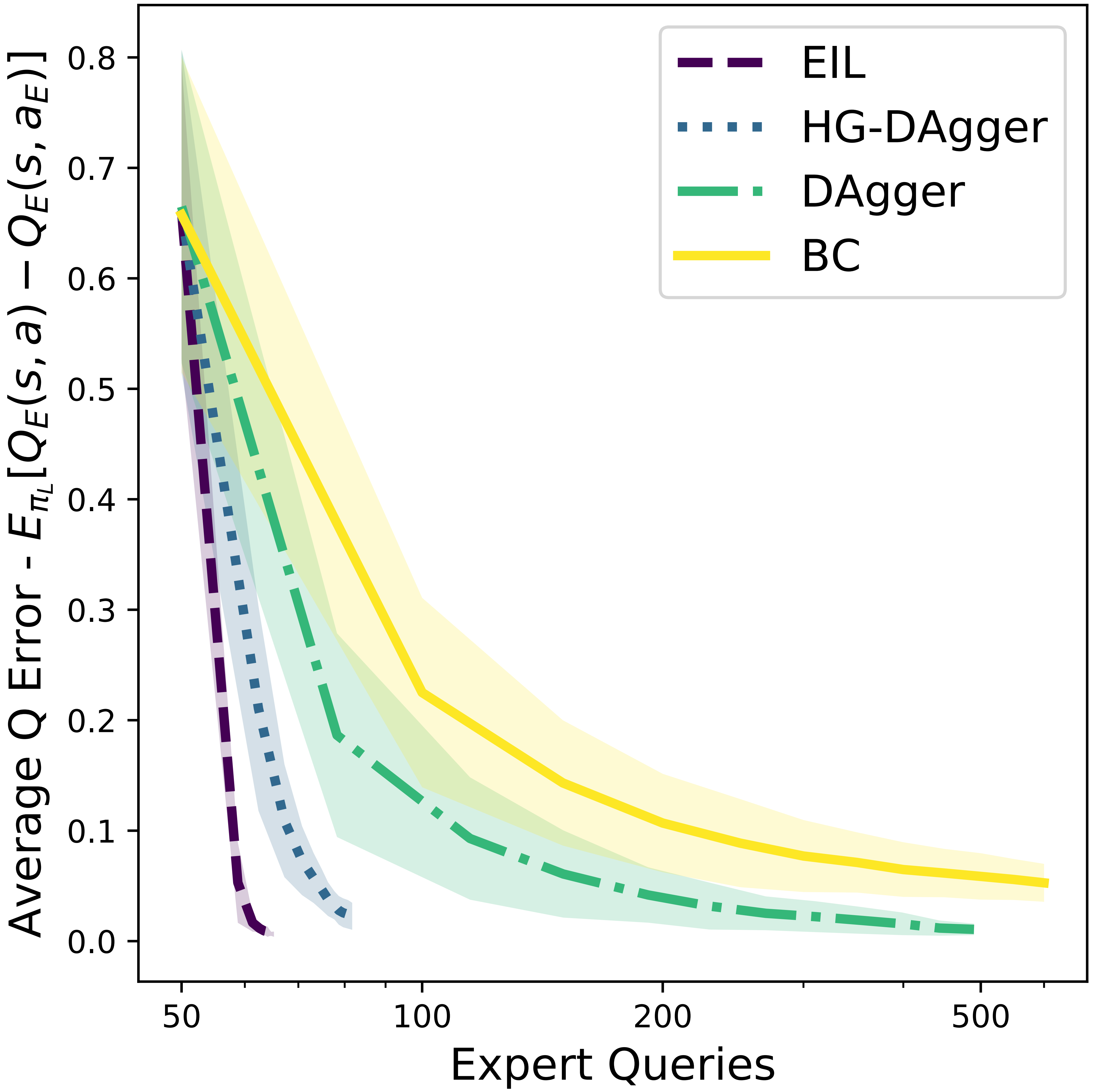
$$Q(s,a) \leq \min_{a'} Q(s,a)$$

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

