A Human-Like Agent Based on a Hybrid of Reinforcement and Imitation Learning

2019-07-17 Rousslan F. J. Dossa – Kobe University, Japan – IJCNN 2019

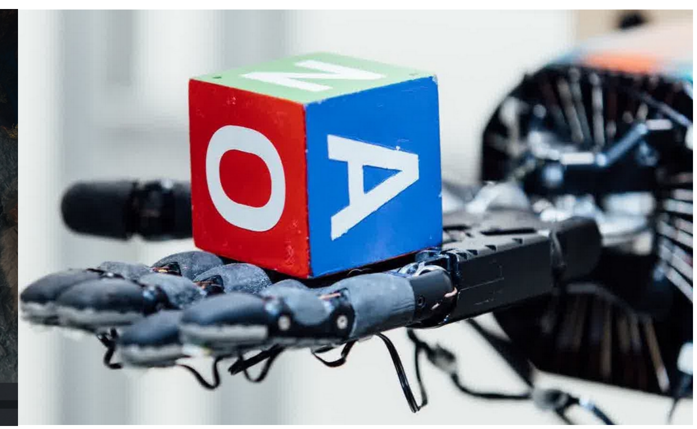
Reinforcement Learning

Pros

- High performance on the task

Cons

- The agent's behavior is likely to be uncanny



- DeepMind's "Alpha Go" (2015) and "AlphaStar" (2019) beat the human experts in the "GO Game" and "Starcraft 2", respectively
- OpenAI's "Dactyl Project" uses RL methods to achieve object manipulation with a robot arm

Background

Imitation Learning

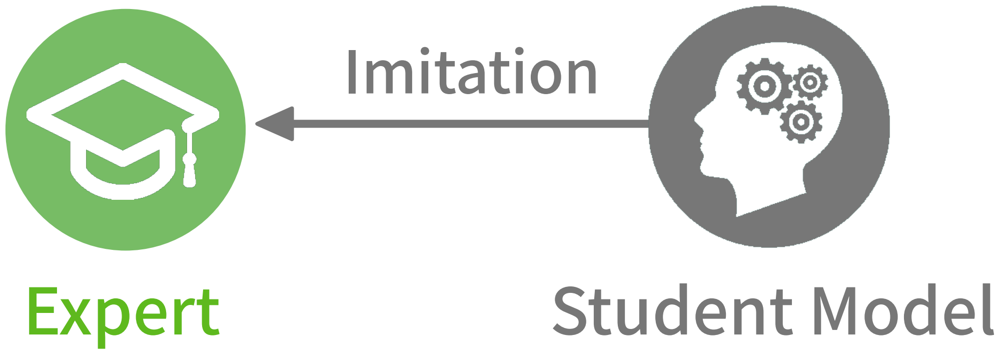
Data provided by an expert is used to training a mimicking agent in a **Supervised Learning** fashion

Pros

- With a human expert, exhibits a relatively **human-like behavior.**

Cons

- The agent's performance is limited to the expert's



Goal

An agent which exhibits high performance
while maintaining a human-like behavior
(based on both Reinforcement and Imitation Learning)

