# Gamification to Advance Reinforcement Learning


Student: Daniel Nagura

Student Email: dn946718@ohio.edu

Faculty: Jundong Liu

Faculty Email: liuj1@ohio.edu

AFRL Sponsor: Trevor Bihl

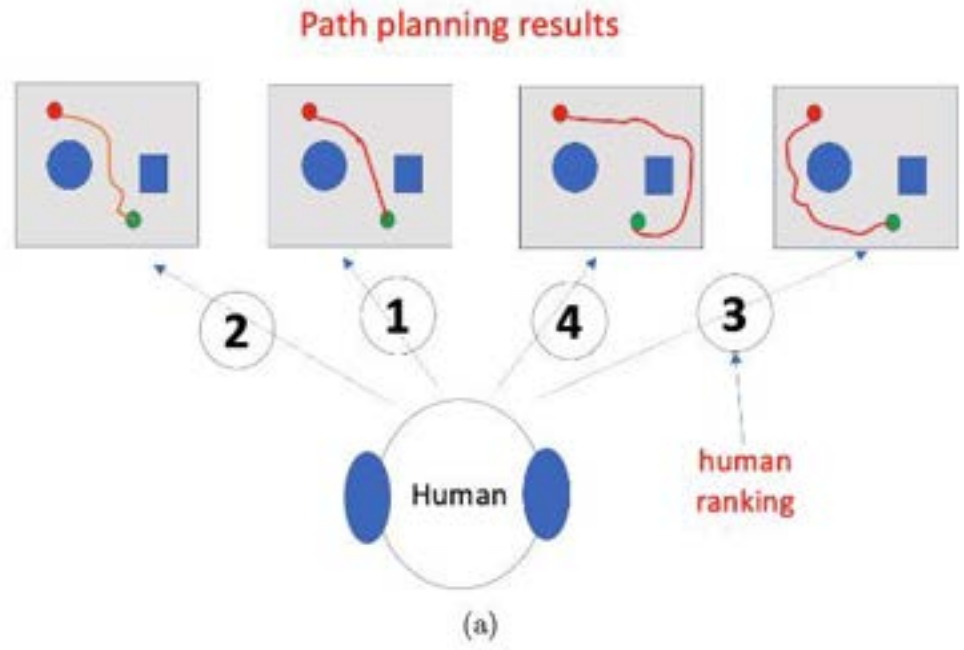
AFRL Directorate: AFRL/RY



PA #: AFRL-2024-6019

# Objective: Gamification to Advance Reinforcement Learning

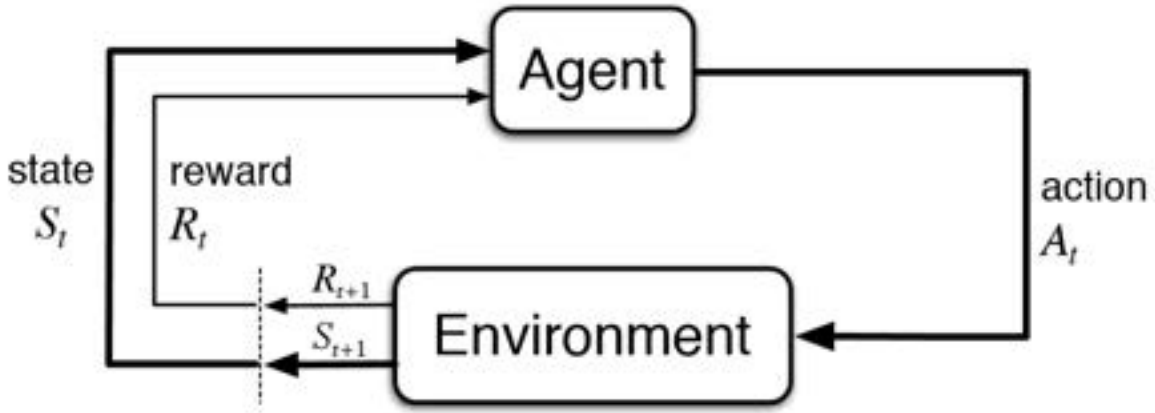
- Proposed plan items
  - User provides feedback, and *reinforcement learning from human feedback* (RLHF) technique will be used to improve the motion planning algorithm.
  - Gamification to improve user engagement.