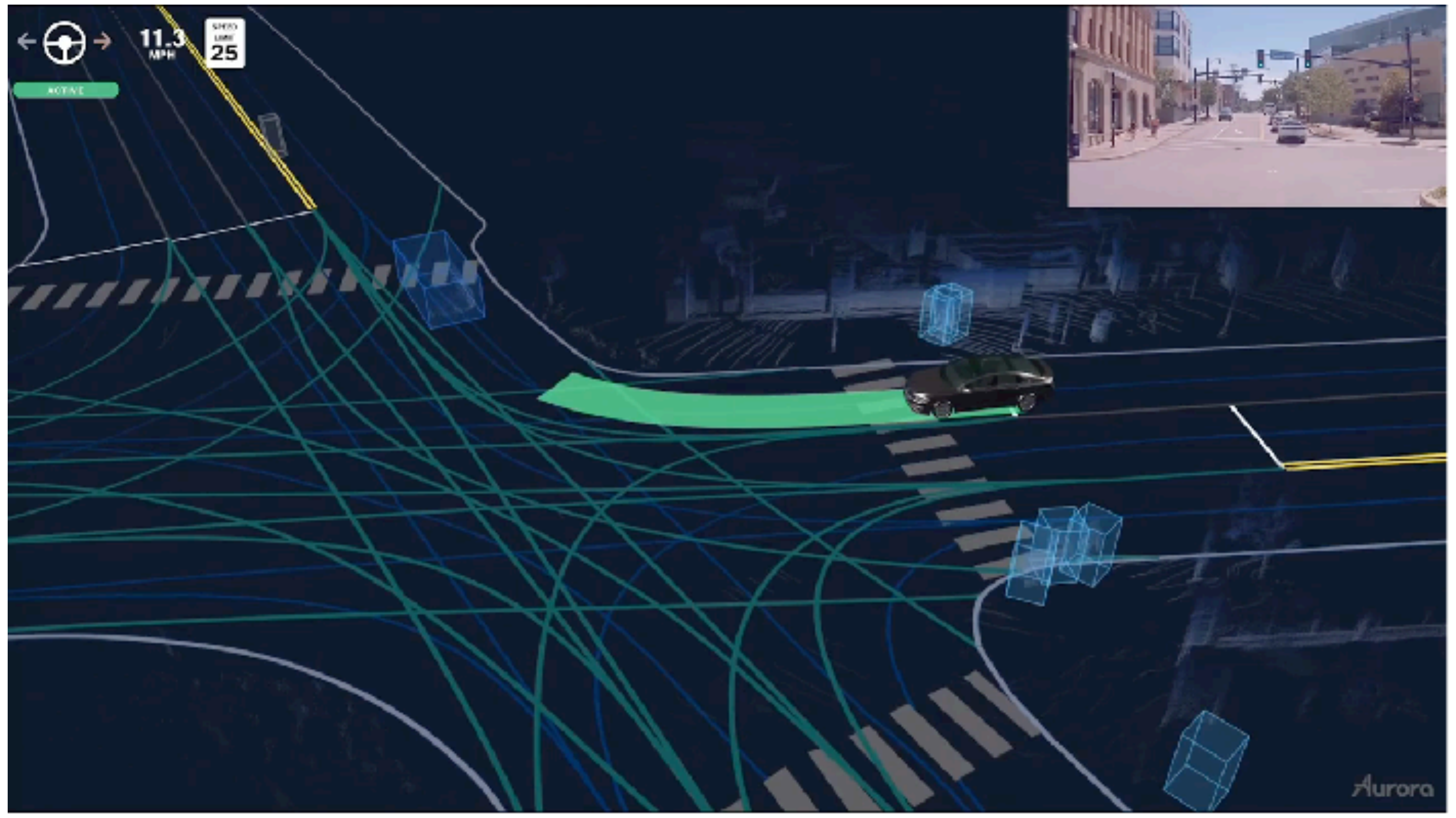
Reduce to online, convex optimization! $O(\epsilon T)$

