強化学習と模倣学習の融合による人間らしいエージェント

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

2019-06-18

神戸大学　システム情報学研究科　ドッサ　ルスラン

研究背景

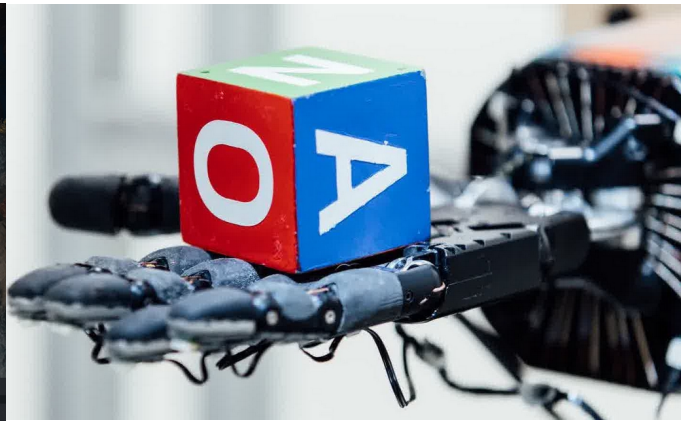
強化学習では

長所
- **性能が高い**

短所
- 振る舞いが**不自然**である恐れがある







囲碁や Starcraft 2 というゲームで世界トッププレイヤに勝利できる
また，作業ロボットも操作できている

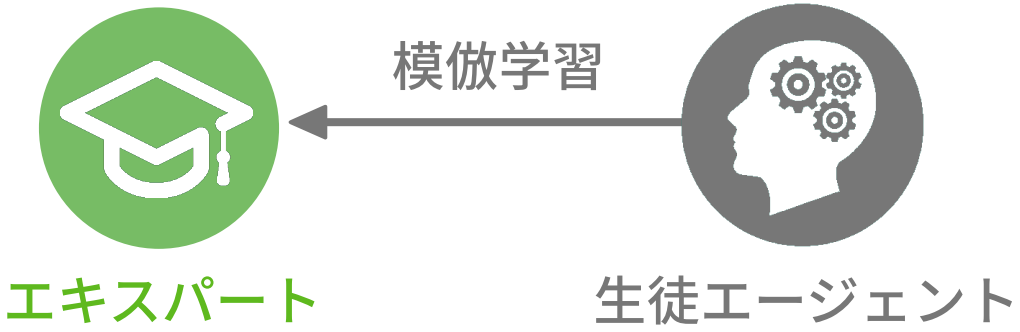
研究背景

模倣学習では

あるエキスパートのデータを教師データとして，**教師あり学習**を行う

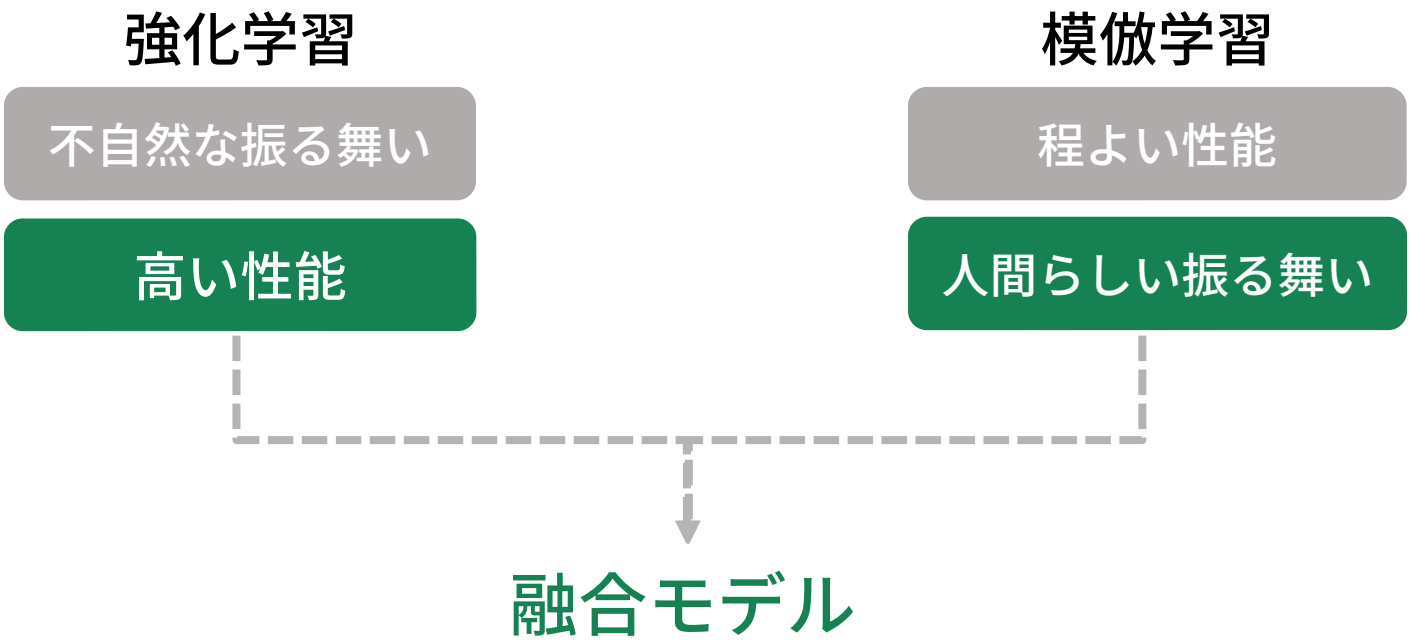
長所
- 人間を真似ているので，かなり**自然な振る舞い**

短所
- 強化学習の**性能より悪い**，エキスパートの性能を超えられない



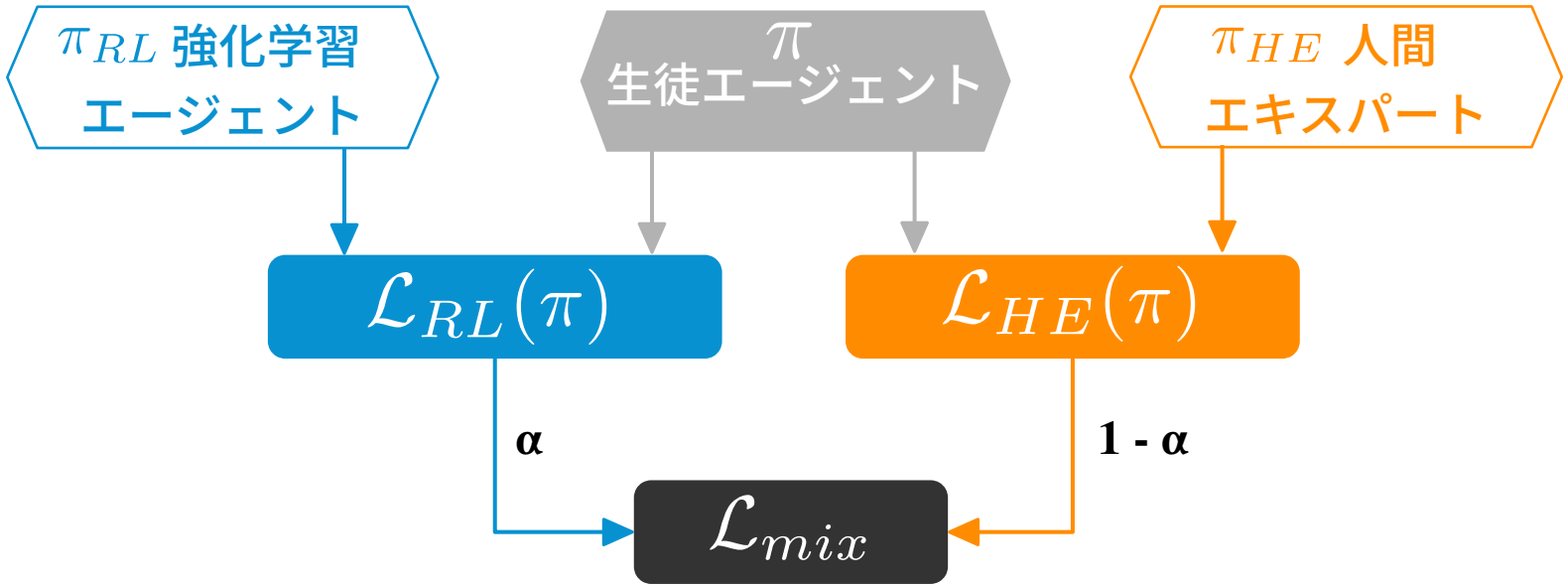
本研究の目的

強化学習と人間の模倣学習の融合により，それぞれのエキスパートの長所を持っているエージェントの実現となっている．



提案手法

融合の損失関数



$$\mathcal{L}_{mix}(\pi;\pi_{RL},\pi_{HE}) = \alpha\underbrace{\mathcal{L}_{\pi_{RL}}(\pi)}_{\text{強化学習エキスパートに対する損失}} + (1-\alpha)\underbrace{\mathcal{L}_{\pi_{HE}}(\pi)}_{\text{人間エキスパートに対する損失}}$$