Reinforcement Learning

Uncanny behavior

High Performance

Imitation Learning

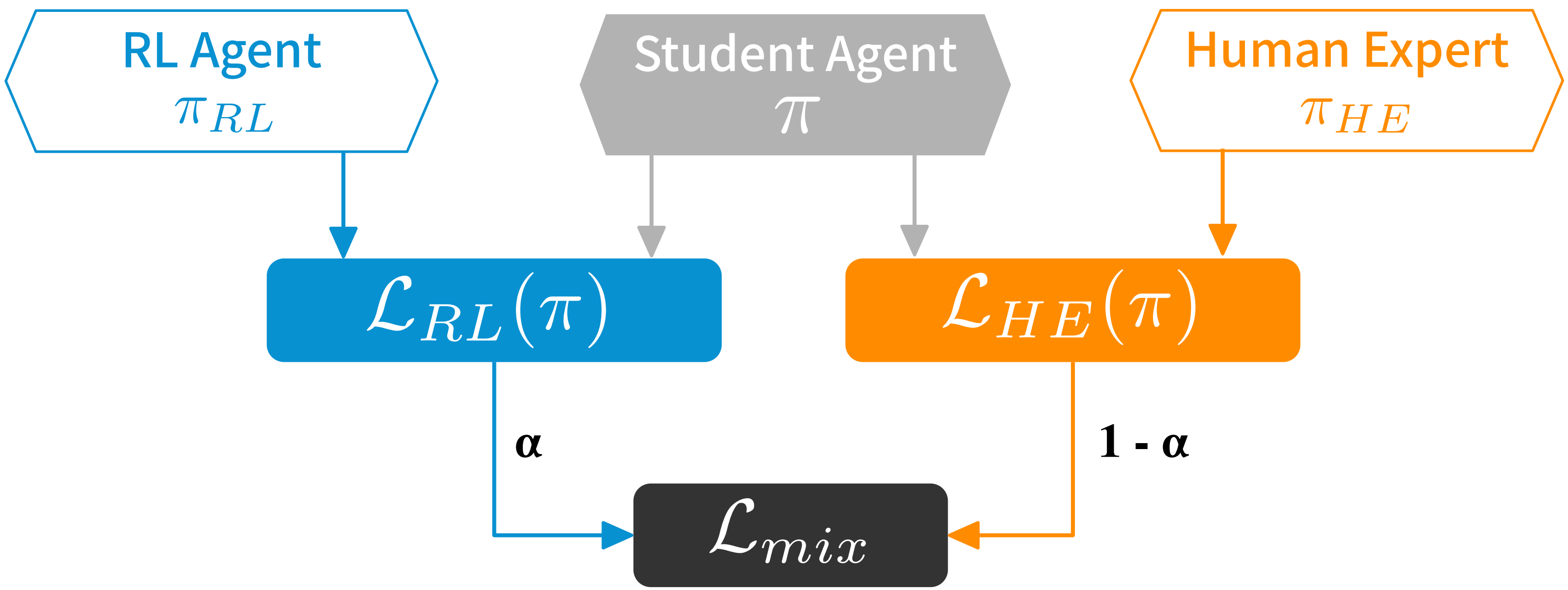
Moderate Performance

Human likeliness

Hybrid Agent

Proposed Method

Hybrid Loss Function

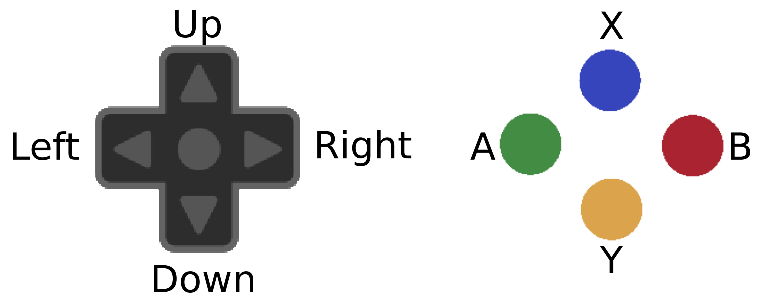


$$\mathcal{L}_{mix}(\pi; \pi_{RL}, \pi_{HE}) = \alpha \underbrace{\mathcal{L}_{\pi_{RL}}(\pi)}_{\text{Loss w.r.t. the RL agent}} + (1-\alpha) \underbrace{\mathcal{L}_{\pi_{HE}}(\pi)}_{\text{Loss w.r.t. the human expert}}$$

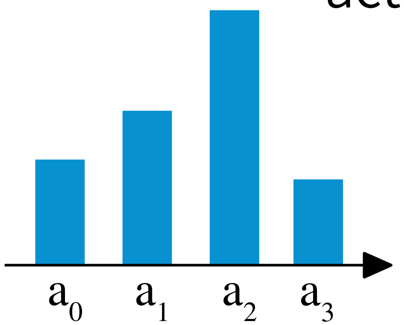
α Trade-off coefficient

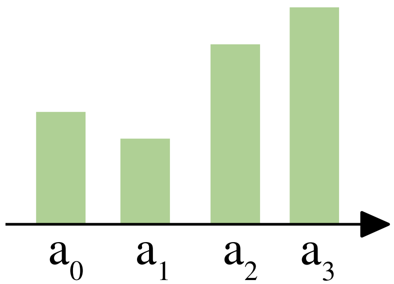
Proposed Method / Discrete Action Space

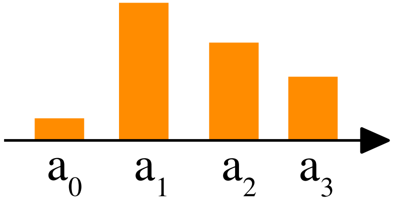
In the Discrete Action Space case:



An agent's policy can be considered as a probability distribution over the action space, given an arbitrary state.







Thus, leveraging the cross-entropy loss function, we get:

$$\mathcal{L}_{mix}(\cdot) = \alpha \mathbb{E}_s \left[-\sum_a \pi_{RL}^{(T)}(a|s) log \pi(a|s) \right] + (1-\alpha) \mathbb{E}_s \left[-\sum_a \pi_{RL}(a|s) log \pi(a|s) \right]$$

Proposed Method / Continuous Action Space

In the Continuous Action Space case:

Tight left turn: -1.0 Neutral: 0.0 Moderate right turn: 0.3

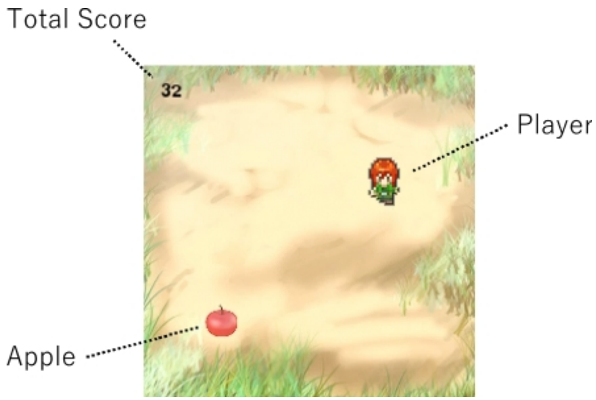
- An agent's policy can be expressed as the marginal probability distribution over the state-action space.
- The difference between two policies can thus be measure using a discriminator function (as in Generative Adversarial Networks)]

Rewriting the hybrid function around an adversarial loss function thus gives:

$$\mathcal{L}_{mix}(\cdot) = \mathbb{E}_{\tau\sim\pi}\left[log(D_w(s,a))\right] +$$
$$\alpha\mathbb{E}_{\tau_{RL}\sim\pi_{RL}}\left[log(1 - D_w(s,a))\right] + (1-\alpha)\mathbb{E}_{\tau_{HE}\sim\pi_{HE}}\left[log(1 - D_w(s,a))\right]$$

Experiment / Discrete Action Space case

Discrete Action Space Tasks



Apple game

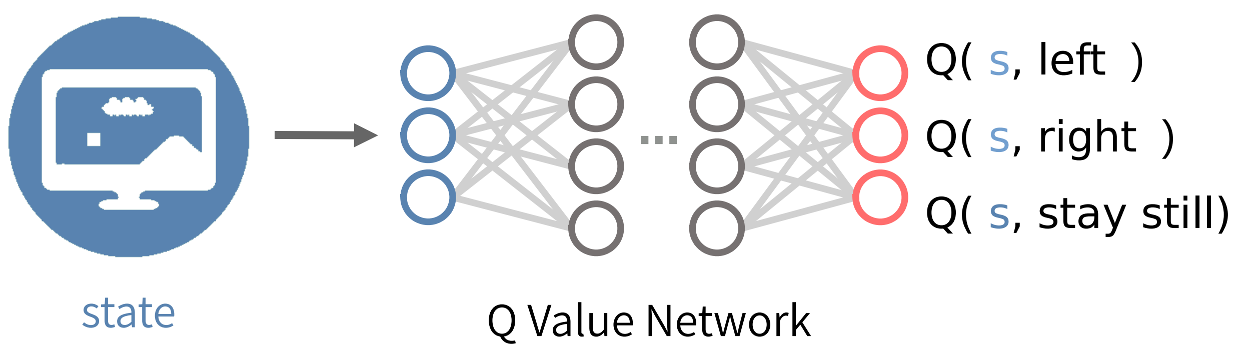
Goal
Collect randomly spawning apples by moving the avatar.