- Updated additional item
  - Integrate human plays to improve RL policy learning.

# Reinforcement Learning (RL)

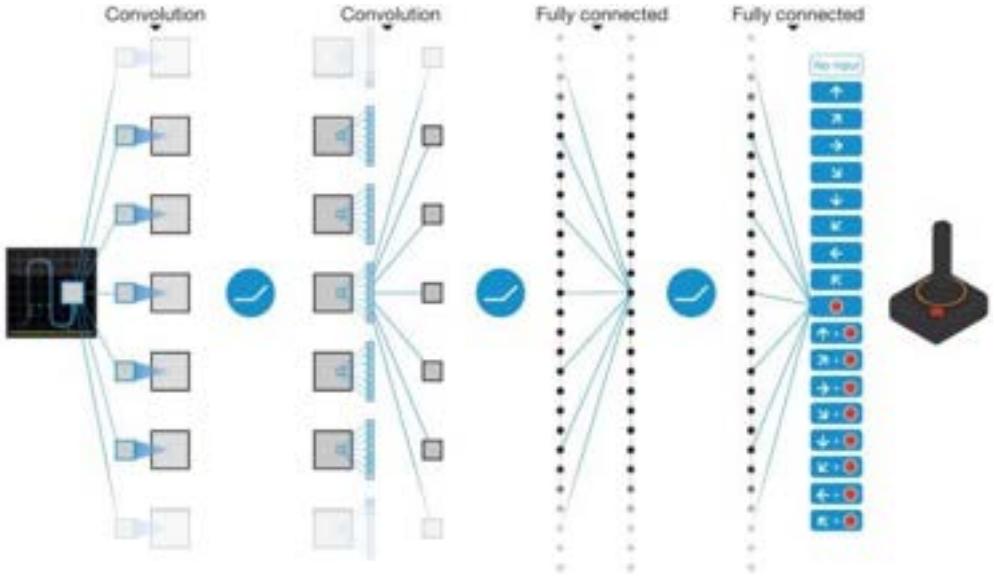
- RL: train an agent to make decisions (policy) in an environment to maximize cumulative reward.
  - State $S_t$
  - Action $A_t$
  - Reward $R_t$




- RL vs. supervised learning

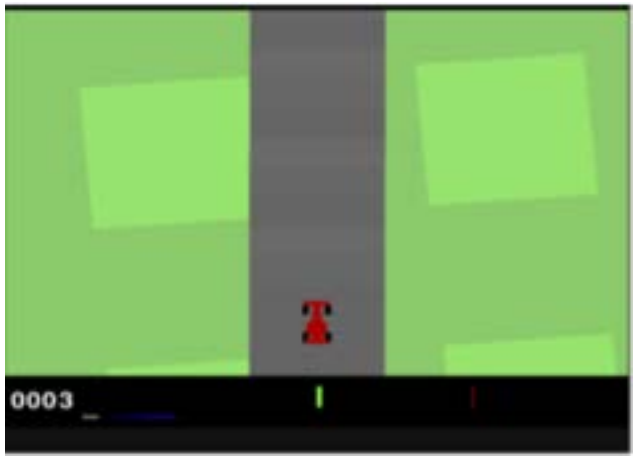
# Deep Reinforcement Learning (DRL)

- In practice, the states could be high-dimensional, e.g., images from camera.
- The basic idea of DRL is to use neural networks, as approximation functions, to model and learn the state and action interactions.
- Successful stories: AlaphGo (20160, AlaphDogFight (2019), ChatGPT (2022).

# Race Car Environment (RCE) and our RL setup

- Car Racing environment (CRE) originally created by OpenAI Gym, which uses the Box2D physics engine.
- The racing car's objective is to complete a lap and score the highest number of points.