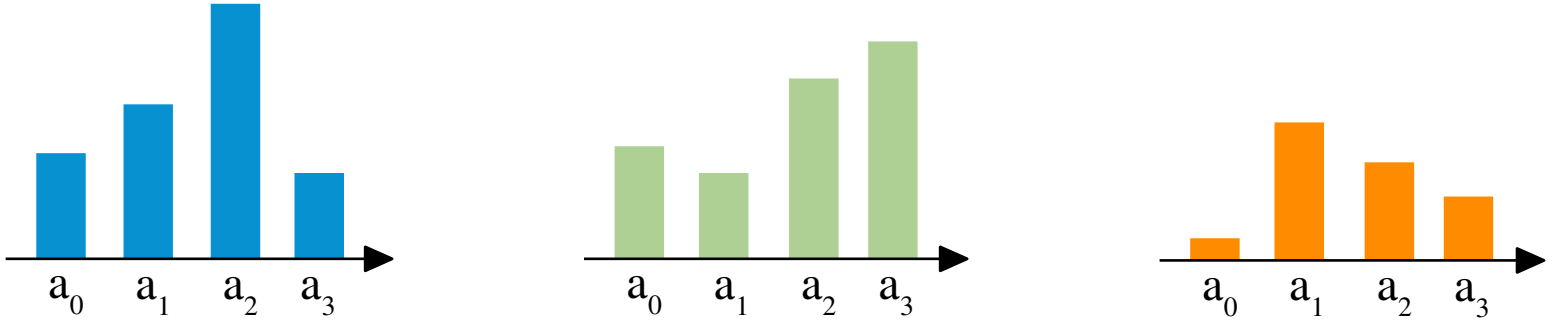
α トレードオフ係数で調整する

提案手法・離散行動空間

離散行動空間タスクの場合



各エージェントの方策は全体の行動空間上の確率分布として考えらるため



交差エントロピー損失を用いる

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

提案手法・連続行動空間

連続行動空間タスクの場合



強く左折する
-1.0



ニュートラル
0.0



少し右折する
0.3

エージェントの方策は 状態・行動 空間上の marginal probability distribution とみなす

Generative Adversarial Network (GAN) のような識別器を用い，生徒エージェントとそれぞれのエキスパートとの差を測定できる

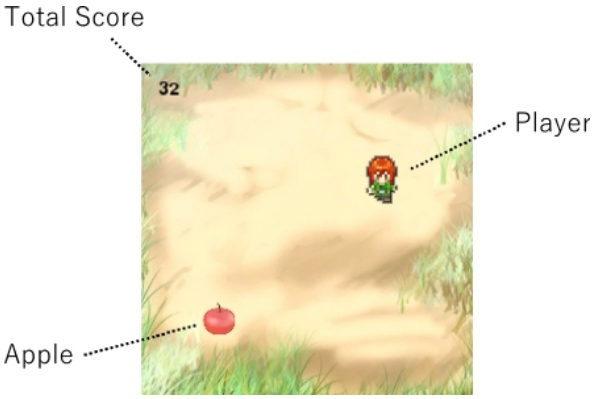
以前の一般融合損失関数を敵対損失関数に置き換える

$$\mathcal{L}_{mix}(\cdot) = \mathbb{E}_{\tau\sim\pi}\left[log(D_w(s,a))\right] +$$

$$\alpha\mathbb{E}_{\tau_{RL}\sim\pi_{RL}}\left[log(1-D_w(s,a))\right] + (1-\alpha)\mathbb{E}_{\tau_{HE}\sim\pi_{HE}}\left[log(1-D_w(s,a))\right]$$

