Hierarchical Reinforcement Learning for **Air-to-Air Combat**

Sungkwon On
05 – Dec – 2022

Abstract

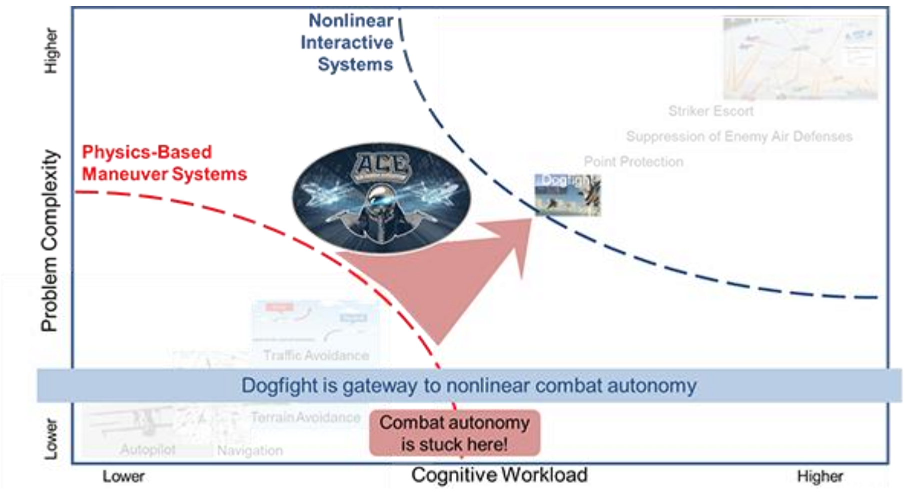
- Artificial Intelligence is becoming a critical component in the defence industry.
- High-fidelity open-source flight dynamics model&simulator available for training

Air Combat Evolution(ACE) program

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

ACE program seeks to increase trust in combat autonomy.
ACE program addresses four primary challenges:

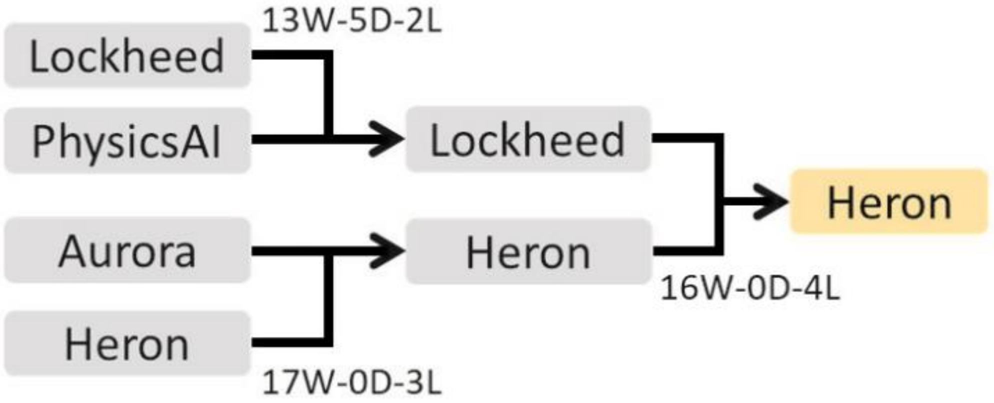
1. Increase air combat autonomy performance in local behaviours (individual aircraft and team tactical)
2. Build and calibrate trust in air combat local behaviours
3. Scale performance and trust to global behaviours (heterogeneous multi-aircraft)
4. Build infrastructure for full-scale air combat experimentation