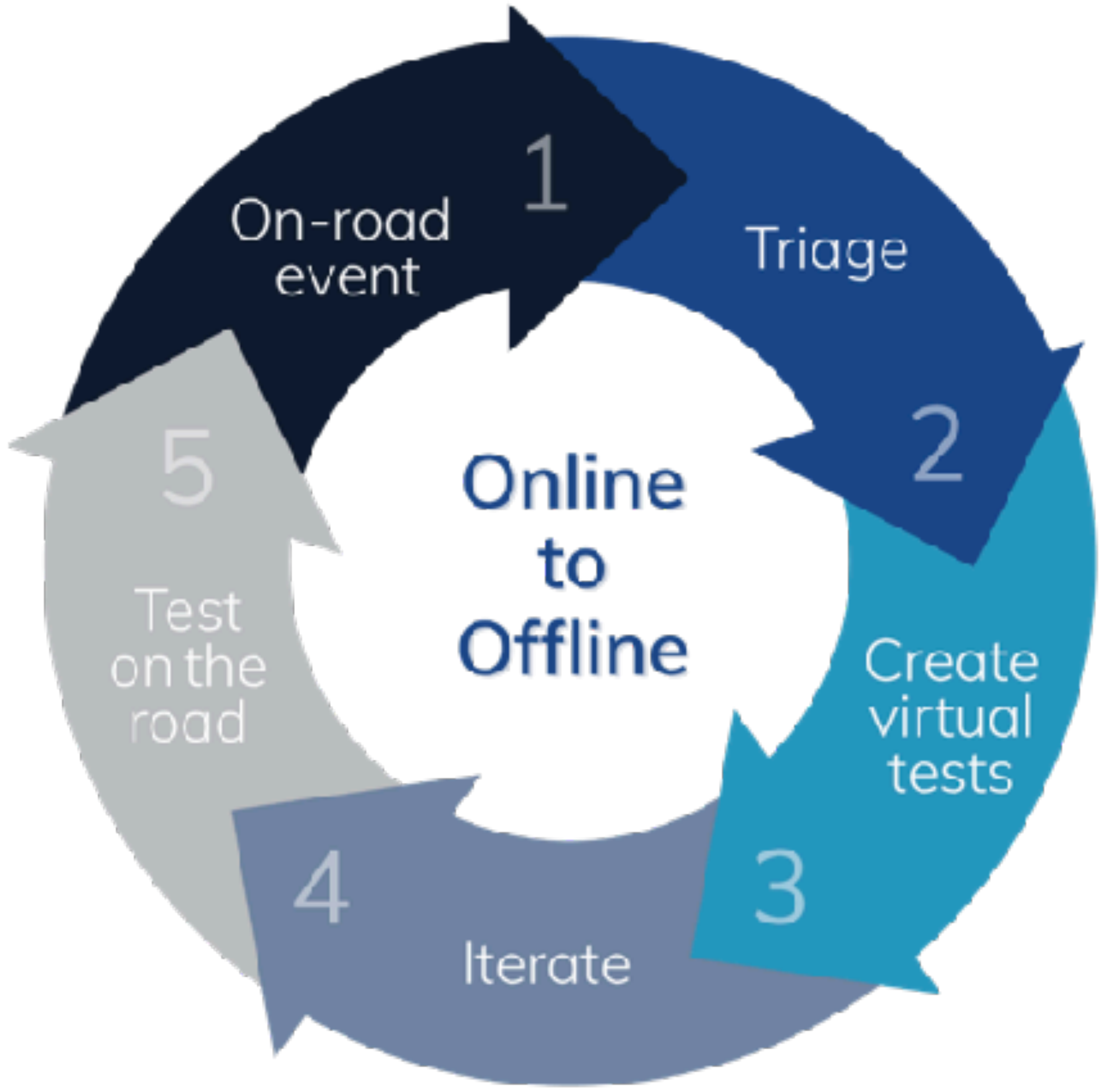
EIL drives down error with less expert query