実験　離散行動空間のタスク

離散行動空間



Apple game

目的 : 任意に現れるリンゴを獲得すること

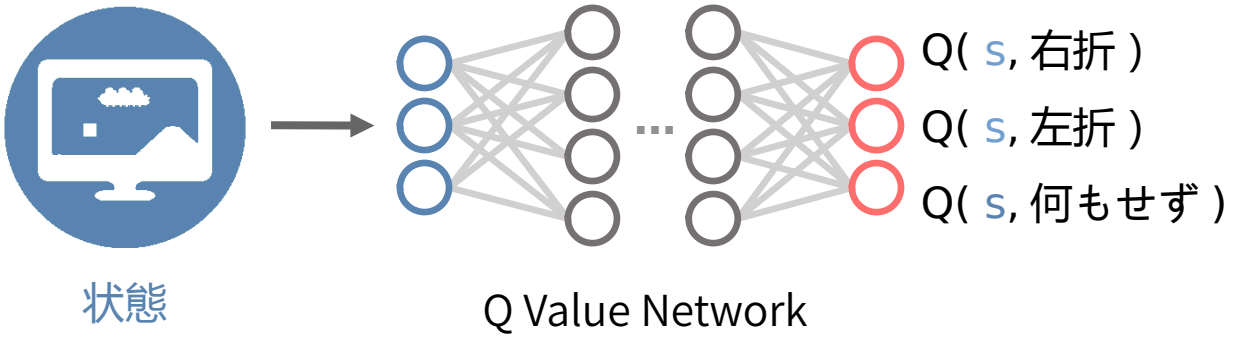


Gopher (Atari2600)

目的 : 地下から這い出したりするネズミに掘られる穴を埋めつつ，人参を守護すること

実験　離散行動空間のタスク

強化学習エージェントには Deep Q-Network (DQN) [Mnih et al., 2015]

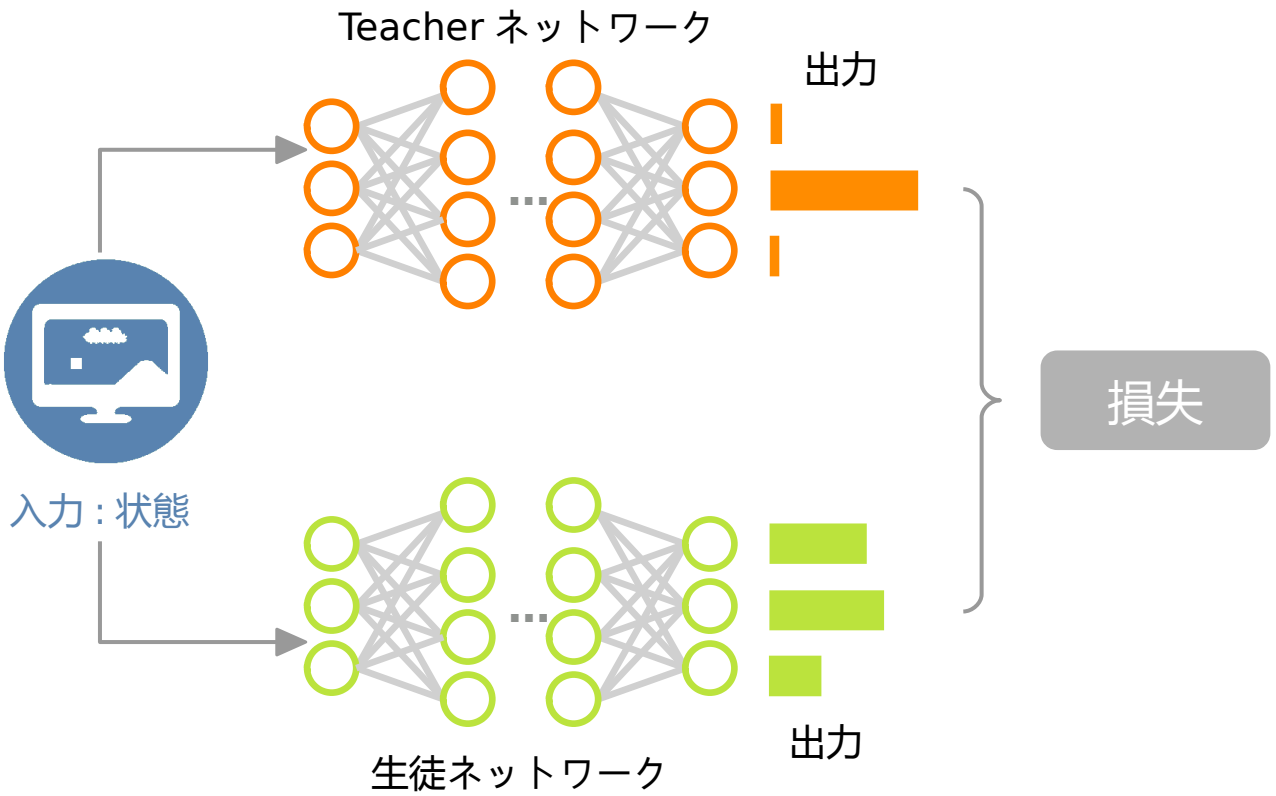


選択する行動

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

実験　離散行動空間のタスク

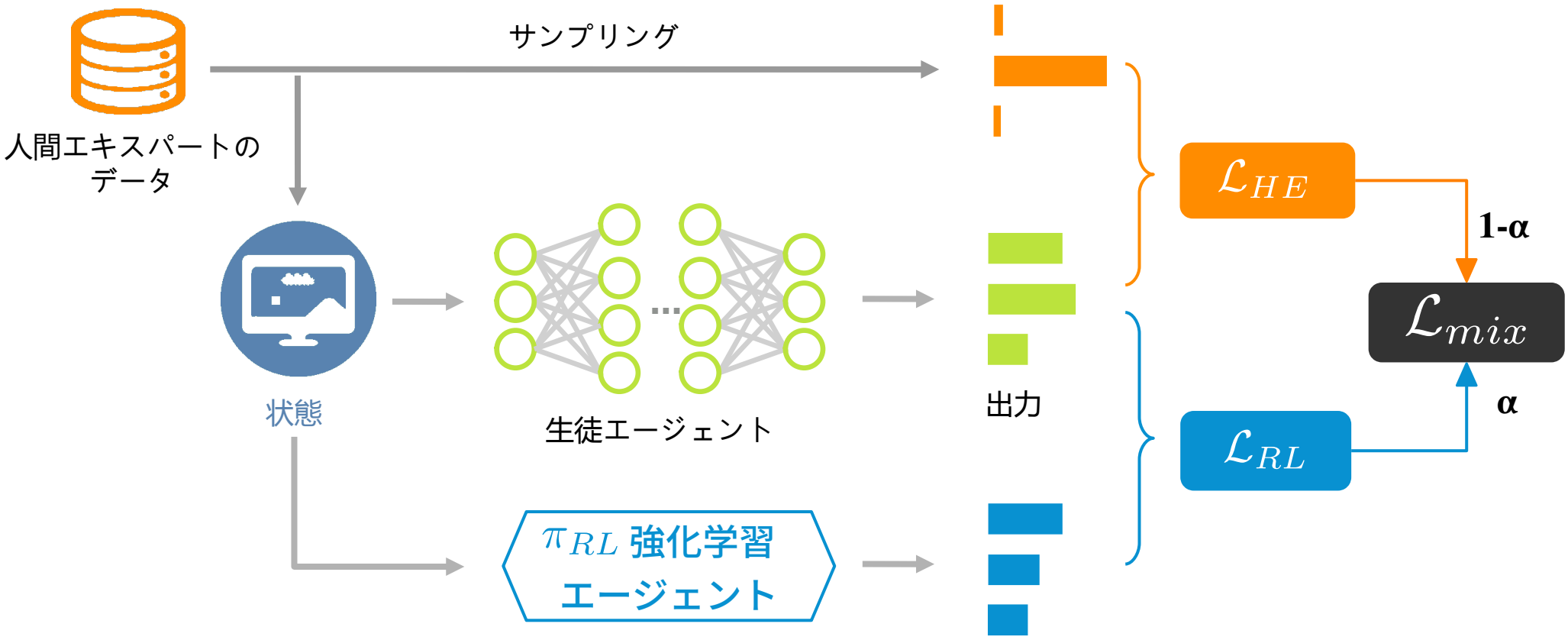
模倣学習には Policy Distillation [Rusu et al., 2015]