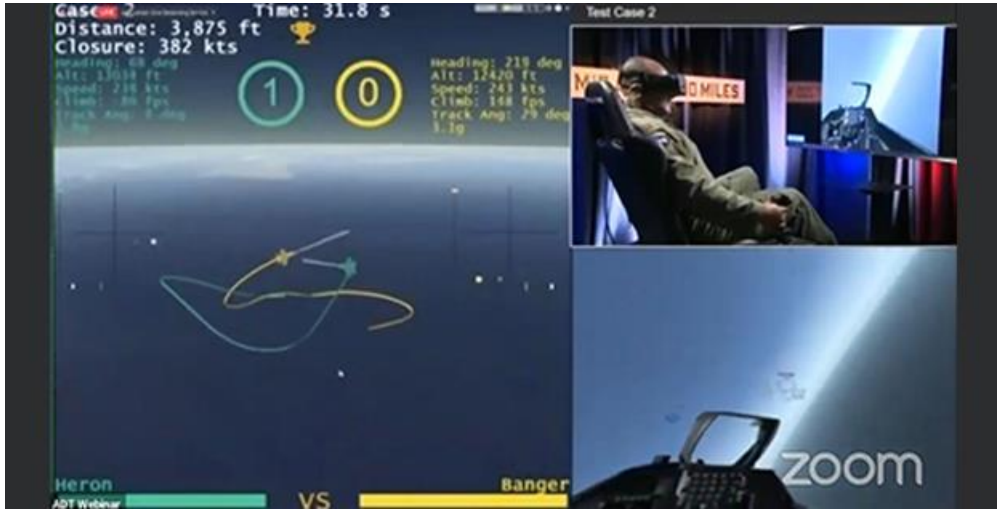


Alpha Dogfight Trials(ADT)

2020August

Virtual finale showcases AI's impressive abilities in simulated F-16 aerial combat



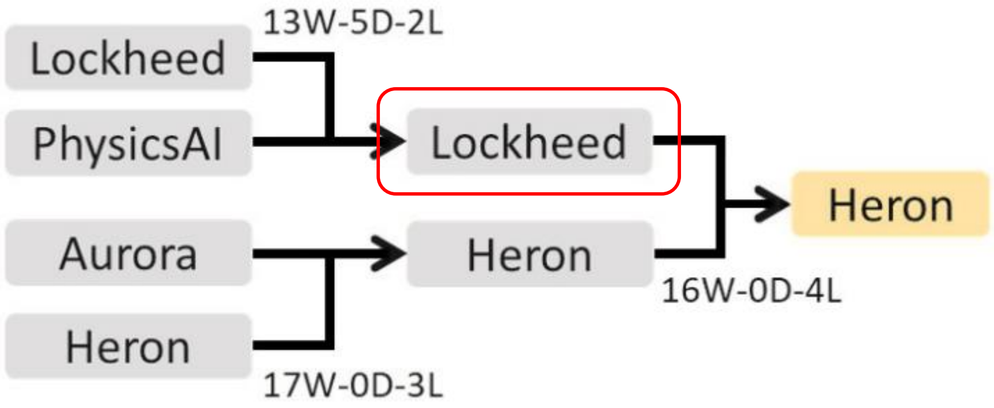


Hierarchical Reinforcement Learning for Air-to-Air Combat

Adrian P. Pope*, Jaime S. Ide*, Daria Mićović, Henry Diaz, David Rosenbluth,
Lee Ritholtz, Jason C. Twedt, Thayne T. Walker, Kevin Alcedo and Daniel Javorsek II[†]
Applied AI Team, Lockheed Martin, Connecticut, USA
[†]U.S. Airforce, Virginia, USA

2021 June

- Achieved 2nd place in the final tournament of ADT (among 8 total competitors)
- Defeated a graduate of the US Air Force’s F-16 Weapons Instructor Course in match play (5W-0L)
- Used Hierarchical RL structure by integrating expert knowledge for reward shaping