Gopher (Atari2600)

Goal
Filling back the holes being dug out by the gopher under the ground, thus keeping the latter from stealing the carrots.

Experiment / Discrete Action Space case

As the expert RL agent: Deep Q-Network (DQN) [Mnih et al., 2015]

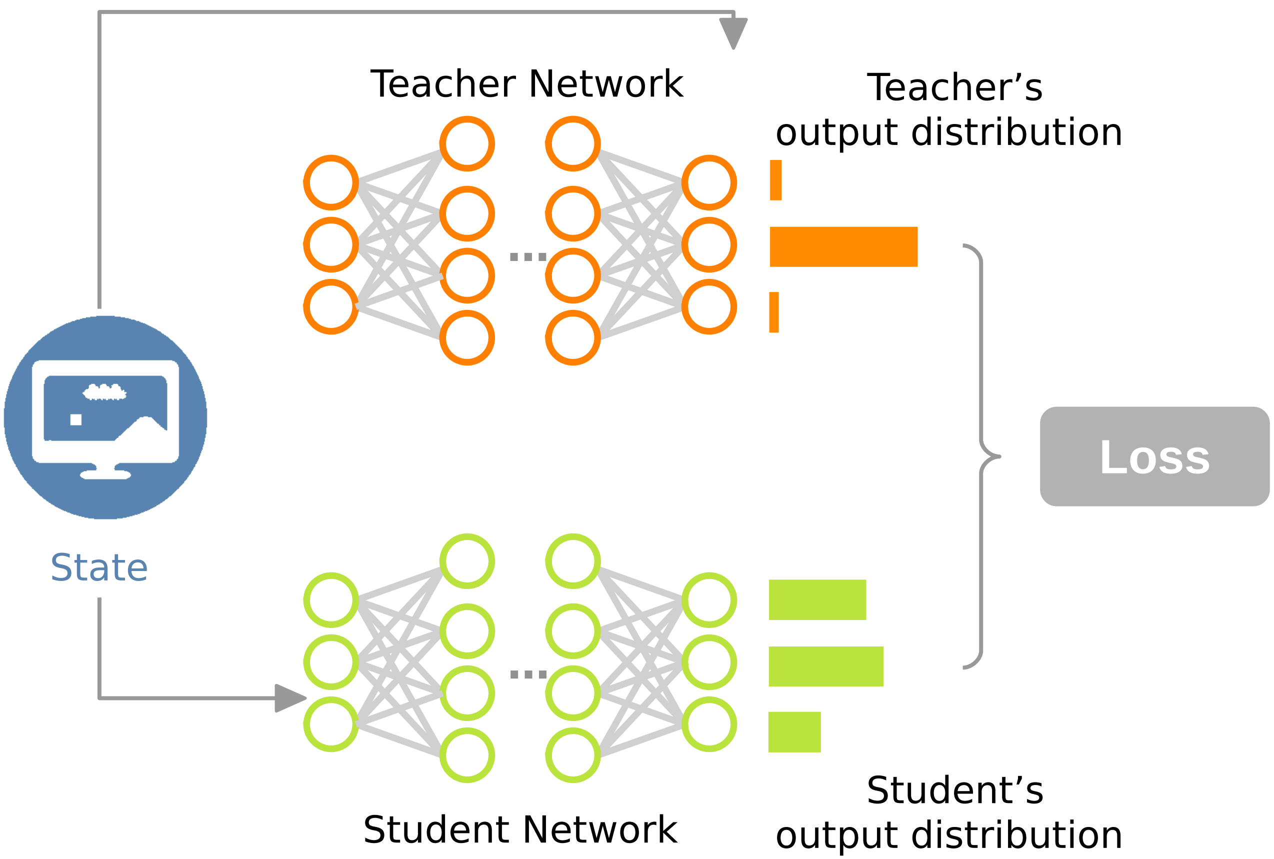


Selected Action

$$a = argmax_a Q(s, \cdot)$$

Experiment / Discrete Action Space case

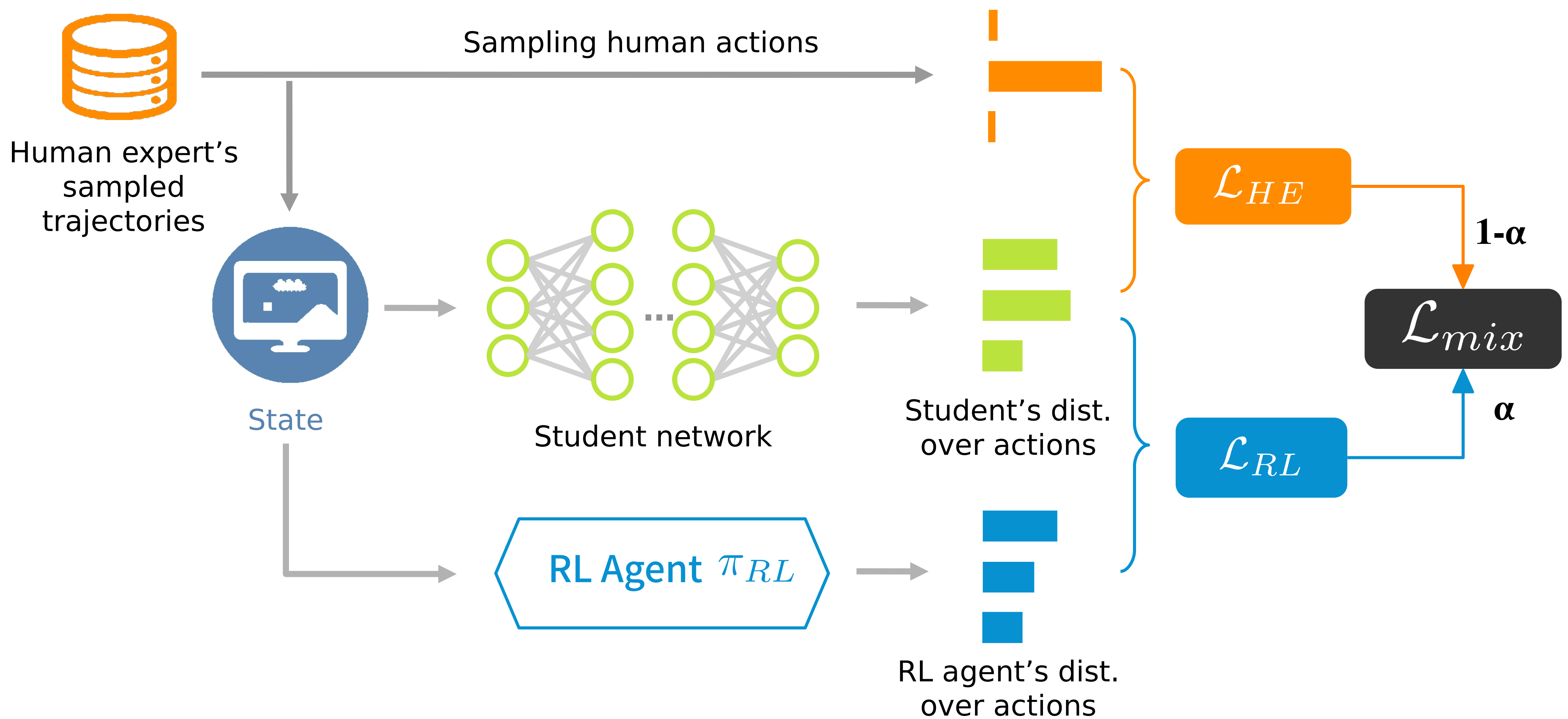
Imitation Learning method: Policy Distillation [Rusu et al., 2015]



