オープンソース仮想シミュレーション環境「箱庭」による強化学習の対応の検討

細合晋太郎，周莎，高瀬英希 (東京大学)，
福田竜也 ((株)インテック)，高田 光隆(名古屋大学)，
久保秋 真((株)チェンジビジョン)，森 崇((合)箱庭ラボ)

TOPPERS
Toyohashi OPen Platform for Embedded Real-time Systems

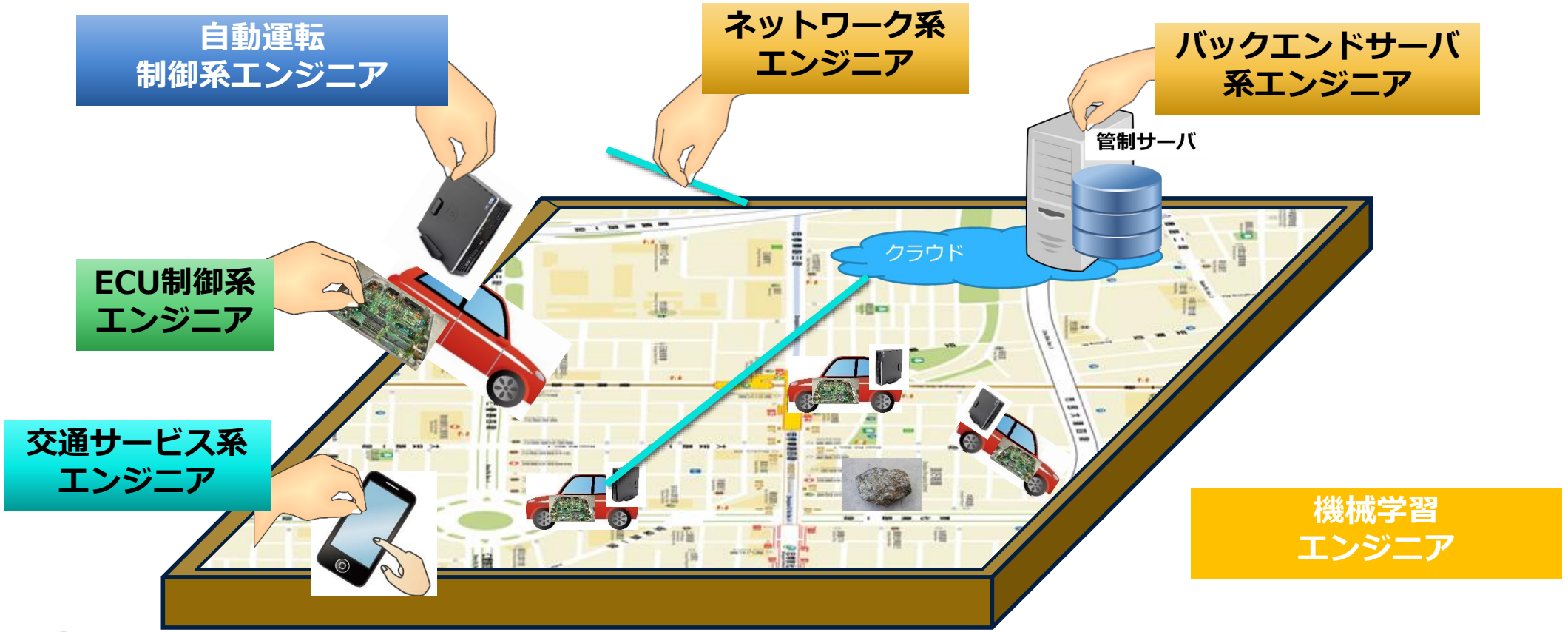
本研究の一部は，JST さきがけJP-MJCR21D2の支援により得られたものである.