C/C++ build passing | pypi v1.1.7
pypi downloads 69k | conda-forge v1.1.7

JSBSim
Open Source Flight Dynamics

Environment

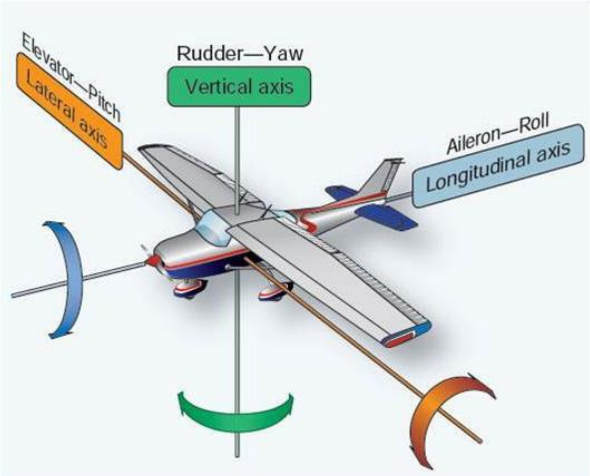
Observation space:

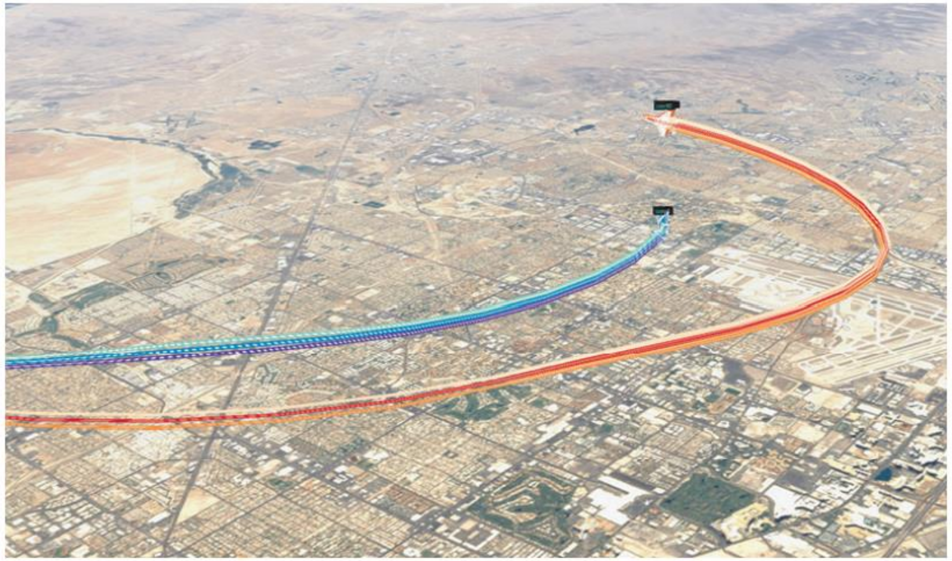
- Ownship info(fuel load, thrust, control surface deflection, health)
- Aerodynamics(alpha and beta angles)
- Position(local plane coordinates, velocity, acceleration) of ownship & opponent
- Altitude of ownship & opponent

(All state provided without noise)

Action space:

- Aileron
- Elevator
- Rudder
- Throttle

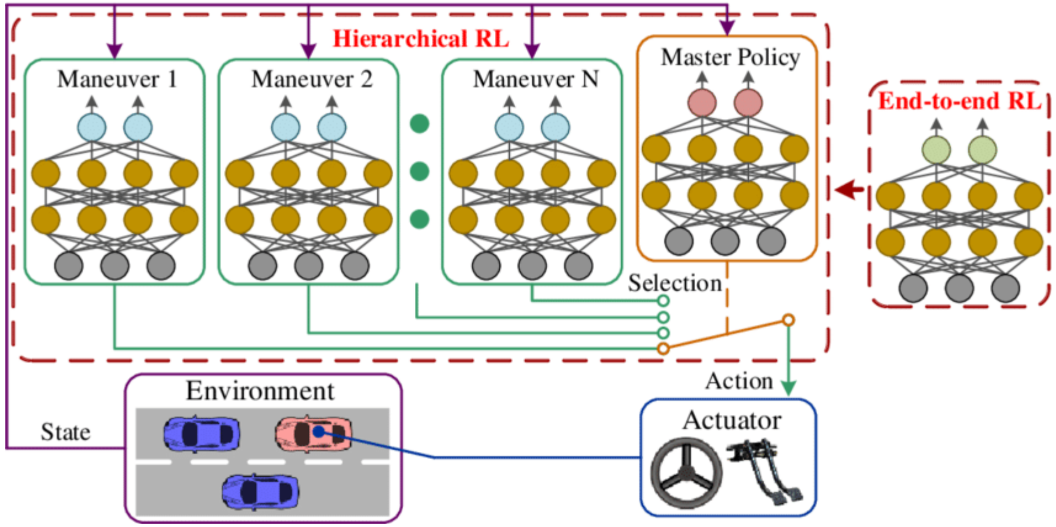


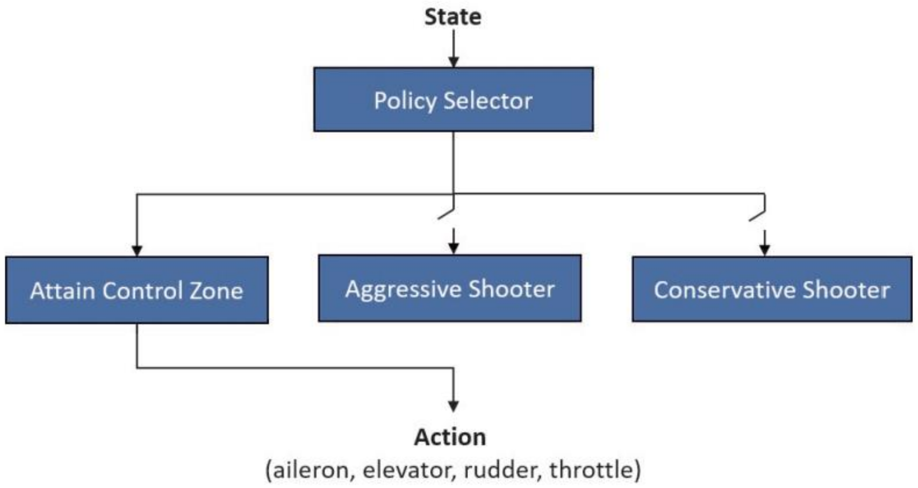


Hierarchical Reinforcement Learning

Divides a complex task into smaller sub-tasks by having different policies in layered-form.

The “Policy Selector” policy selects which one of the sub-policies executes an action.





Soft Actor-Critic (SAC)

$$J_\alpha = \mathbb{E}_{a_t \sim \pi_t}[-\alpha \log \pi_t(a_t \mid s_t) - \alpha \mathcal{H}_0]$$

Off-policy actor-critic RL method

Added entropy(of state) term increases exploration during training

---

Algorithm 1: Soft Actor Critic

Initialize Q, policy and α network parameters;
Initialize the target Q-network weights;
Initialize the replay buffer $\mathcal{D}$;
for *each episode* **do**
  for *each environment step* **do**
    Sample the action from the policy $\pi(a_t|s_t)$, get the next state s_{t+1} and reward r_t from the environment, and push the tuple (s_t, a_t, r_t, s_{t+1}) to $\mathcal{D}$;
  end
  for *each gradient step* **do**
    Sample a batch of memories from $\mathcal{D}$ and update the Q-network (Equation 6), the policy (Equation 7), the temperature parameter α (Equation 8), and the target network weights (soft-update).
  end
end

Weapon engagement zone (WEZ)

WEZ: locus of points that lie within a spherical cone of 2 degree aperture, which extends out of the nose of the plane, that are also 500-3000ft away.

WEZ is also thought to be the position where when engaging weapons(eg, taking shots), there is a high chance of shooting the enemy plane down.