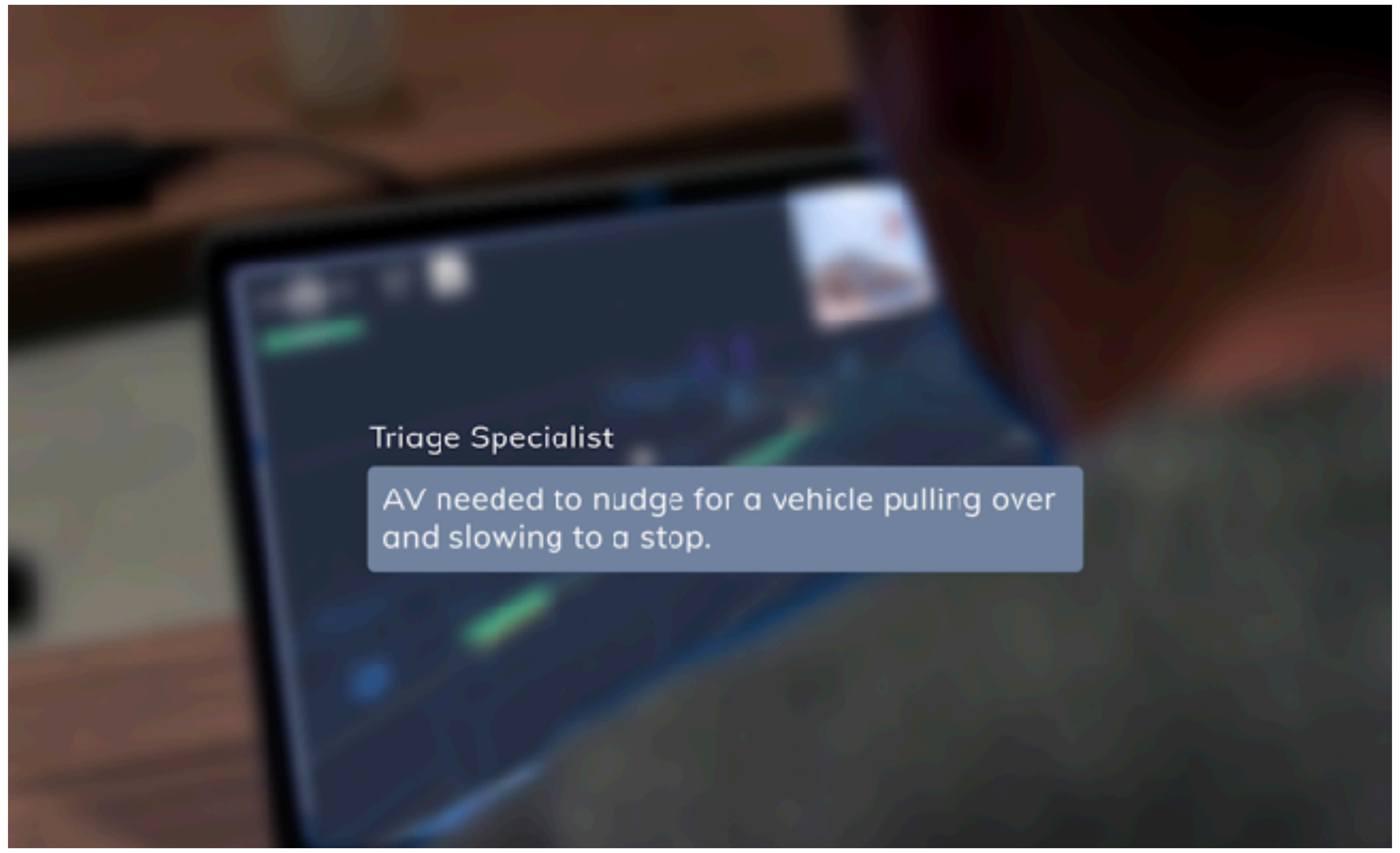


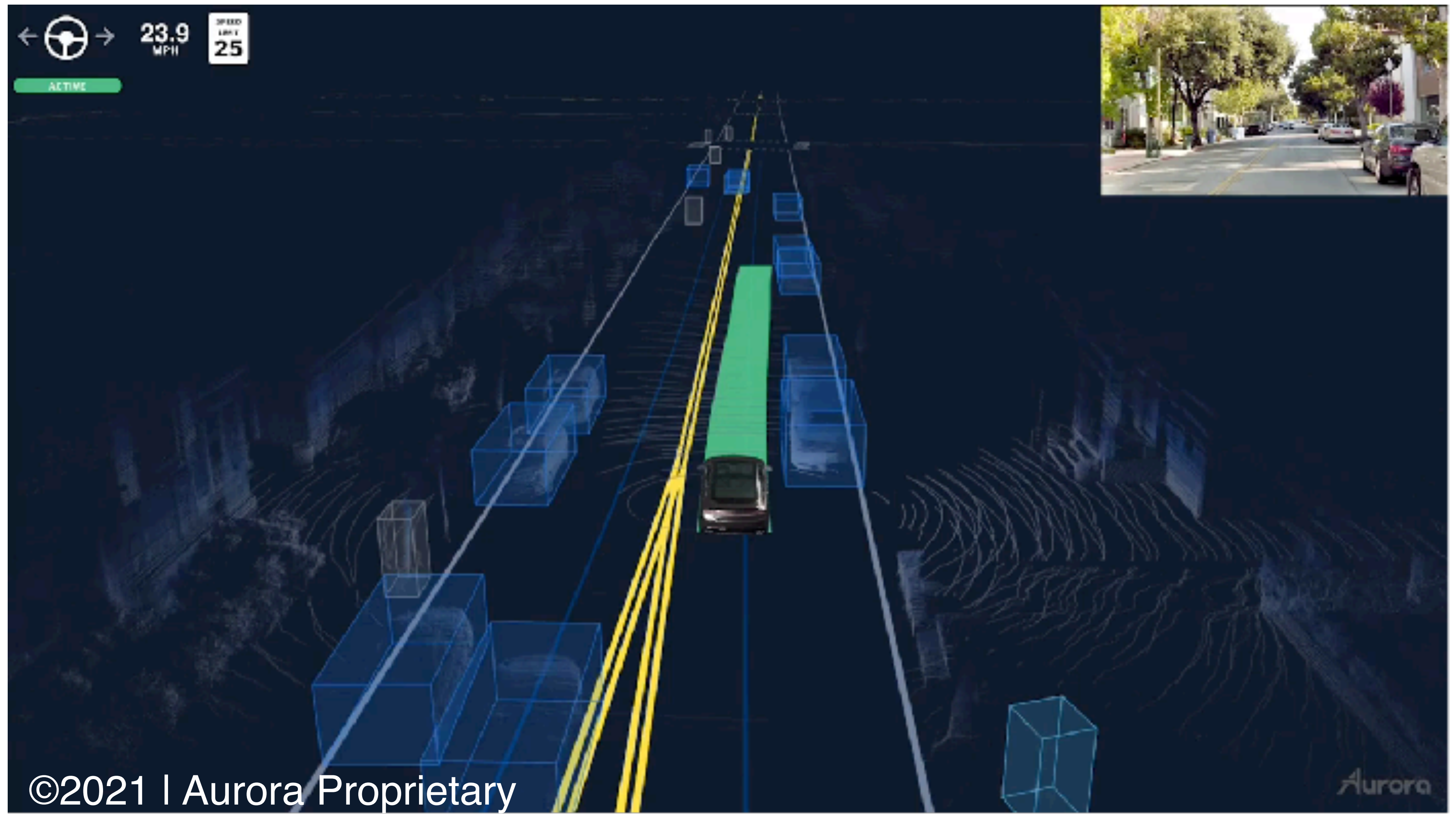


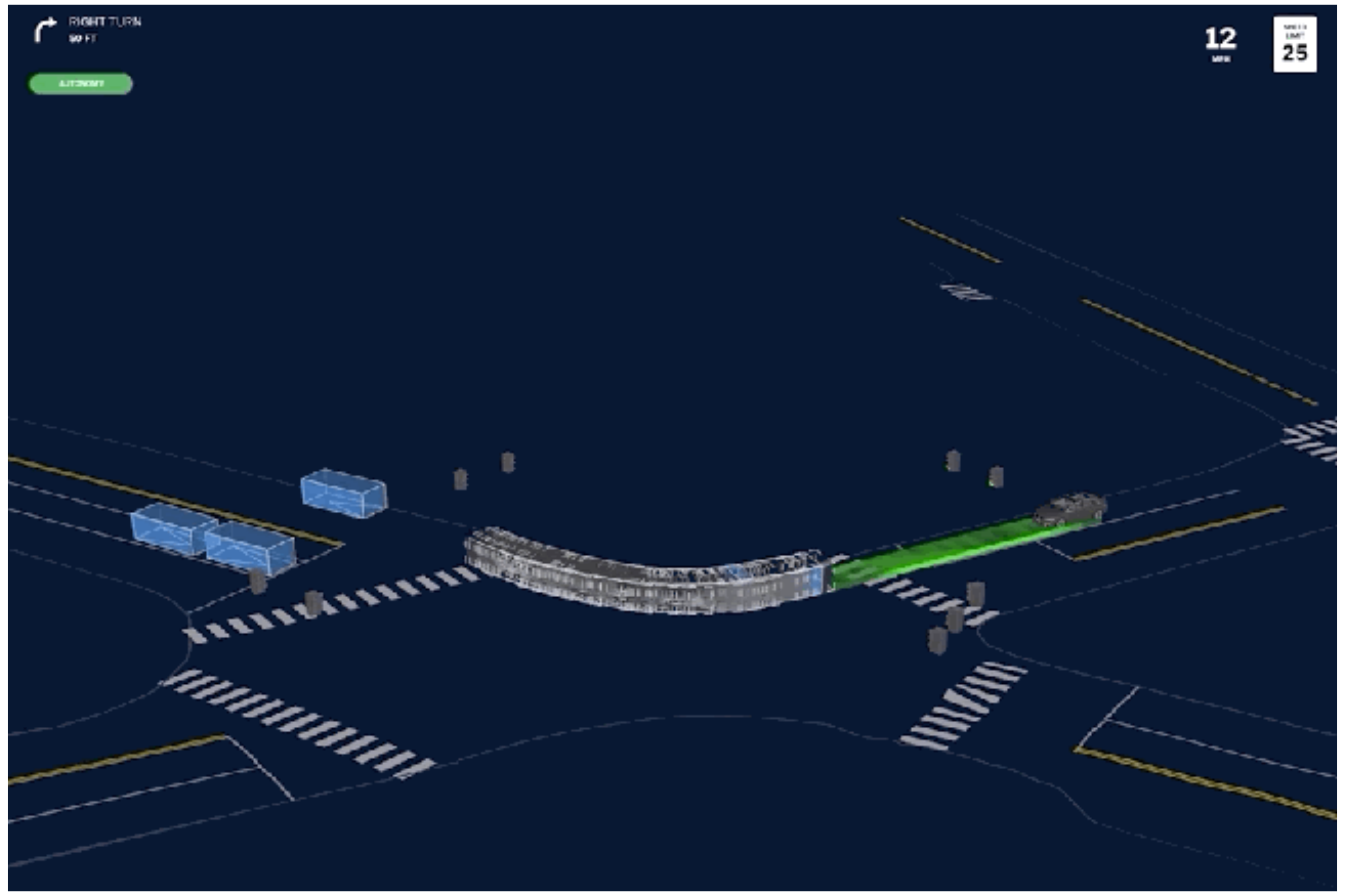
Turning interventions to