強化学習と仮想環境

- ロボットのような実世界とのインタラクションを伴う学習には仮想環境が必要
- Gazebo，AirSim，Carlaなど，ROSや自動運転に特化したシミュレーション環境も多く存在する
- 要件に合わせて環境やシステム構成を作る必要があり，Unityで自作している例も見られる．特に多様なシステムが混載している場合，特定用途に特化したシミュレータでは対応が難しい

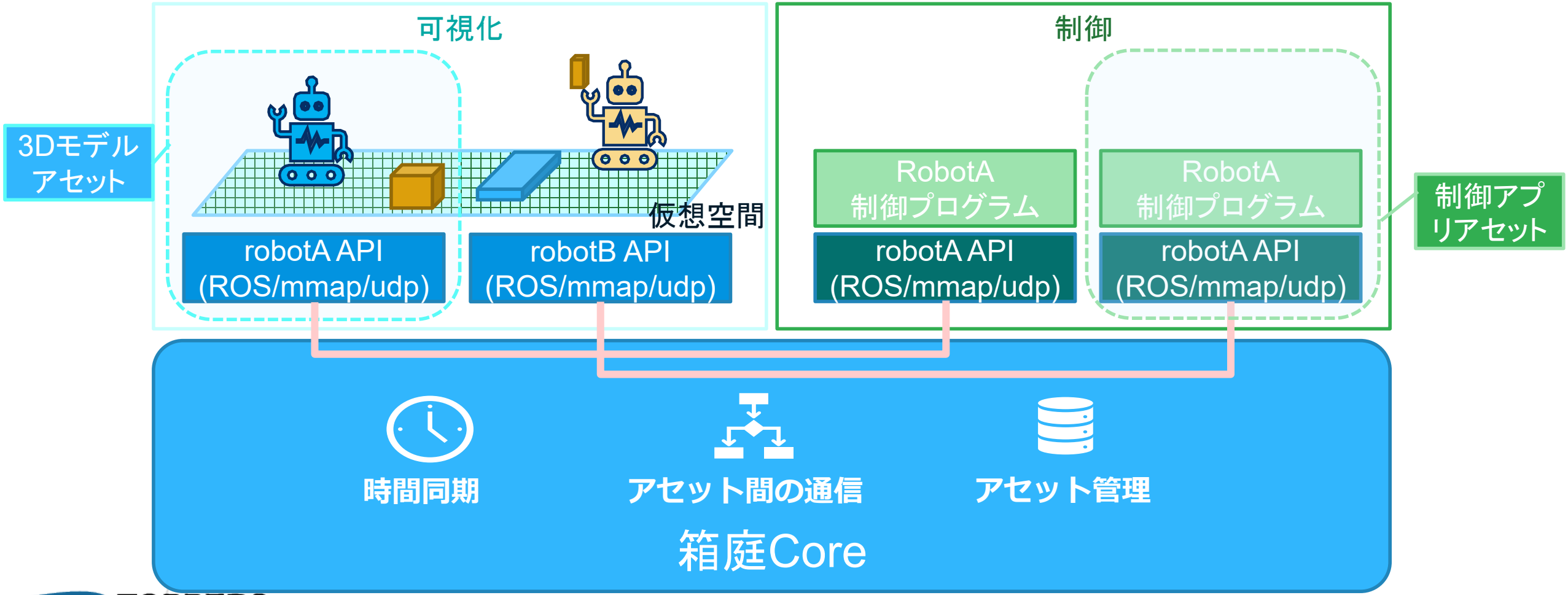
『箱庭』とは？ コンセプトと狙い

- 箱の中に，様々なモノをみんなの好みで配置して，いろいろ試せる！
  ⇒ 各技術者が開発対象と興味(＝アセット)を持ち寄って，机上で実証実験



箱庭のしくみ

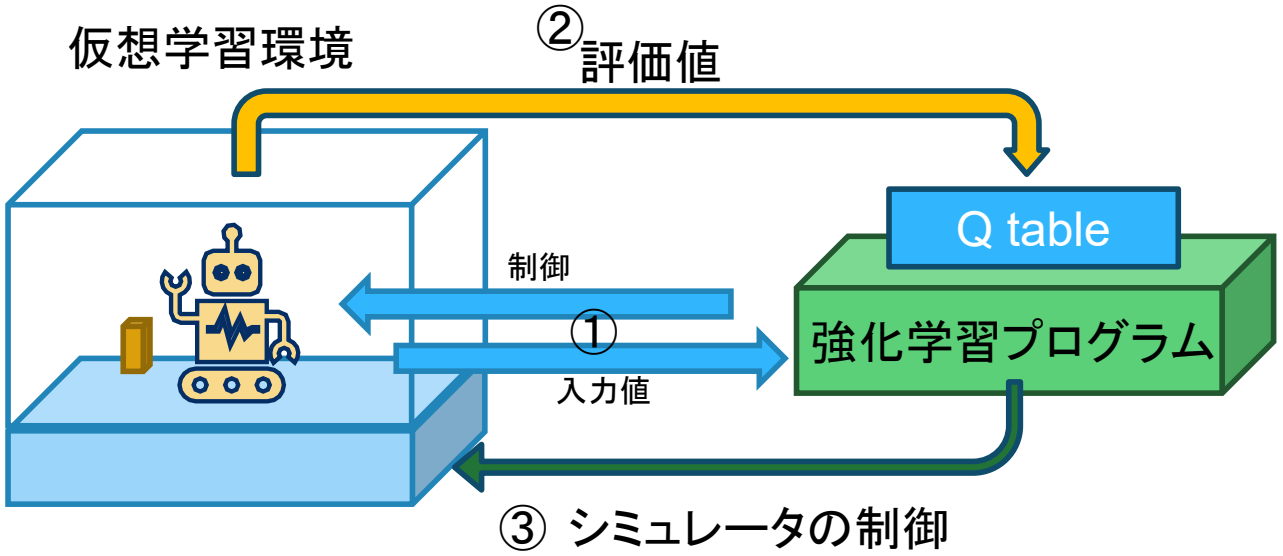
箱庭は、単なるシミュレータではなくシミュレータを作るためのフレームワーク



『箱庭』と強化学習

- 学習済みモデルから制御するだけであれば，制御プログラムを機械学習のプログラムに差し替えるだけ