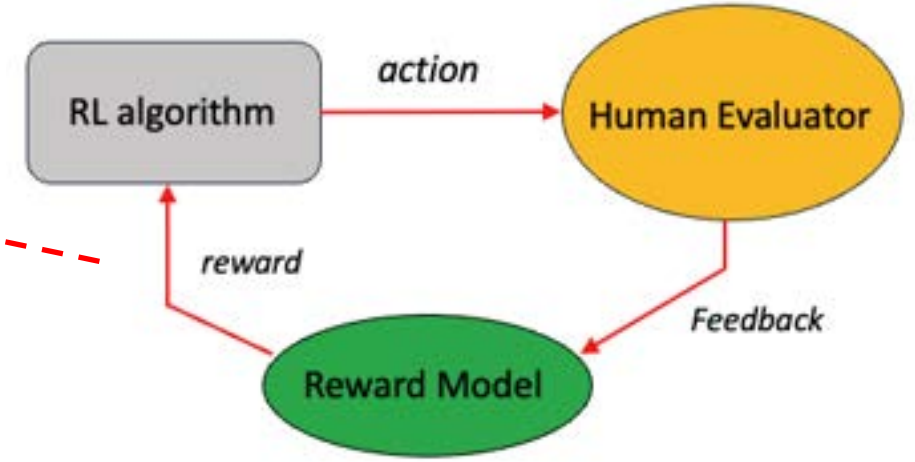

- Our RL setup:
  - State $s_t$: normalized gray-valued image of the current patch of size 96×96
  - Action $a_t$: sampled from the action space
  - Reward function $r_t$:
    - Rewarded for every track segment (a short distance) visited --- encourage following the track
    - Every time-step rewards -0.1 points and -0.2 points when at least one wheel is outside the road.

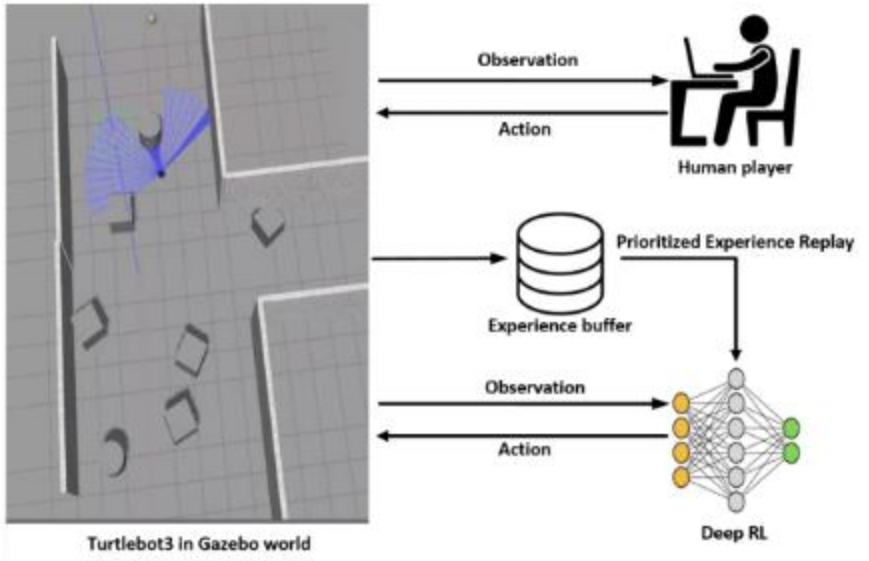
# Project progress

1. Stage 1: RL from AI feedback (RLAIF)
   - Completed. Method and results have been published in NAECON'24 conference.

2. Stage 2: RL with integration of human experience (RLHE)
   - Completed.

3. Stage 3: Multi-player RL and gamification
   - Nov. 2024 – March 2025.
   - Results from 2) and 3) will be submitted for journal publications.