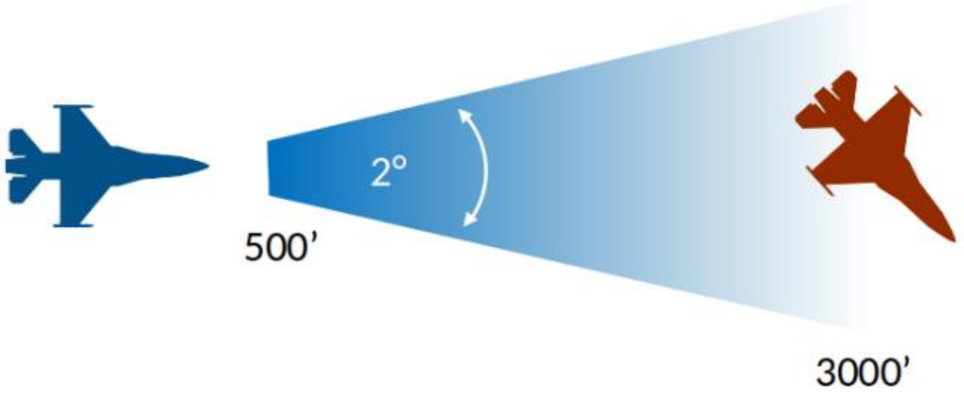
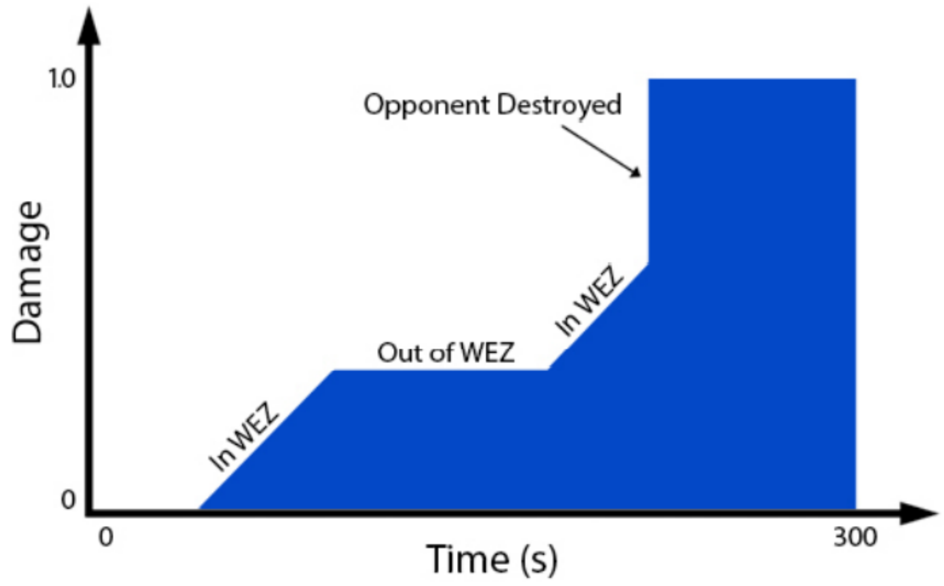


Fig. 2: Weapon Engagement Zone (WEZ)

Reward (Policy Selector)

$$d_{wez} = \begin{cases} 0 & r > 3000ft \\ \frac{3000-r}{2500} & 500ft \leq r \leq 3000ft \\ 0 & r < 500ft \end{cases}$$

$$r_t = \begin{cases} \mathbb{E}_{t' \in [0,T]}[d_{opp}(t')] & d_{self} < 1 \\ 0 & otherwise \end{cases}$$