同じ状態に対して， Teacher と生徒ネットワークそれぞれの出力の差を最小し，生徒ネットワークは Teacher を模倣することができる

Experiment / Discrete Action Space case

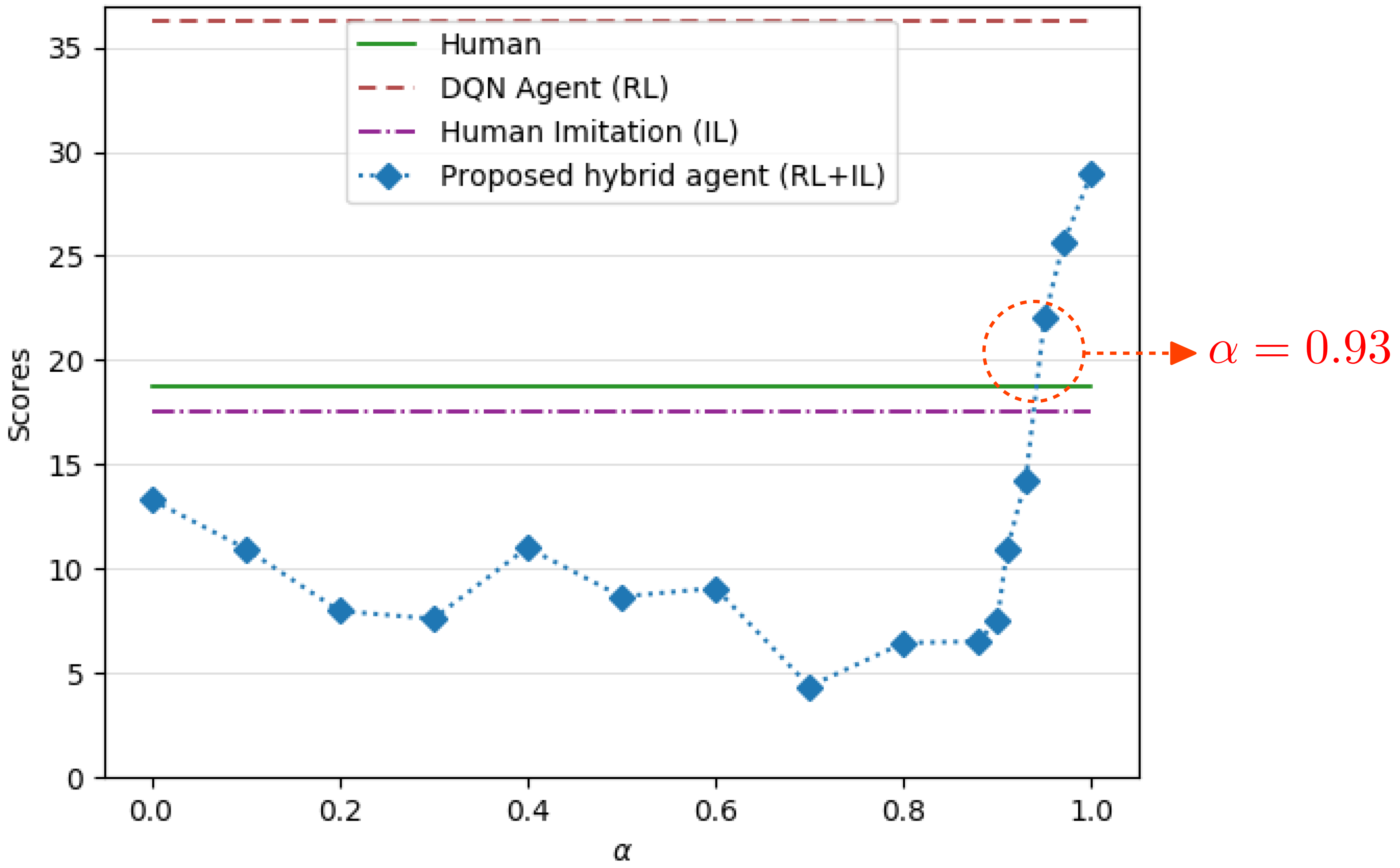
Proposed Hybrid Model



Based on "Policy Distillation", Rusu et al., 2015

Results / Discrete Action Space case

Apple game : Trade-off coefficient's impact on the hybrid agent



Results / Discrete Action Space case

*1st and 2nd places in **bold** and <u>underlined</u> fonts, respectively.