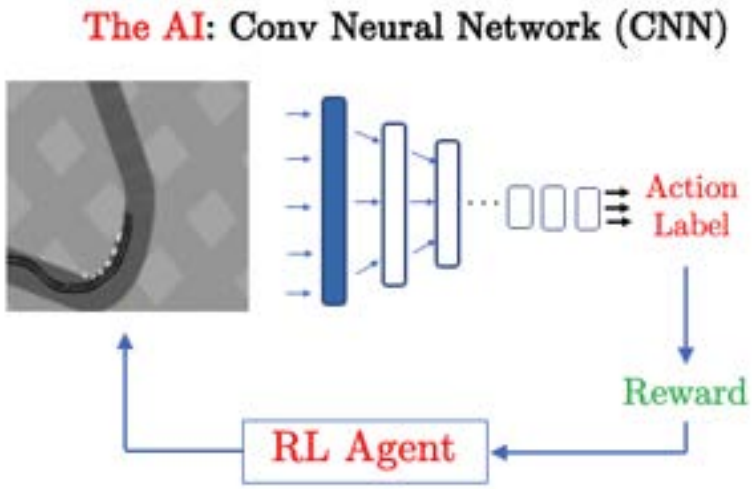


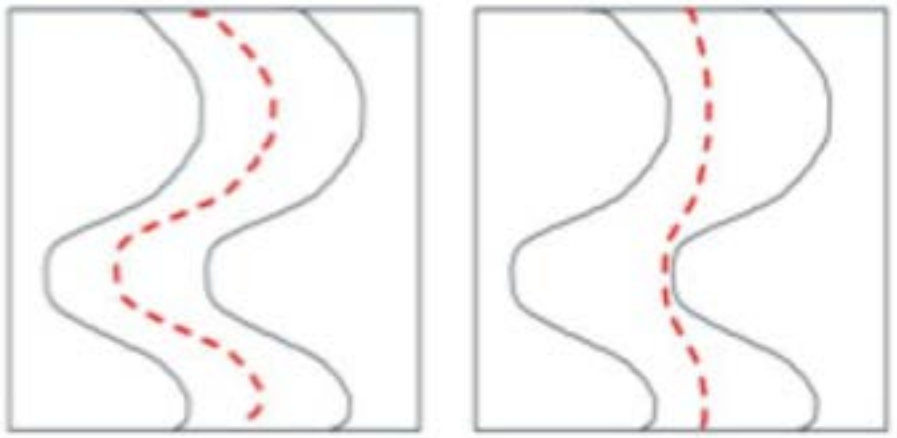

From Niu 2021 et. al. "Accelerated sim-to-real DRL.

# RL from AI Feedback (RLAIF) [Nagura et al., NAECON'24]

- The "AI" is designed as a convolutional neural network (CNN).




- Preferences can be set differently, ***for example***, "staying in the middle of the road" or "running straightly within the boundary, therefore faster".