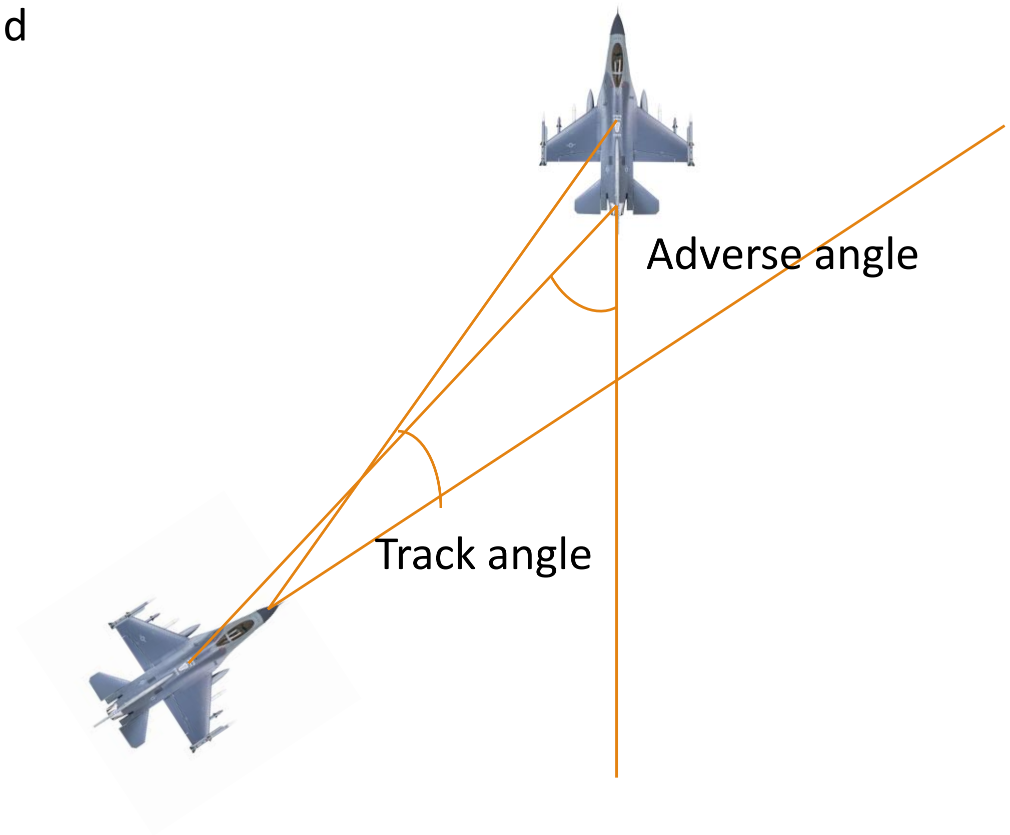
Episode ends when one of the aircrafts' damage is greater than 1 or time step reaches max, T=300

- Win if $d_{opp} > 1$
- Lose if $d_{self} > 1$
- Draw when the time step reaches T=300

Rewards (Low Level Policies)

$\boldsymbol{R_{relative\ position}}$ rewards the agent for positioning itself behind the opponent
$\boldsymbol{R_{track\ \theta}}$ penalizes the agent for having a non-zero track angle
$\boldsymbol{R_{closure}}$ rewards the agent for getting closer to the opponent
$\boldsymbol{R_{gunsnap(blue)}}$ rewards for achieving a minimum track angle and is within particular range
$\boldsymbol{R_{gunsnap(red)}}$ penalizes when opponent achieves a minimum track angle and is within particular range
$\boldsymbol{R_{deck}}$ penalizes when flying below minimum altitude (1000ft)
$\boldsymbol{R_{too\ close}}$ penalizes for violating a minimum distance within a range of adverse angles

PS: the agents do not actually take shots but the gunsnap rewards are given according to position and assumption of shooting

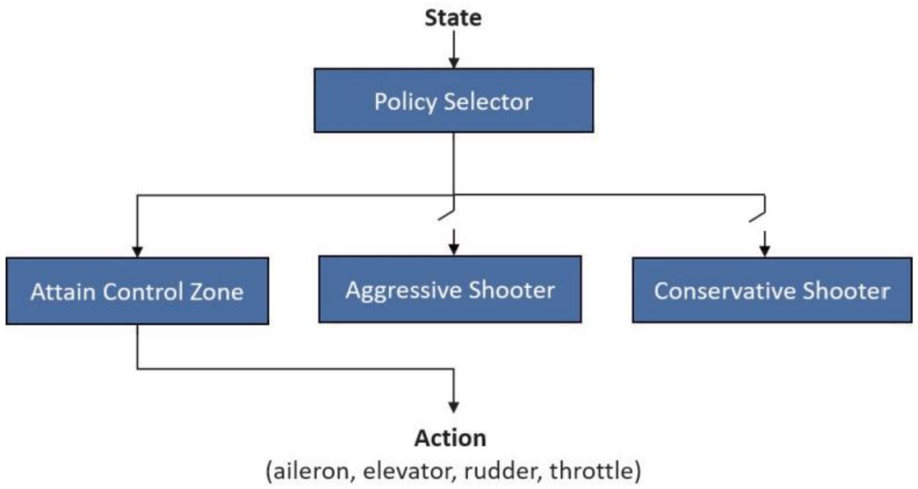


Low Level Policies

Control Zone (CZ): tries to attain a pursuit position behind the opponent (eg, WEZ)