Apple game : Performance and sensitivity evaluations

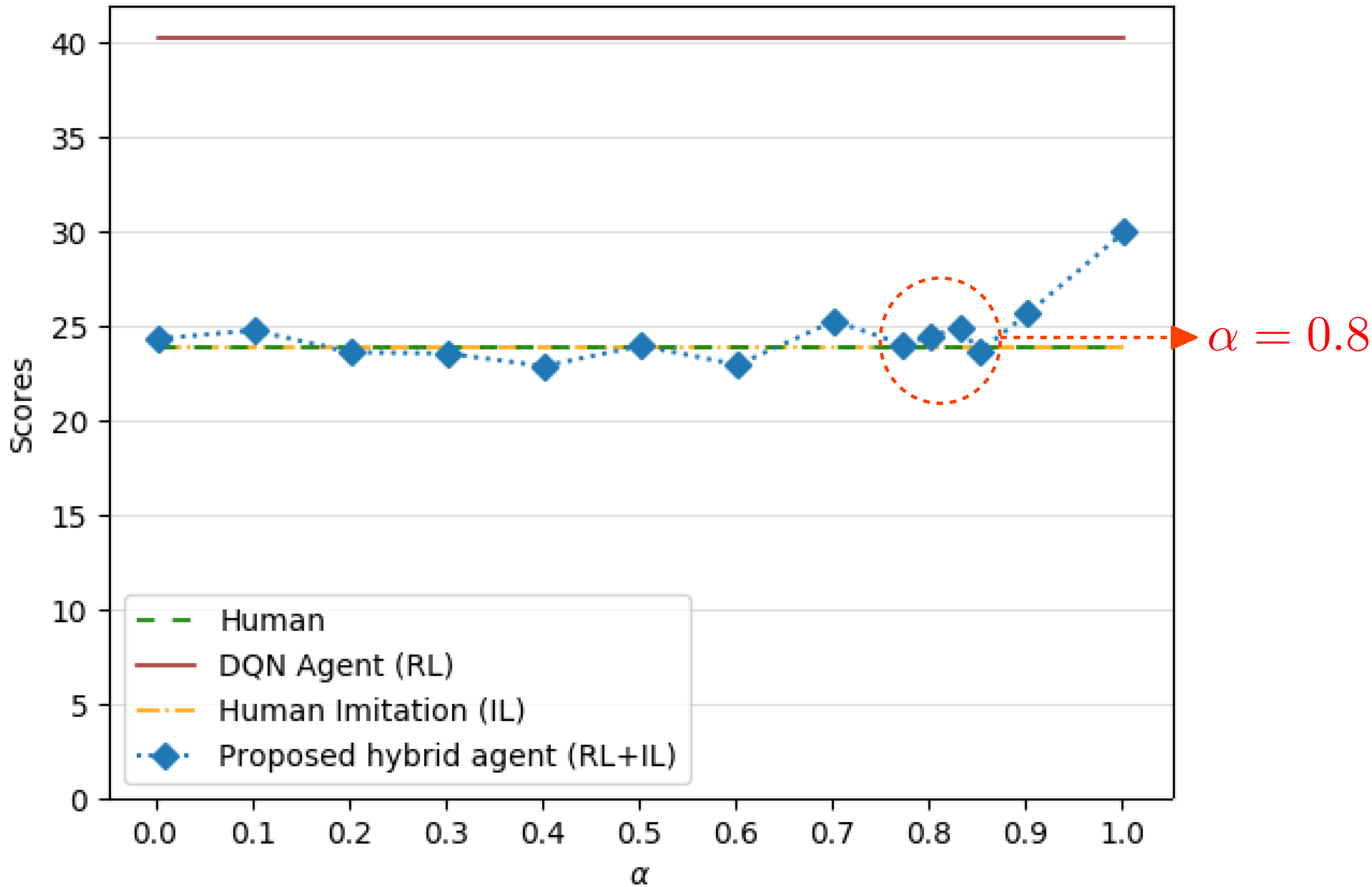
Agent	Score				Sensitivity test
	Max	Min	Mean	Std.	Identified as human
Human expert	27	11	18.71	2.86	**32 out of 50**
DQN(RL)	53	15	**36.27**	5.44	4 out of 50
Behavior cloning (IL)	29	3	17.57	4.37	22 out of 50
Proposed hybrid agent (RL+IL)	35	11	<u>22.02</u>	3.70	<u>27 out of 50</u>

Score DQN > Hybrid agent > Human > Human Imitation

Human-likeliness Human > Hybrid agent > Human Imitation > DQN

Results / Discrete Action Space case

Gopher : Trade-off coefficient's impact on the hybrid agent



Results / Discrete Action Space case

*1st and 2nd places in **bold** and underlined fonts, respectively.

Apple game : Performance and sensitivity evaluations

Agent	Score				Sensitivity test
	Max	Min	Mean	Std.	Identified as human
Human expert	81	2	23.87	19.81	<u>23 out of 52</u>
DQN(RL)	246	0	**40.30**	36.81	17 out of 52
Behavior cloning(IL)	126	0	23.91	23.79	**31 out of 52**
Proposed hybrid agent (RL+IL)	138	0	<u>26.05</u>	24.31	**31 out of 52**