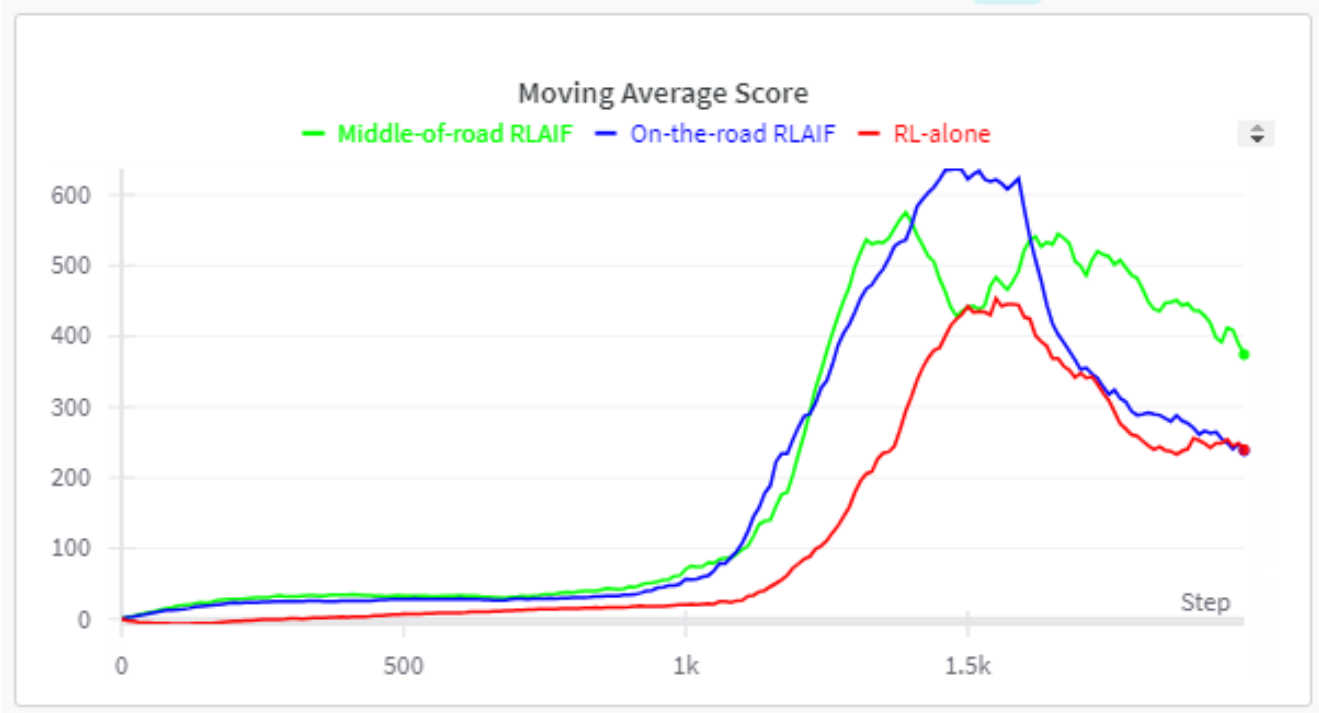
Preference of (a) **stay in the middle of the road**; (b) **run straightly**.

# RLAIF [Nagura et al., NAECON'24]

- Scores from testing the RL models:

| | 5-Class CNN | Stay-in-road | Middle-of-road |
|---|---|---|---|
| RL-alone | 216.15 | 176.41 | 98.15 |
| RLAIF | 394.50 | 248.56 | 194.15 |

Column: evaluation CNNs;
Row: trained RL models.




Moving average scores for the rewards: The models are middle-of-road RLAIF (green), on-the-road RLAIF (blue) and RL-alone (red), respectively.

# RL with Human Experience (RLHE)

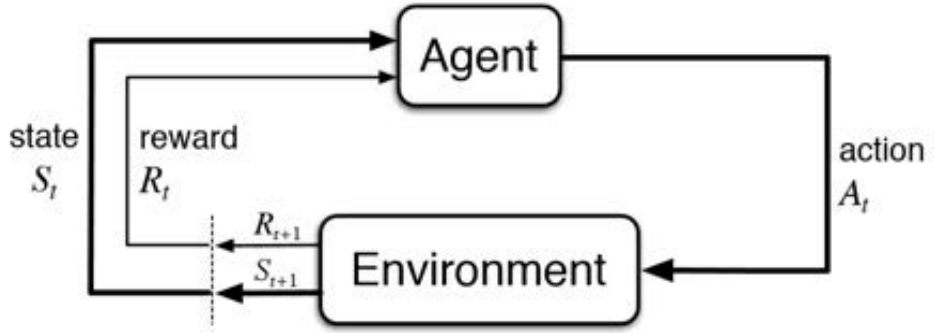
- A human plays with the environment and generate plays to be integrated into the experience replay buffer.
- The experience replay buffer is then used in PPO model to enhance the learning procedure of the agent.

logp denotes logarithm probability.