- 学習のためには仕組みが必要
  - ①プログラム連携
  - ②シミュレーション情報の利用
  - ③シミュレータの制御



・強化学習のためには，評価関数が必要
・繰り返し学習のために，シミュレータのリセット等の制御が必要

強化学習の実装事例

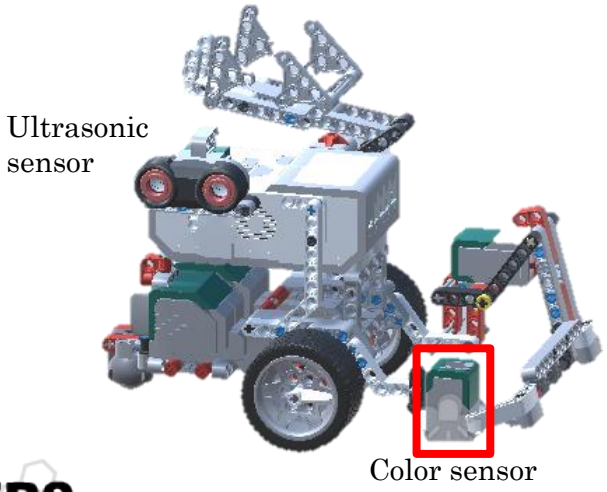
- 黒線に沿って走行するロボット
- Q学習を使ってライントレースを学習させたい
- 入力は前方下部のカラーセンサ，出力は左右のモータ2つ
- 距離センサーは制御には用いず，即時報酬の値として利用

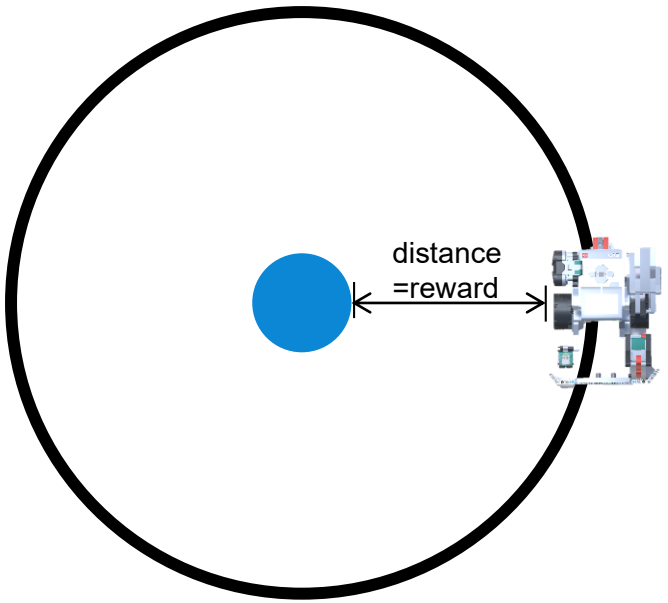
s:state, a:action,
α:学習率
γ:割引率

Qテーブルの更新式

$$Q(s,a) = (1-\alpha)\cdot Q(s,a) + \alpha \cdot \left(r + \gamma \max_{a'} Q(s',a')\right)$$