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

実験　離散行動空間のタスク

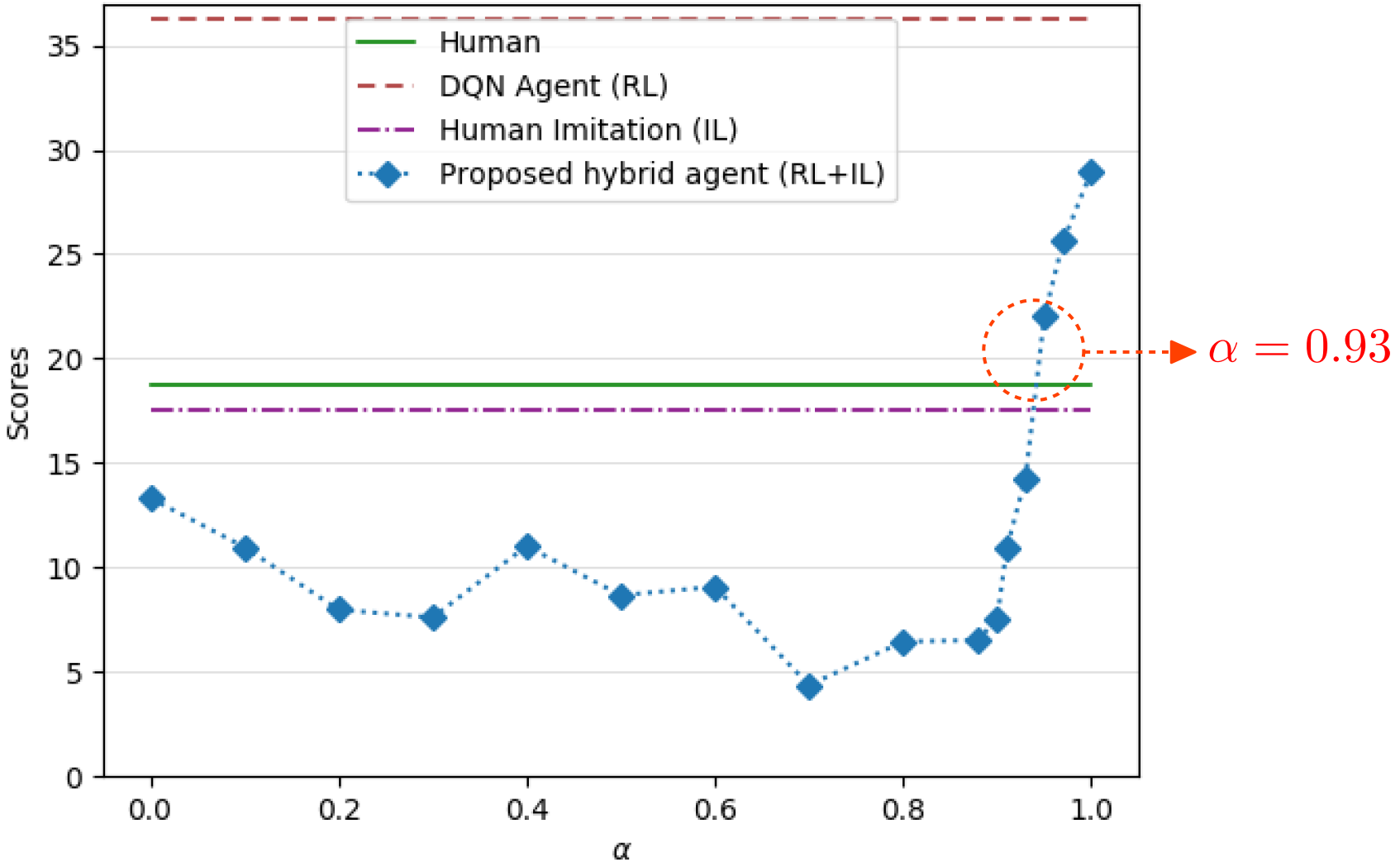
提案の融合モデル



* Policy Distillation, Rusu et al., 2015 を参考にする

結果　離散行動空間のタスク

Apple game ・トレードオフ係数の影響



結果　離散行動空間のタスク

* 1 位と 2 位をそれぞれ太字と下線で示す
スコアは獲得されたりんごの数

Apple game ・性能評価及び感性評価

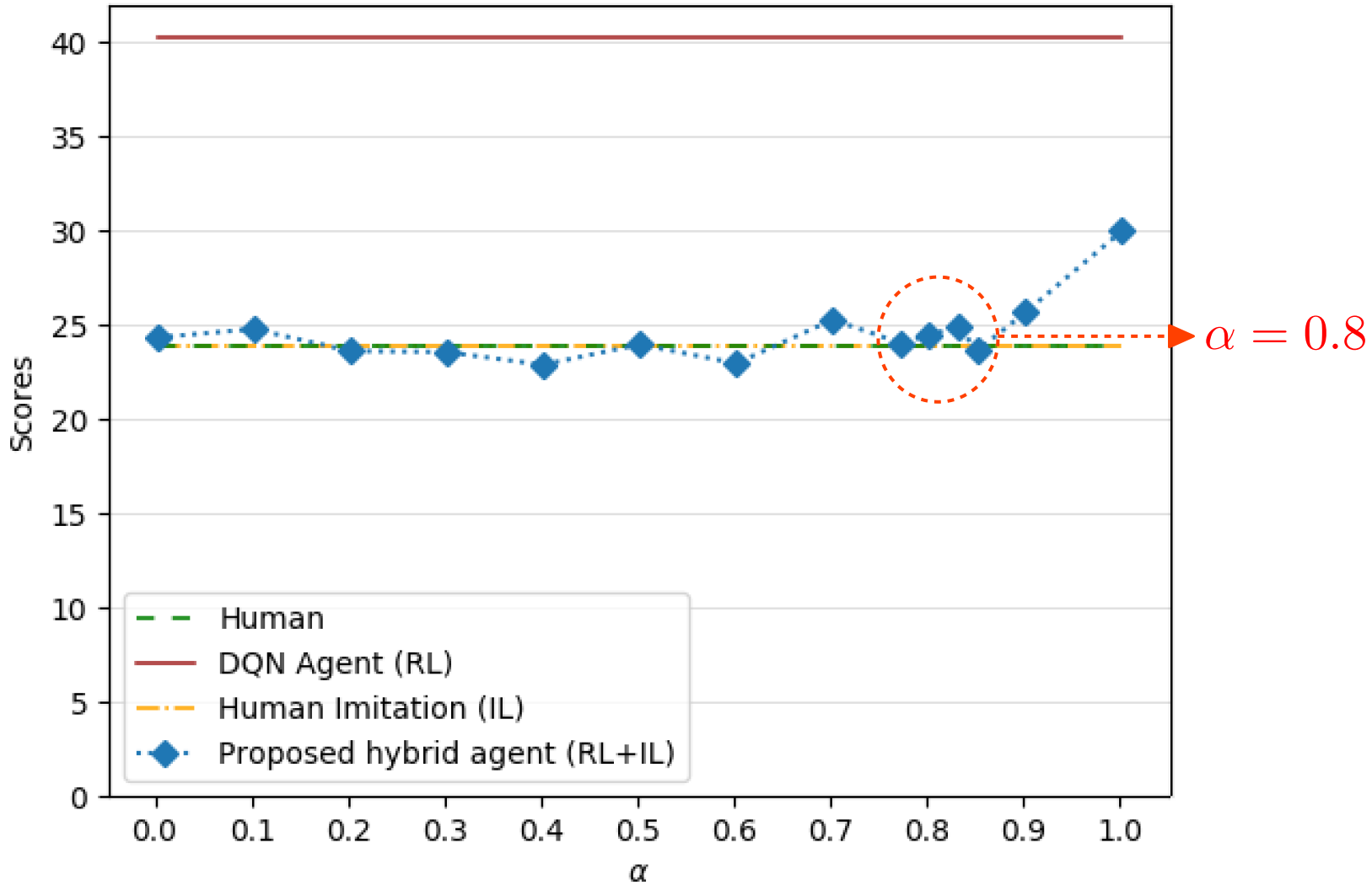
モデル	点数				感性評価
	Max	Min	平均	標準偏差	人間と判断された
人間エキスパート	27	11	18.71	2.86	**32 out of 50**
強化学習 (RL)	53	15	**36.27**	5.44	4 out of 50
人間の模倣 (IL)	29	3	17.57	4.37	22 out of 50
融合モデル (RL+IL)	35	11	<u>22.02</u>	3.70	<u>27 out of 50</u>

性能　強化学習 > 提案モデル > 人間 > 人間の模倣

人間らしさ　人間 > 提案モデル > 人間の模倣 > 強化学習

結果　離散行動空間のタスク

Gopher ・トレードオフ係数の影響



結果　離散行動空間のタスク

* 1 位と 2 位をそれぞれ太字と下線で示す

Gopher ・性能評価及び感性評価

モデル	点数				感性評価
	Max	Min	平均	標準偏差	人間と判断された
人間エキスパート	81	2	23.87	19.81	<u>23 out of 52</u>
強化学習 (RL)	246	0	**40.30**	36.81	17 out of 52
人間の模倣 (IL)	126	0	23.91	23.79	**31 out of 52**
融合モデル (RL+IL)	138	0	<u>26.05</u>	24.31	**31 out of 52**