$(s_t, a_t, r_t, s_{t+1})$

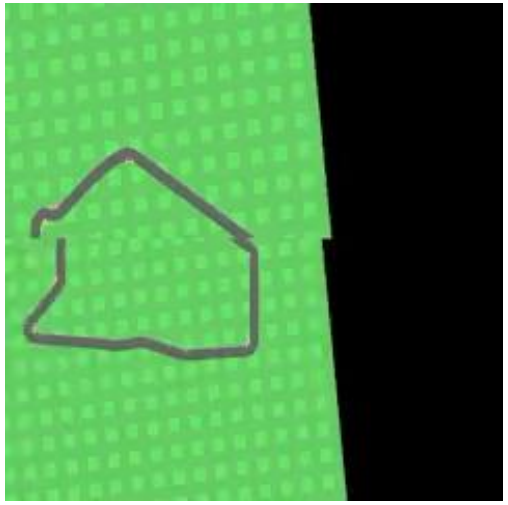
$a_{\text{logp}}(s_t)$

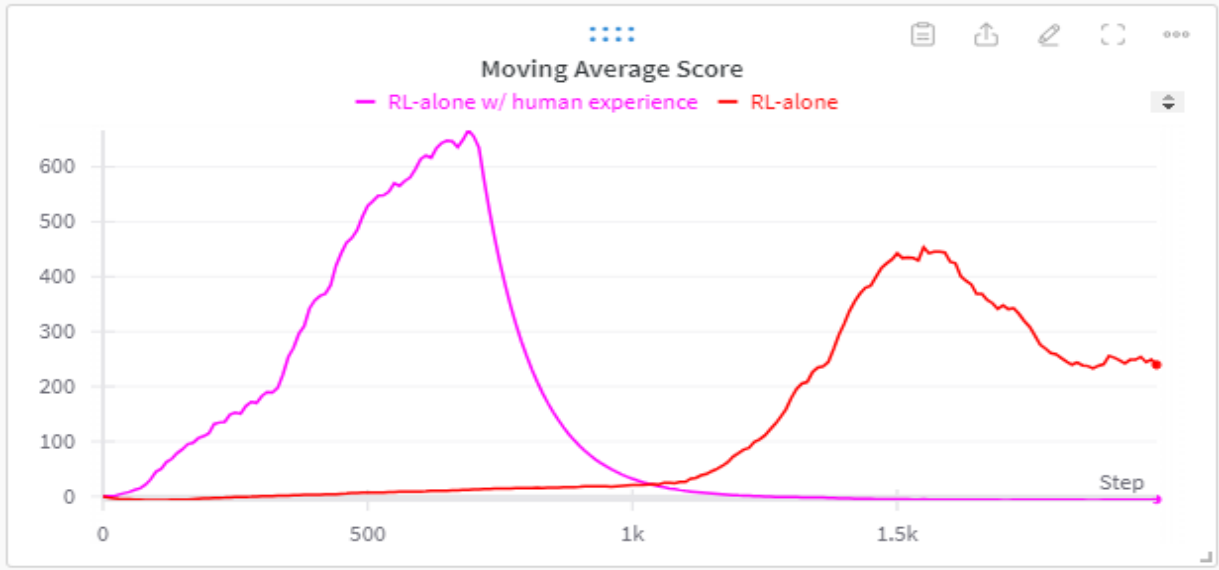
$(s_t, a_t, a_{\text{logp}}, r_t, s_{t+1})$



# RLHE: Results

- Car motion: RL-alone (left) vs. RLHE (right)






Training Procedure: RLHE (purple) vs RL-alone (red). X-axis: epoch; Y-axis: score.

- Results

| | No RLAIF | 5-Class | Stay-in-road | Middle-of-road |
|---|---|---|---|---|
| RL-alone | 515.21 | 508.81 | 527.31 | 440.61 |
| RLHE | 859.81 | 928.61 | 875.15 | 832.69 |

Table: Column: evaluation CNNs, Row: trained models.