Score DQN > Hybrid agent > Human Imitation > Human

Human-likeliness Hybrid agent = Human Imitation > Human > DQN

Experiment / Continuous Action Space case

Continuous Action Space Task

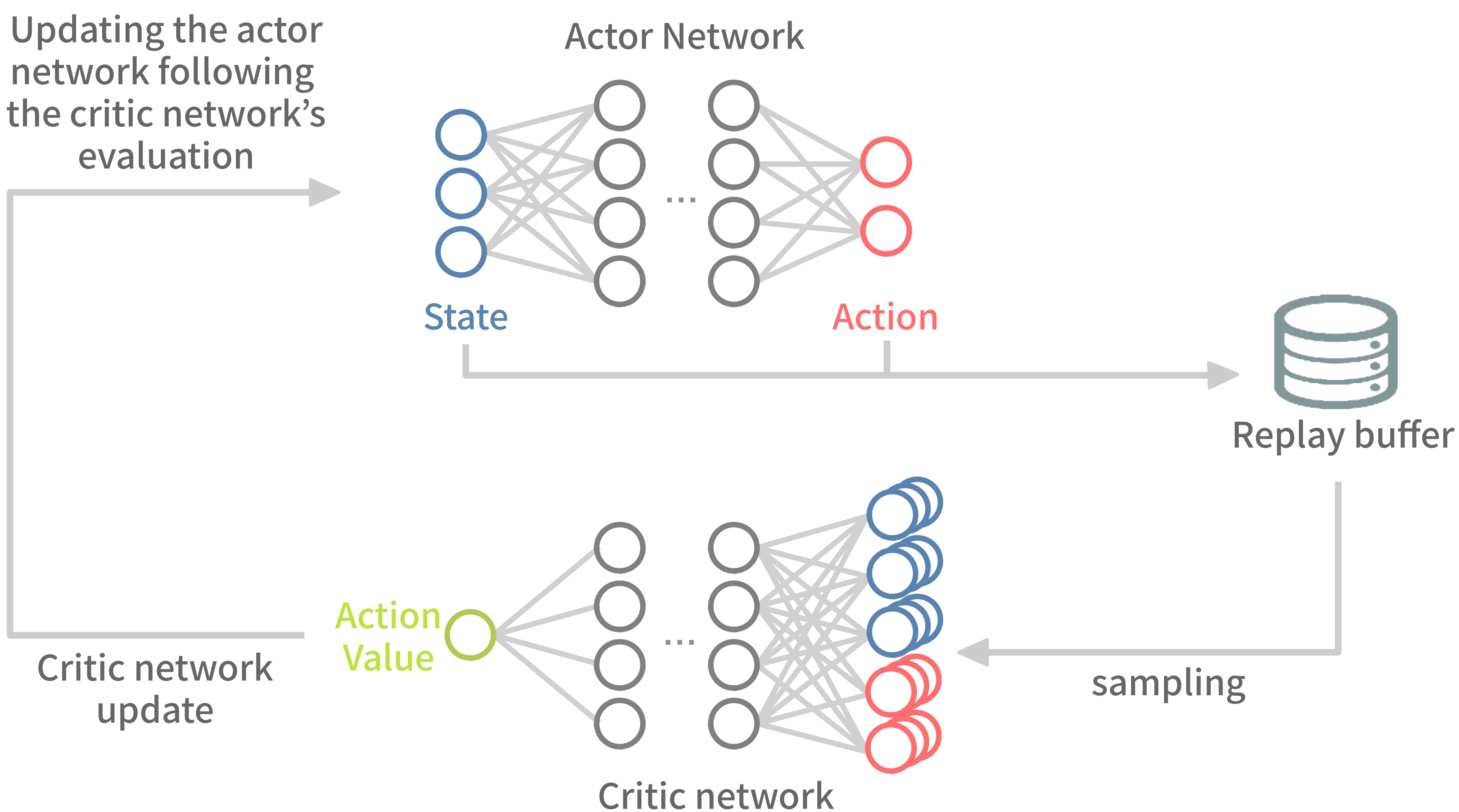
TORCS Racing Car Simulator

Goal

Drive a full lap while avoid obstacles
and exiting the track

Experiment / Continuous Action Space case

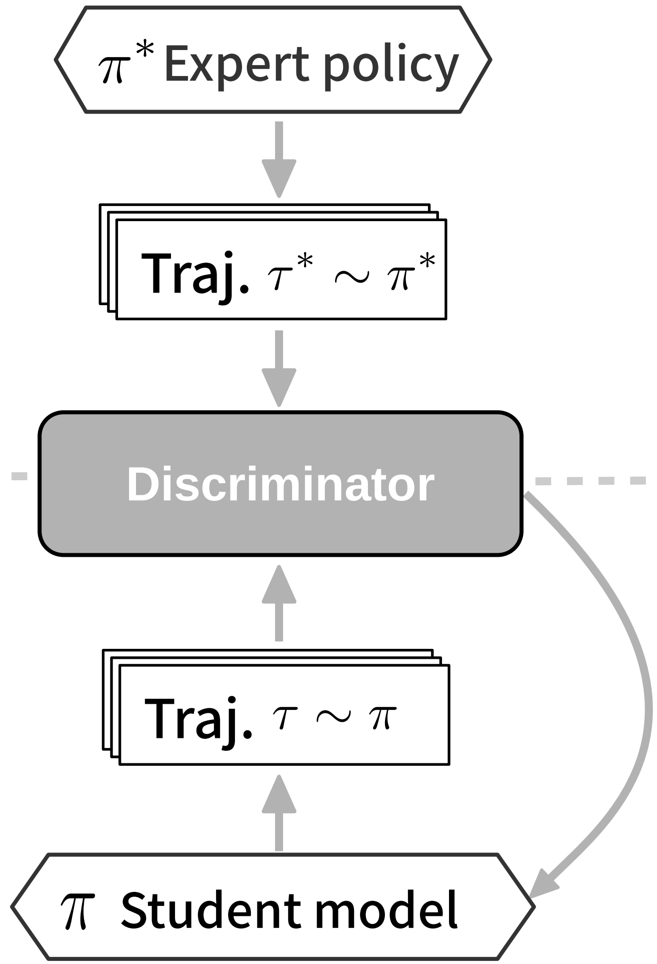
As the RL Agent : Deep Deterministic Policy Gradients [Lillicrap et al., 2015]



Experiment / Continuous Action Space case

IL method : Generative Adversarial Imitation Learning [Ho et al., 2016]