simulations for learner











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!

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

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!

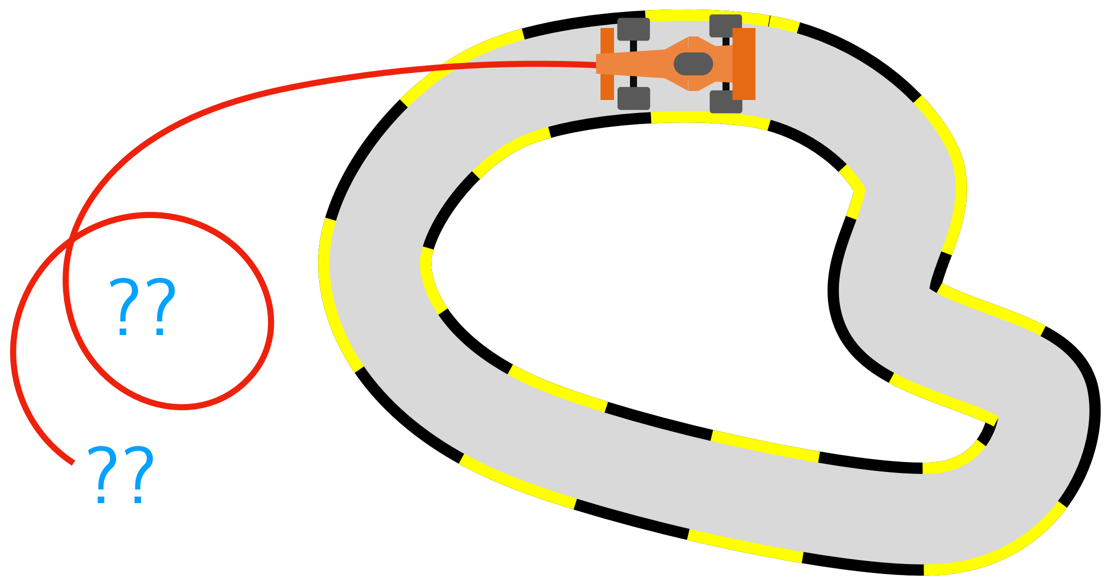
✓ Loss

"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)