性能　強化学習 > 提案モデル > 人間の模倣 > 人間

人間らしさ　提案モデル = 人間の模倣 > 人間 > 強化学習

実験　連続行動空間のタスク

連続行動空間

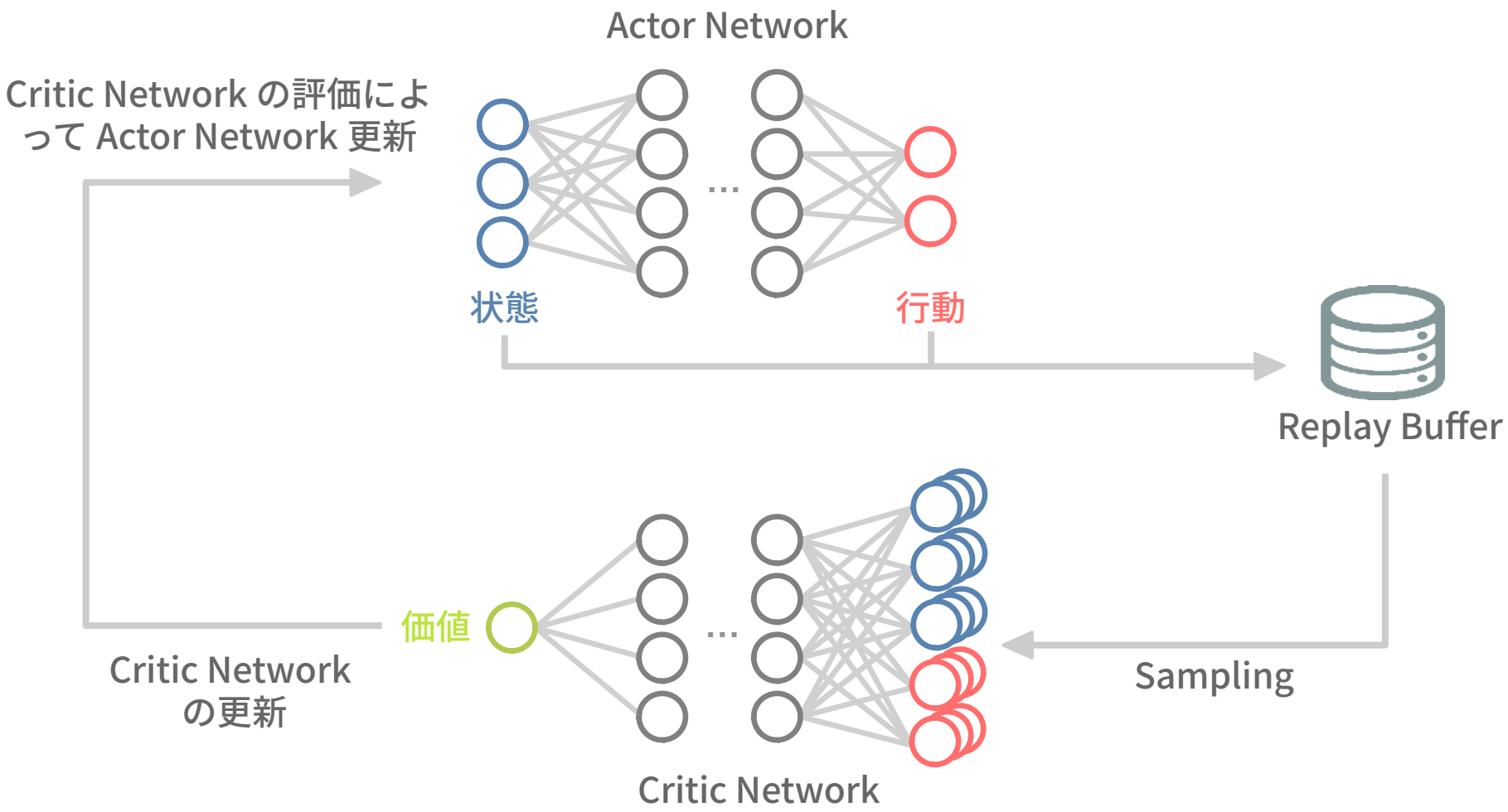


TORCS Racing Car Simulator

目的：障害物を避難し，トラックから脱出せずに走行距離を最大化する

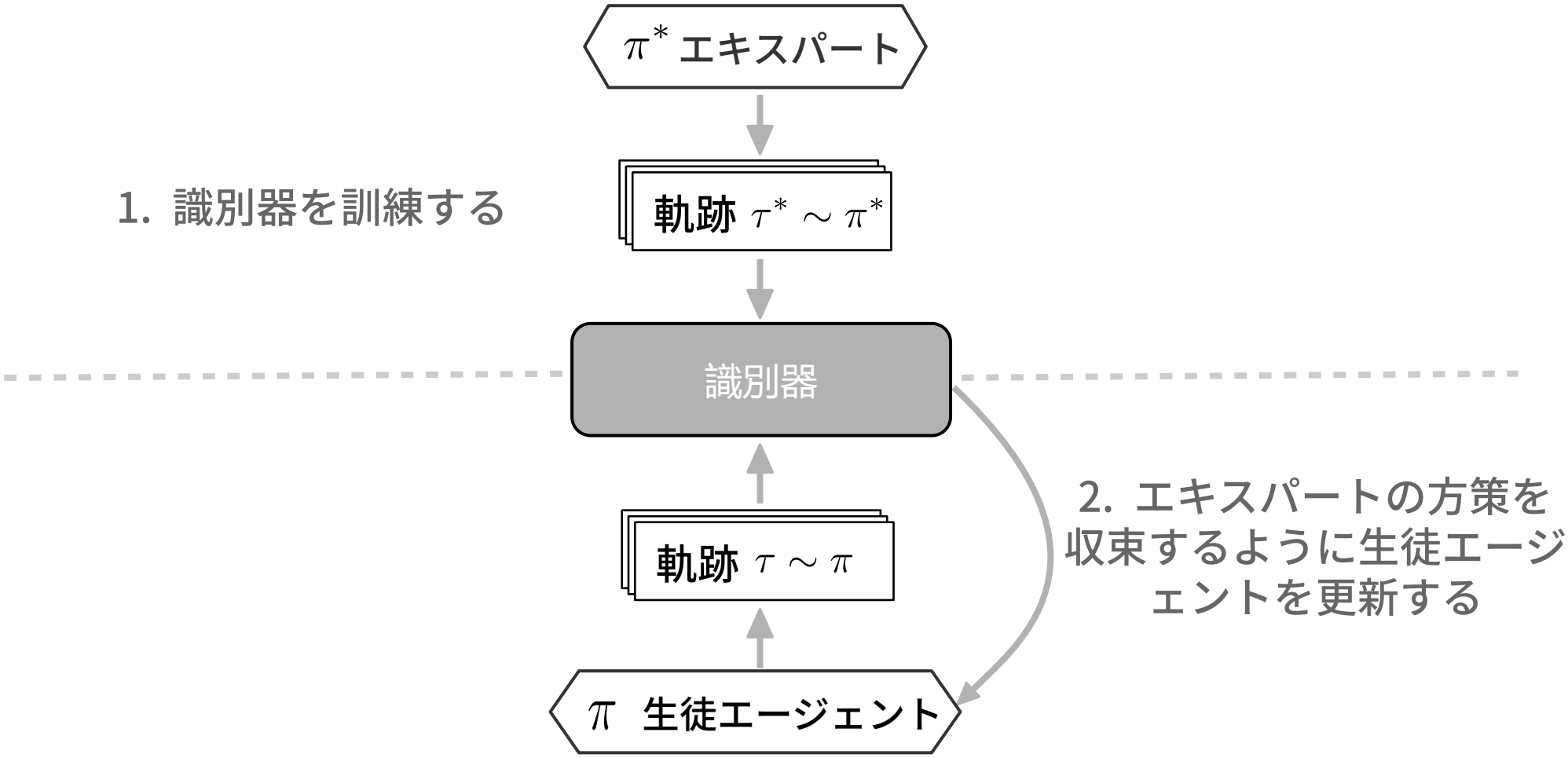
実験　連続行動空間のタスク

強化学習エージェントには Deep Deterministic Policy Gradients [Lillicrap et al., 2015]