上記に加えてε-greedy法で 確率εで別行動を取らせる

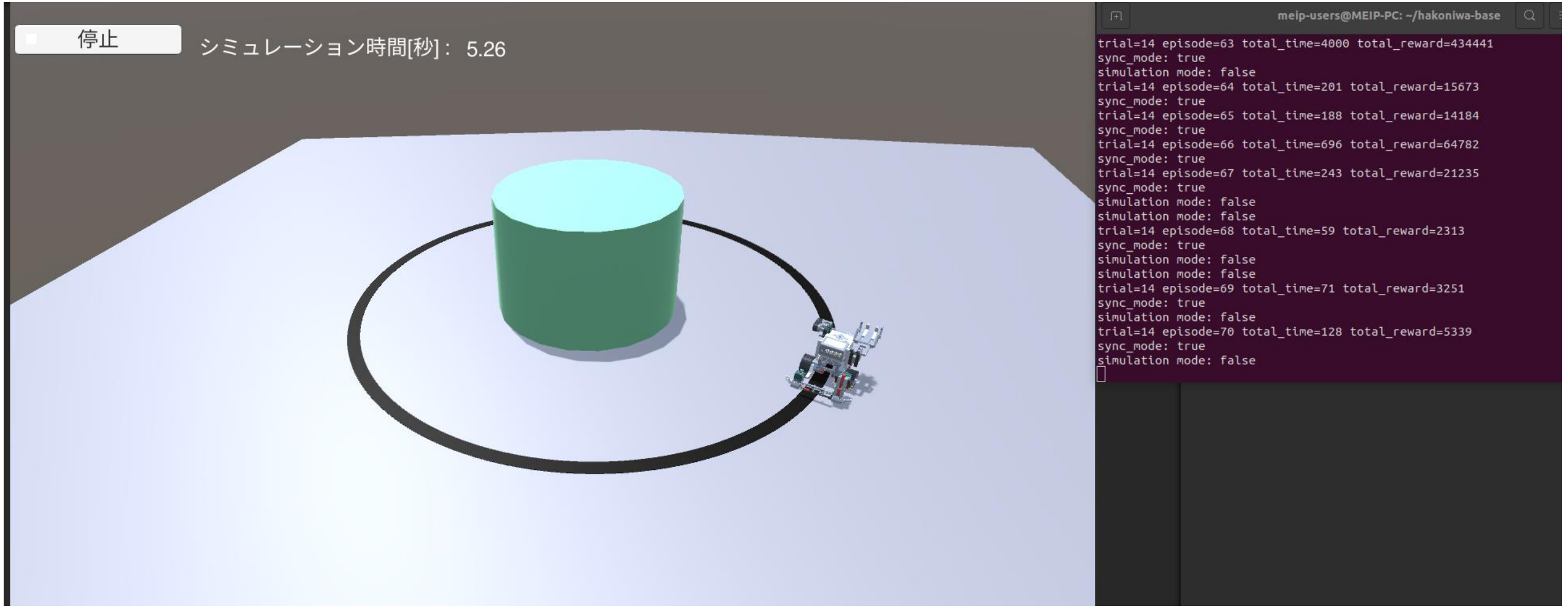




State
4 states derived from dividing the grayscale values of the color sensor.