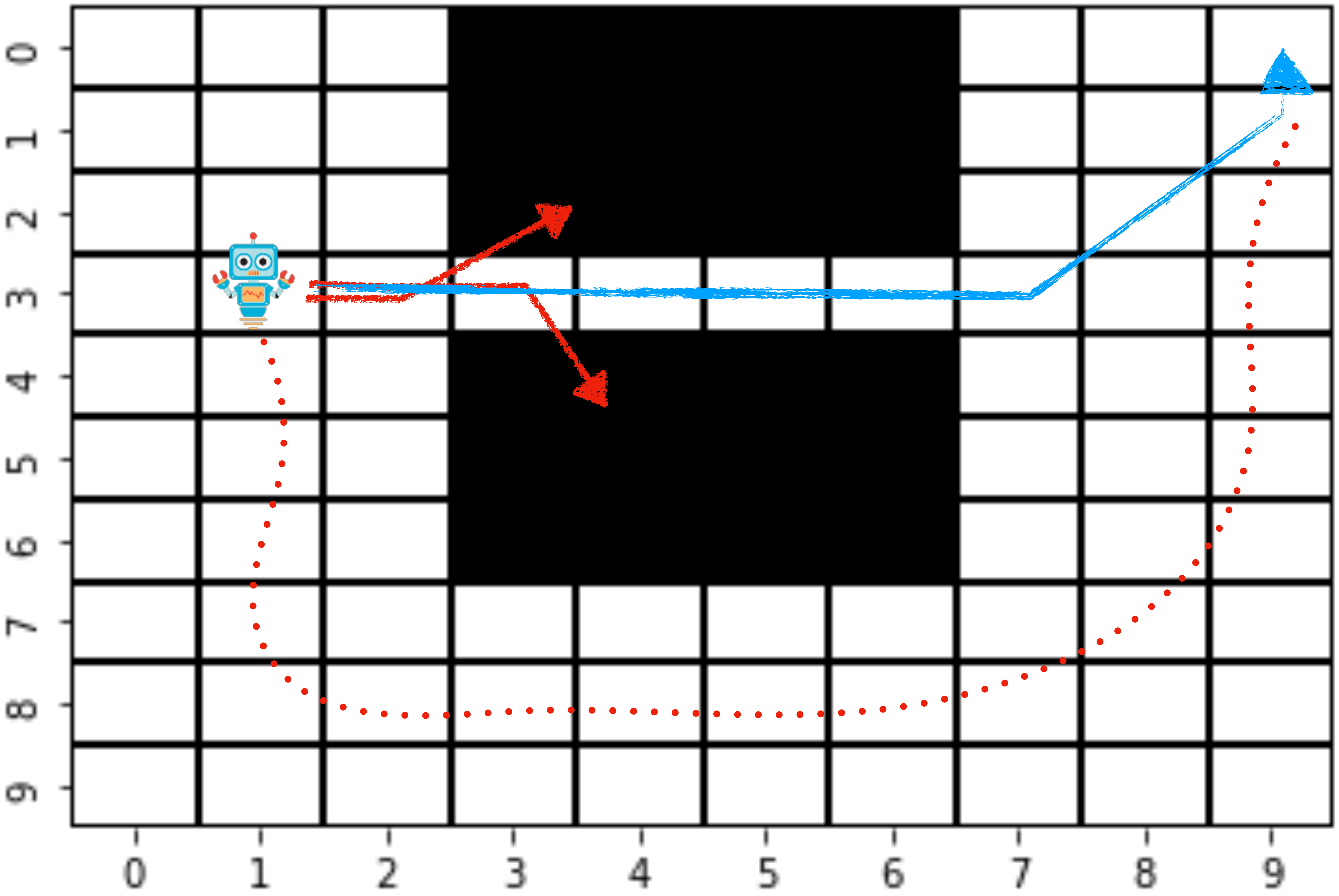
[SCB+ RSS'20]

The expert action-value function is latent ...

... and must be inferred from human interventions

x

Hidden charge #3: Dagger expects at least one policy to be good *everywhere*



Learner simply can't cross the bridge! ...

... but can take the long way round.