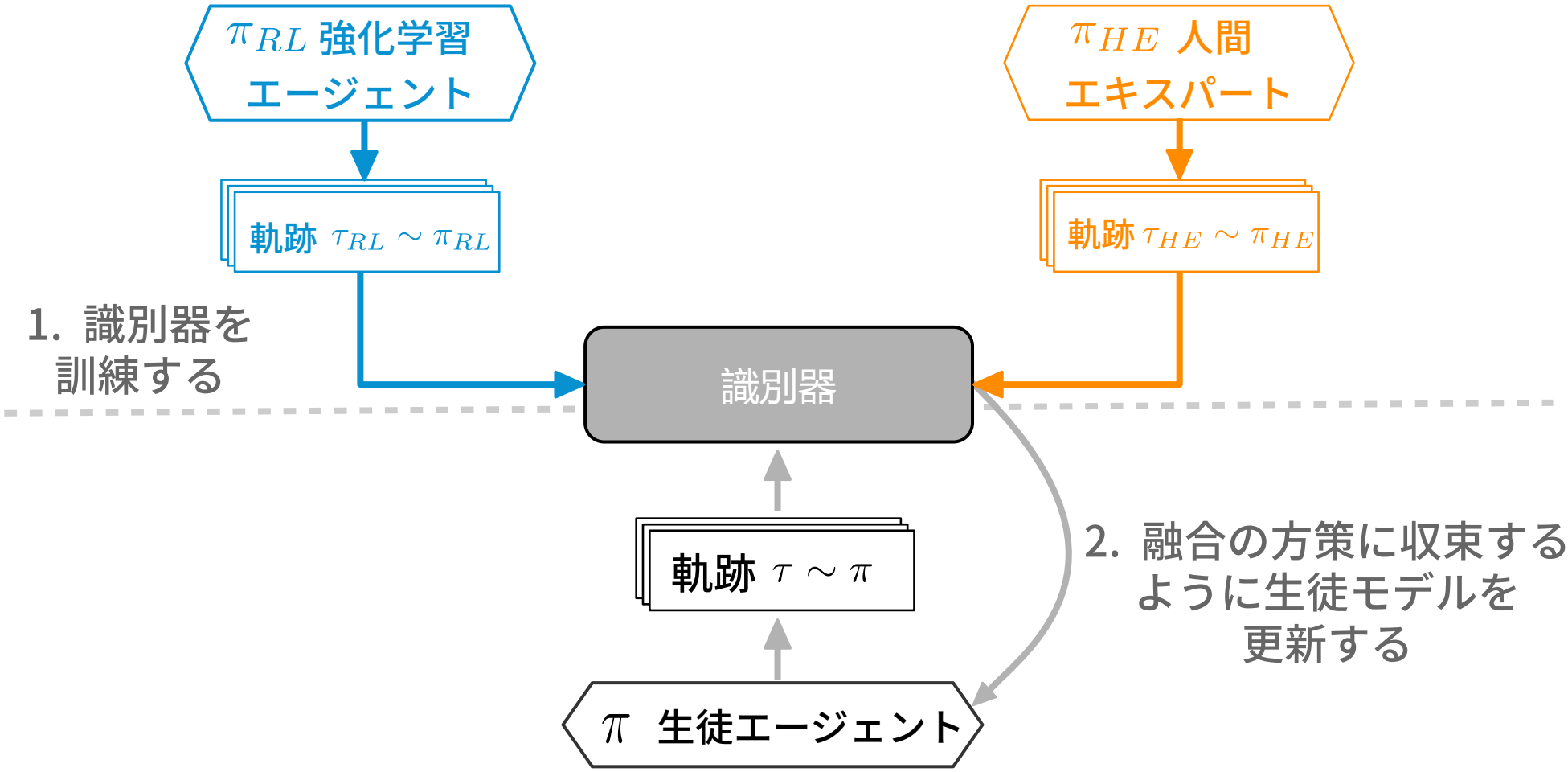
実験　連続行動空間のタスク

模倣学習には Generative Adversarial Imitation Learning [Ho et al., 2016]