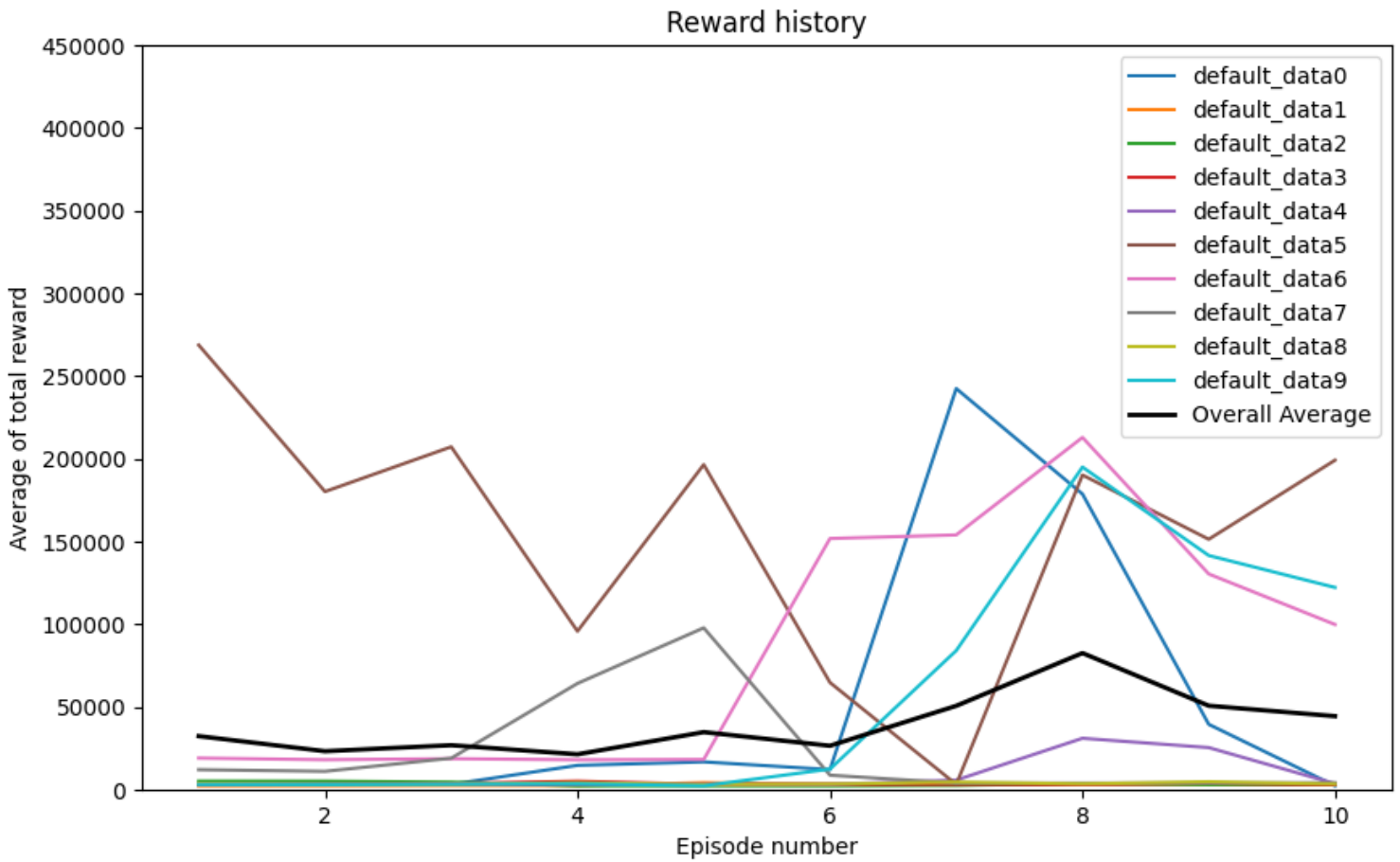
Action
6 actions in total, comprising forward movement, right and left turns, each with two levels of intensity.

学習中の様子


停止
シミュレーション時間[秒]： 5.26
meip-users@MEIP-PC: ~/hakoniwa-base
trial=14 episode=63 total_time=4000 total_reward=434441
sync_mode: true
simulation mode: false
trial=14 episode=64 total_time=201 total_reward=15673
sync_mode: true
trial=14 episode=65 total_time=188 total_reward=14184
sync_mode: true
trial=14 episode=66 total_time=696 total_reward=64782
sync_mode: true
trial=14 episode=67 total_time=243 total_reward=21235
sync_mode: true
simulation mode: false
simulation mode: false
trial=14 episode=68 total_time=59 total_reward=2313
sync_mode: true
simulation mode: false
simulation mode: false
trial=14 episode=69 total_time=71 total_reward=3251
sync_mode: true
simulation mode: false
trial=14 episode=70 total_time=128 total_reward=5339
sync_mode: true
simulation mode: false