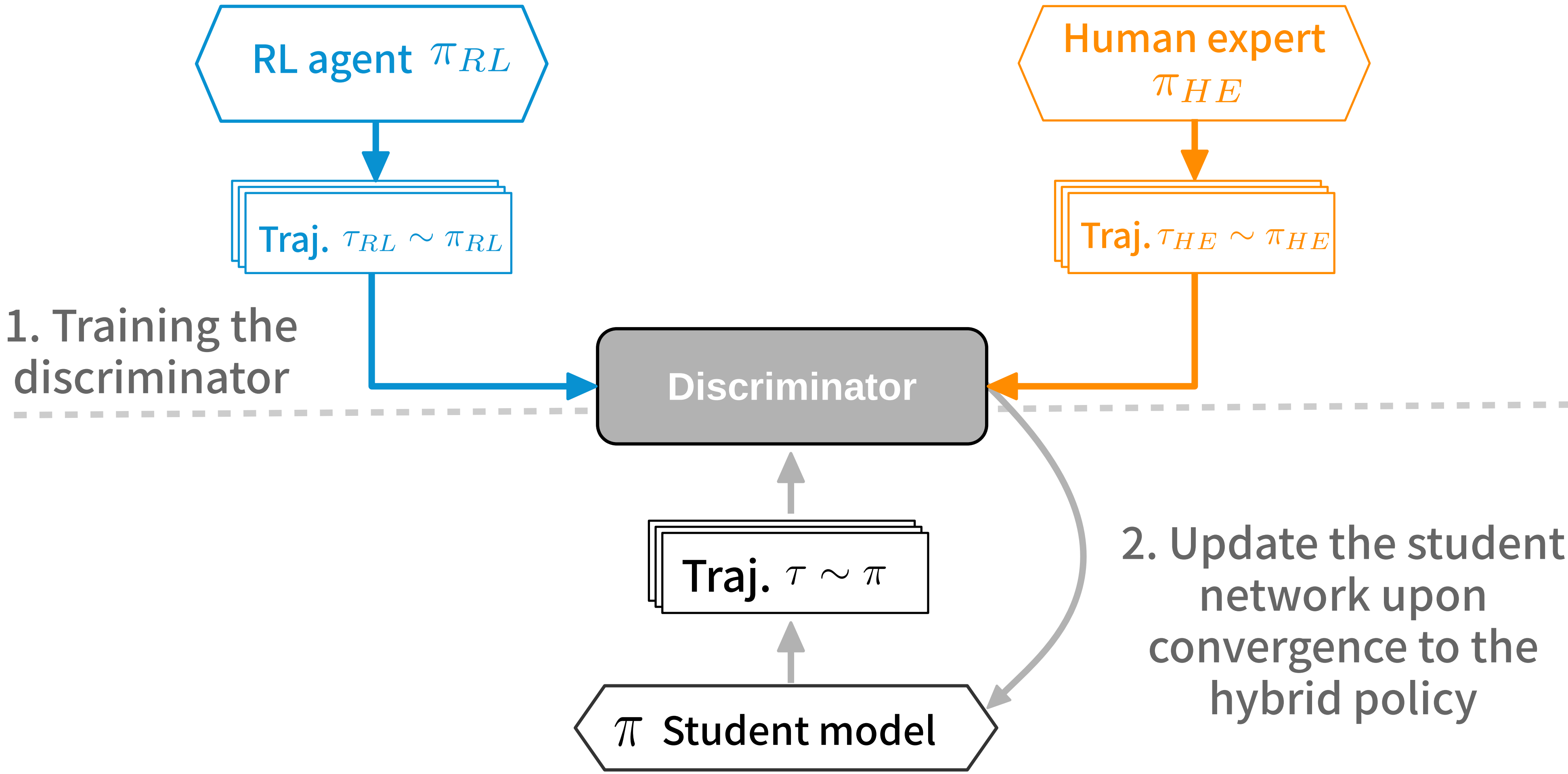
1. Training a discriminator over expert data



2. Update the student network to minimize the difference with the expert's policy

Experiment / Continuous Action Space case

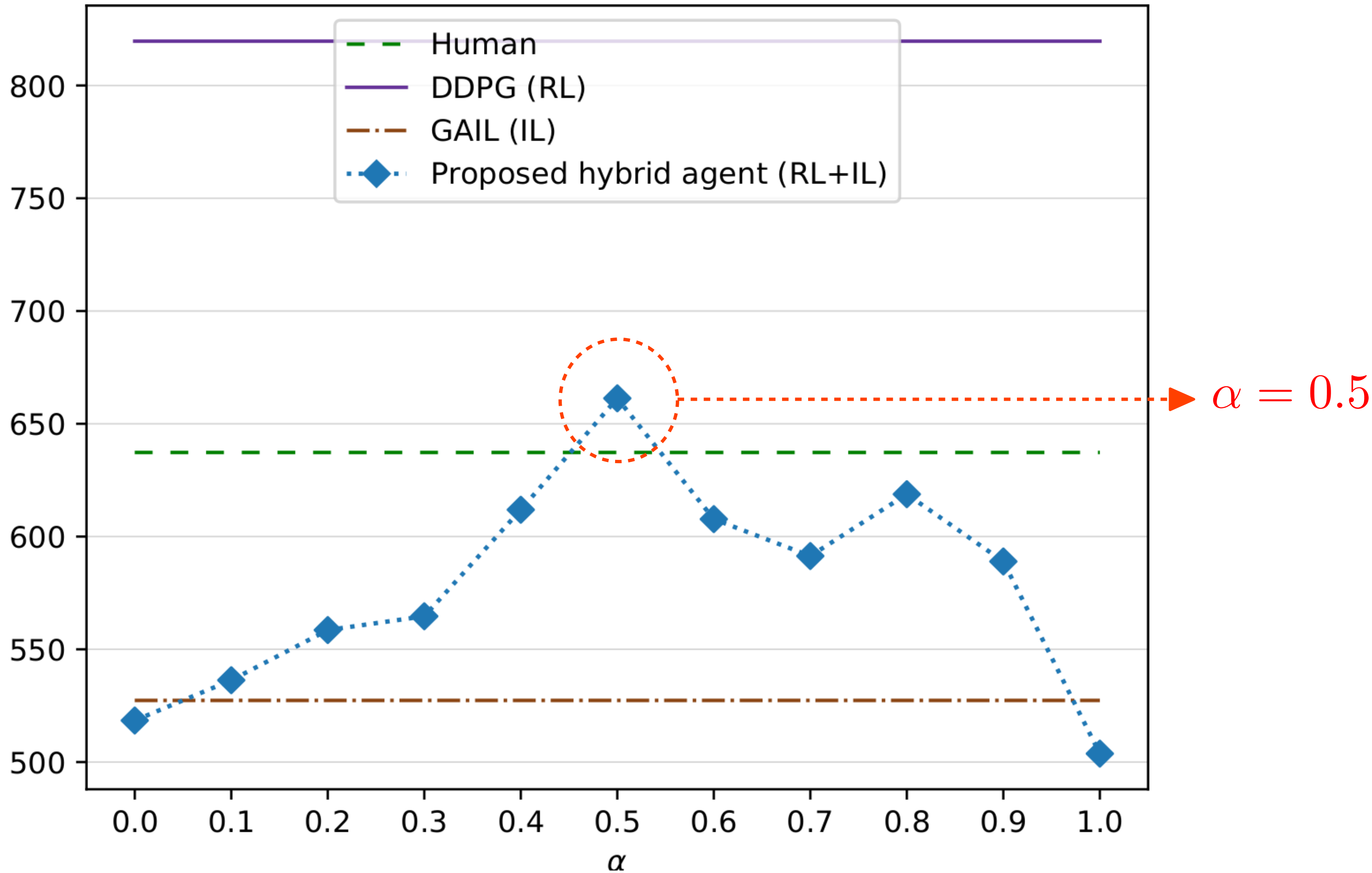
Proposed Hybrid Model



Based on "GAIL", Ho et al., 2015

Result / Continuous Action Space case

Torcs : Trade-off coefficient's impact on the hybrid agent



Result / Continuous Action Space case

*1st and 2nd places in **bold** and <u>underlined</u> fonts, respectively.