実験　連続行動空間のタスク

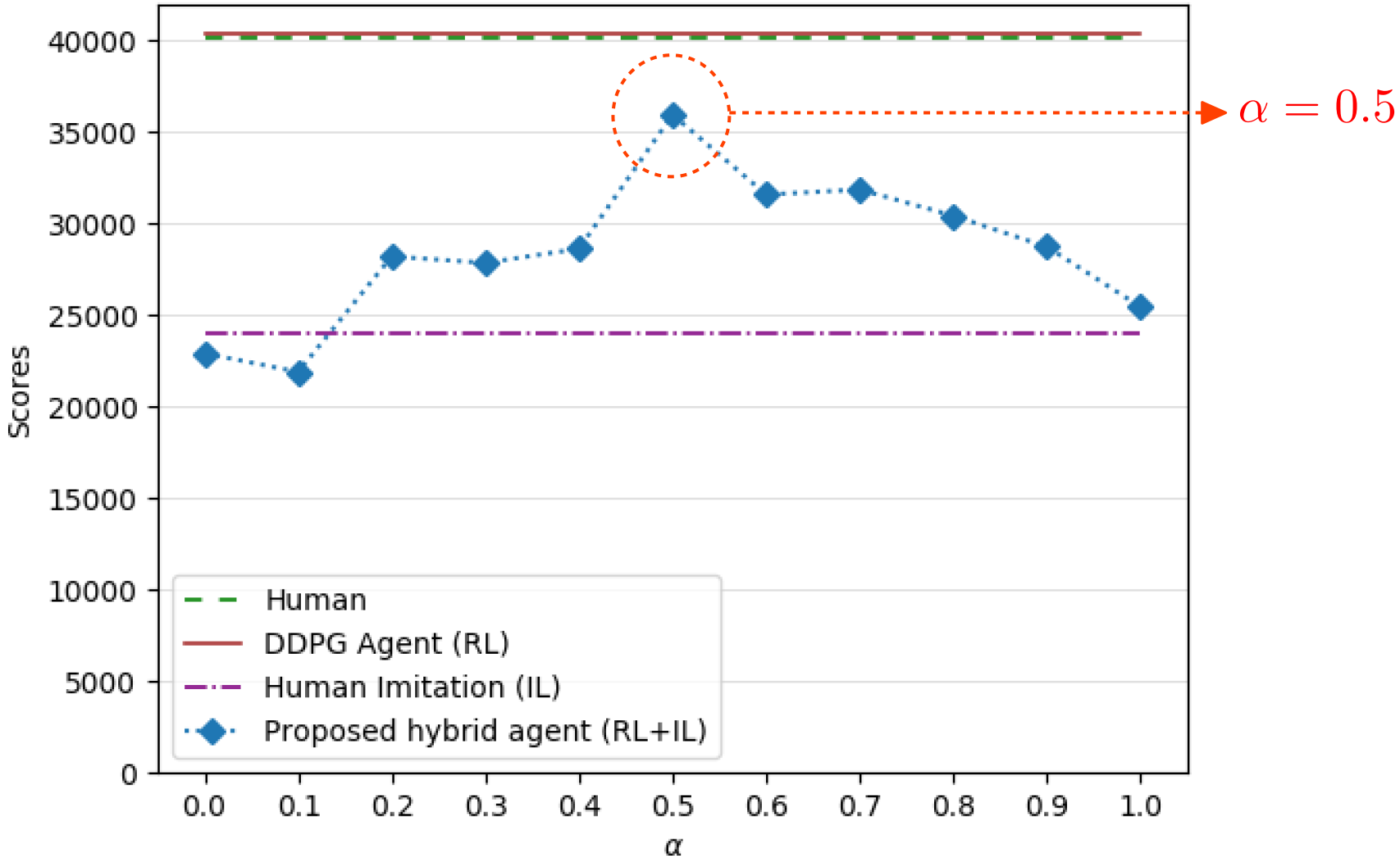
提案の融合モデル



* GAIL, Ho et al., 2015 を参考にする

結果　連続行動空間のタスク

Torcs Racing Car Simulator ・トレードオフ係数の影響



結果　連続行動空間のタスク

* 1 位と 2 位をそれぞれ太字と下線で示す
性能は走行距離

Torcs ・性能評価及び感性試験

モデル	点数 (x10^3)				感性評価
	Max	Min	平均	標準偏差	人間と判断された
人間エキスパート	48.70	27.68	40.17	3.63	26 out of 52
強化学習 (RL)	41.91	39.91	**40.45**	0.43	17 out of 52
人間の模倣 (IL)	39.93	14.88	23.99	1.16	<u>27 out of 52</u>
融合モデル (RL+IL)	40.22	6.71	<u>36.05</u>	1.32	**32 out of 52**

性能　　強化学習 > 人間 > 提案モデル > 人間の模倣