評価結果①

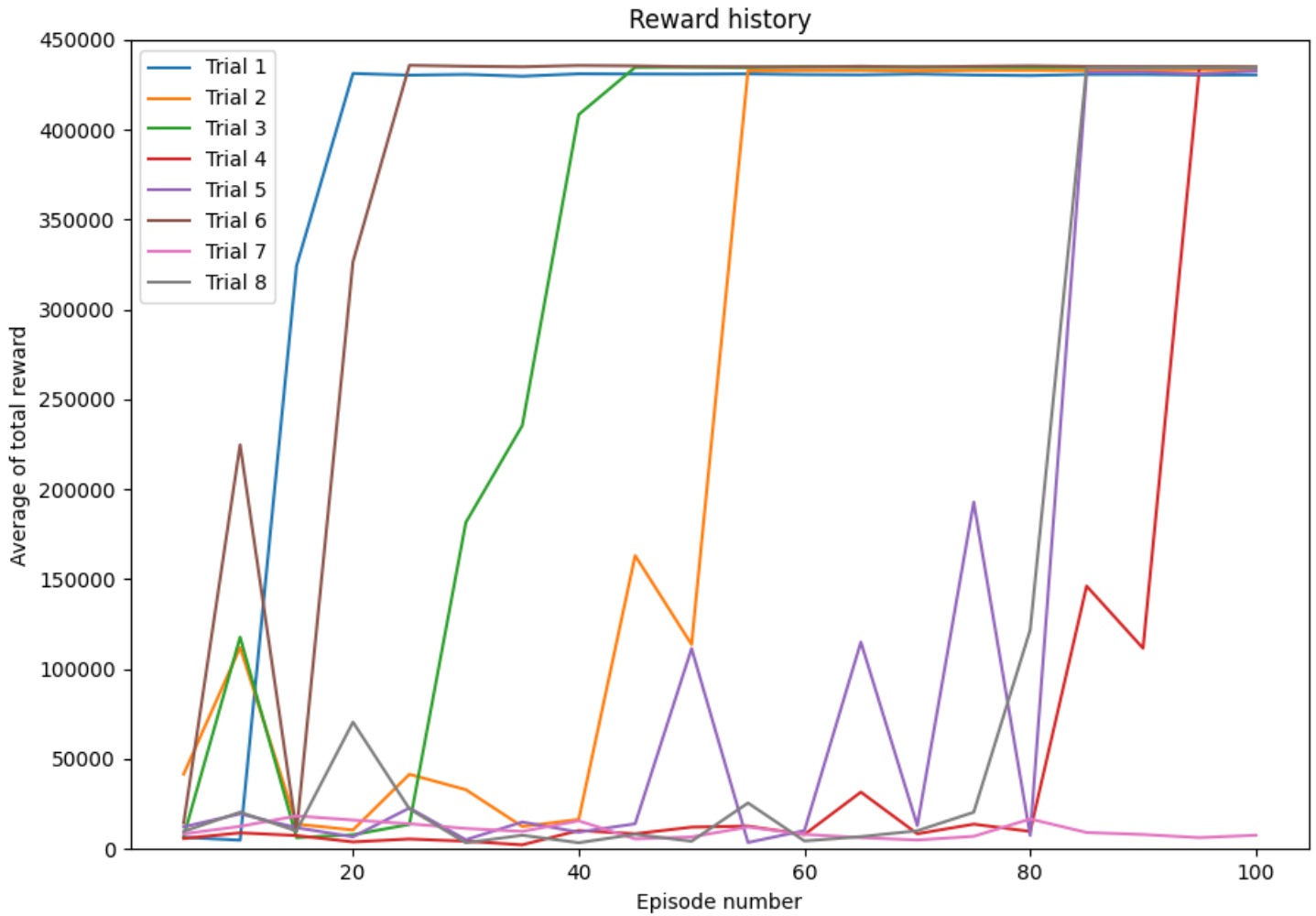
- 初回に定めたパラメータでは，学習を繰り返しても最終報酬（走行距離）は大きく変わらなかった.
- 初期値に大きく影響を受けている.
- 学習が進むにつれ学習率αと，ε-greedyのεを下げる処理を導入した.



評価結果②

- エピソードが進むにつれ，報酬が増加していることが確認できた.

- 多くの試行で最大値(円周上をずっと周回)付近まで学習できている



おわりに

- 仮想シミュレーション環境「箱庭」と強化学習を連携させ，例題を用いて検証した.

- 連携に必要となる機能を検討し分離したことで，同様の事例についてもシミュレーション上で機械学習を容易に実現できることが期待できる.

Web：https://toppers.github.io/hakoniwa/

Github：https://github.com/toppers/hakoniwa


箱
庭