Torcs : Performance and sensitivity evaluations

モデル	Score				Sensitivity test
	Max	Min	Mean	Std.	Identified as human
Human expert	696.7	588.6	637.2	31.1	26 out of 52
DDPG(RL)	823.4	818.8	**819.6**	0.5	17 out of 52
GAIL(IL)	608.8	23.4	527.3	72.4	<u>27 out of 52</u>
Proposed hybrid agent (RL+IL)	817.8	107.4	<u>661.2</u>	179.2	**32 out of 52**

Score DQN > Hybrid agent > Human > Human Imitation

Human-likeliness Hybrid agent > Human Imitation > Human > DQN

Summary

- Proposed an hybrid of reinforcement and imitation learning

- Adapted the proposed hybrid method to both discrete and continuous action space tasks.

- Experimented said method on:
  - 2 discrete action task (**Apple Game** and **Atari 2600's Gopher**)
  - 1 continuous action task (Torcs Racing Car Simulator)

- The proposed hybrid agent
  - achieved similar, if not better performance than the human expert and its imitation
  - Was identified as more human likely than reinforcement learning counterpart.

Appendix / Sensitivity test

Sensitivity test's details

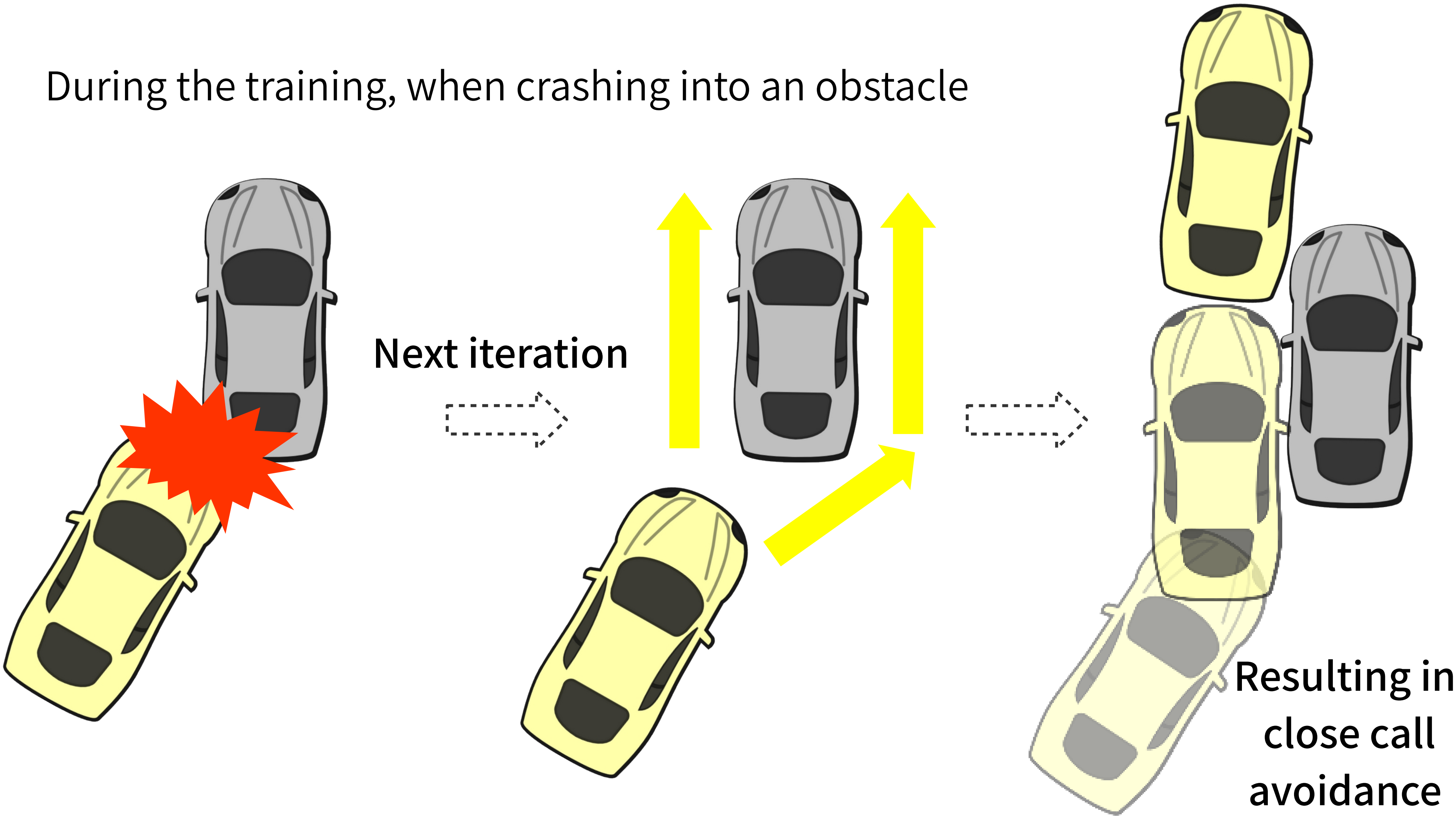
- 26 participants
  - 23 males
  - 03 females

- Each participant
  - was first instructed on the different games as well as an opportunity to try by himself
  - then provided with 2 game play video of every agent (human – RL agent – human imitation – Proposed hybrid agent) for each game
  - and request to label each one of the video as either "human" or "AI".

Appendix / Causes of some undesirable RL agent's behaviors

1. Close call avoidance

During the training, when crashing into an obstacle



Appendix / Causes of some undesirable RL agent's behaviors

1. Edge proximity