人間らしさ　　提案モデル > 人間の模倣 > 人間 > 強化学習

まとめ

- 強化学習エージェントと人間の模倣学習により融合モデルを提案
- タスクの種類によって，融合モデルの変種を提案
- 離散行動空間も連続行動空間のタスクに応用
- 提案した融合モデルは
  - 性能評価によって，人間，及びその模倣の性能をほとんど超えている
  - 感性評価によって，高い割合で人間だと判断されている

付録

感性試験の詳細

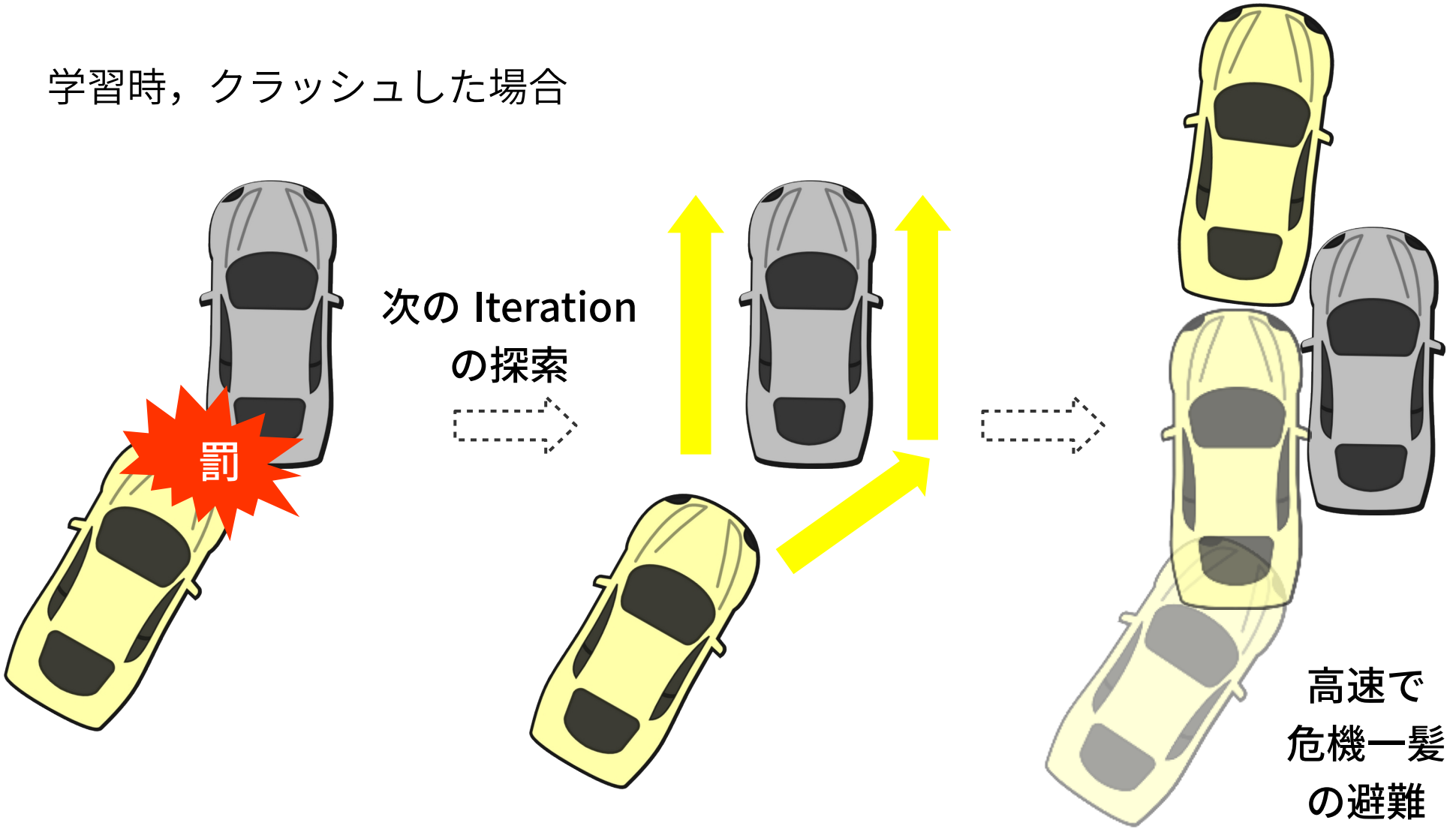
- 審査員は計 26 人で
  - 男　23
  - 女　03

- 各審査員には
  - ゲームの体験会に参加
  - モデル（人間・人間の模倣・融合エージェント・強化学習）のゲームプレイ動画 2 本を提供（ Apple Game と Gopher は 15 秒， Torcs は 30 秒）
  - 「人間」か「 AI 」か判断してもらう

付録

強化学習エージェントの不自然・望ましくない振る舞いの起源（１）

学習時，クラッシュした場合



強化学習エージェントの不自然・望ましくない振る舞いの起源（２）

学習時，トラックを脱出する場合