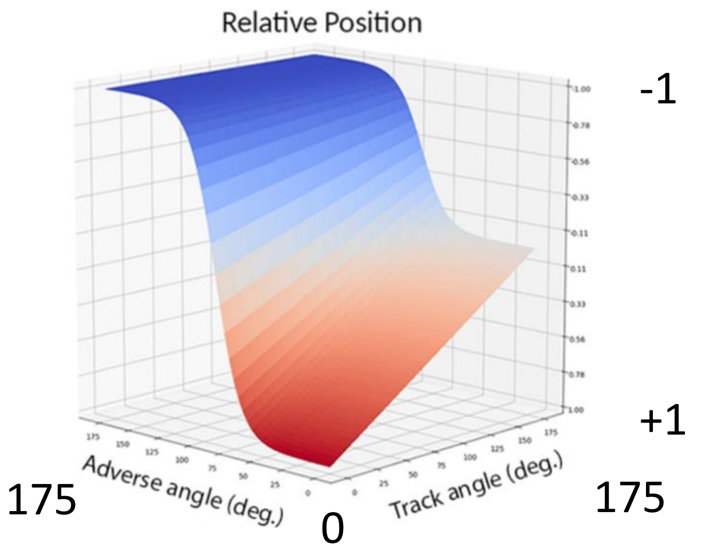
Aggressive Shooter (AS): Encourages to take aggressive shots. Gunsnap rewards are greater at a closer distance

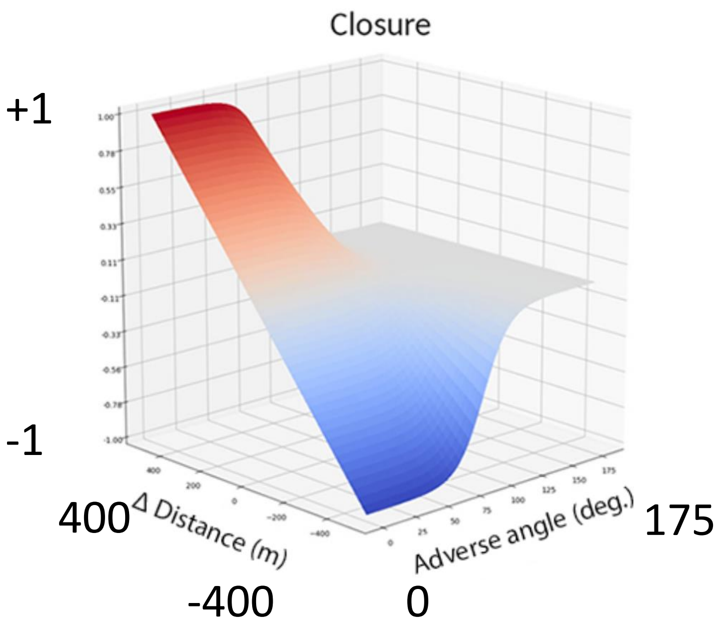
Conservative Shooter (CS): Values gunsnap from near and far equally → maintains an offensive scoring position

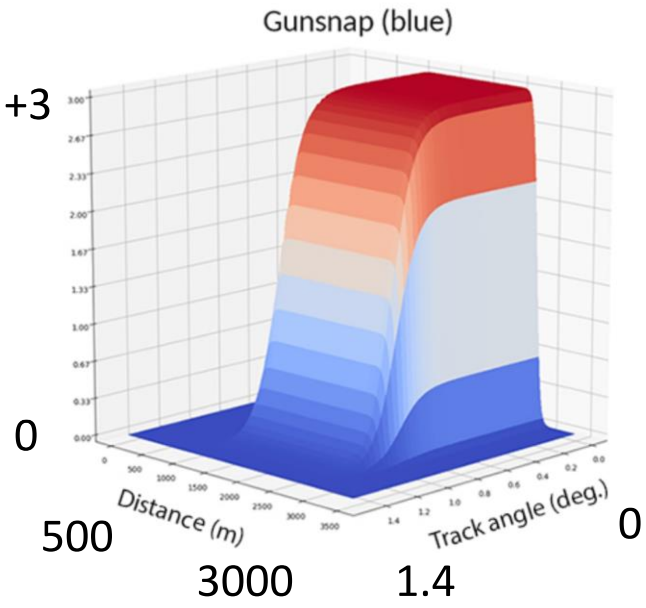


Control Zone (CZ)

$$R_{total} = R_{relative\ position} + R_{closure} + R_{gunsnap(blue)} + R_{gunsnap(red)} + R_{deck} + R_{too\ close}$$

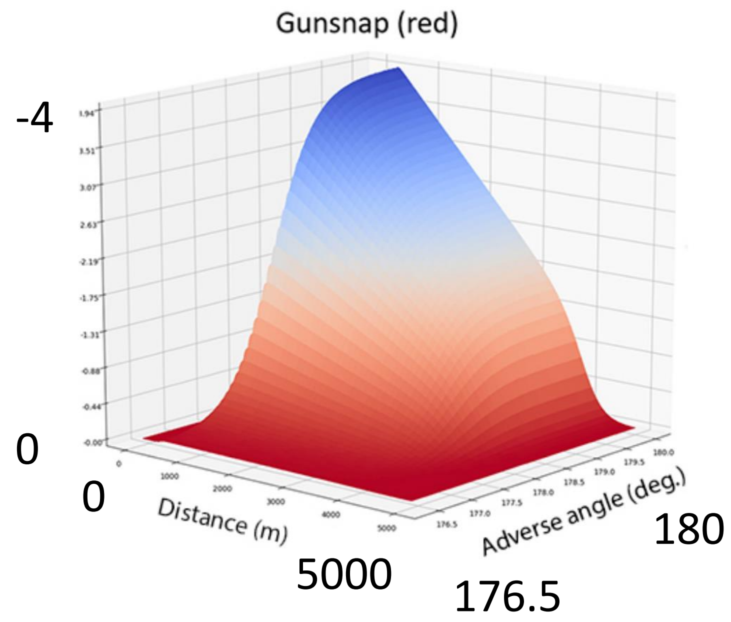


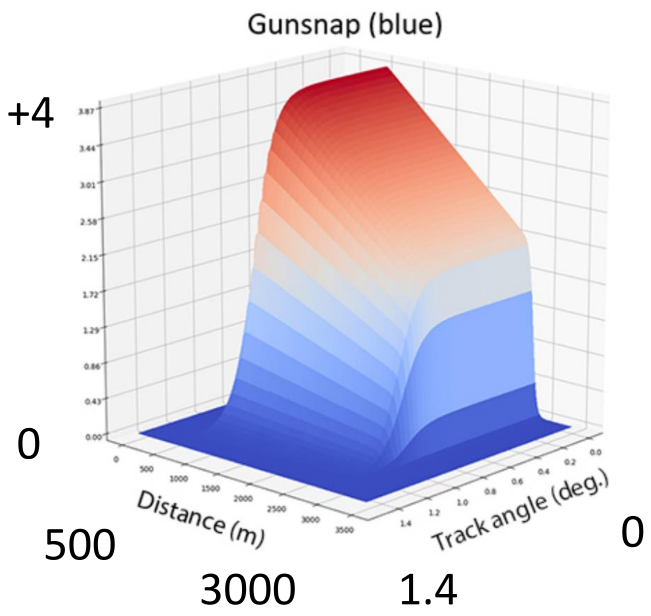




Aggressive Shooter (AS)

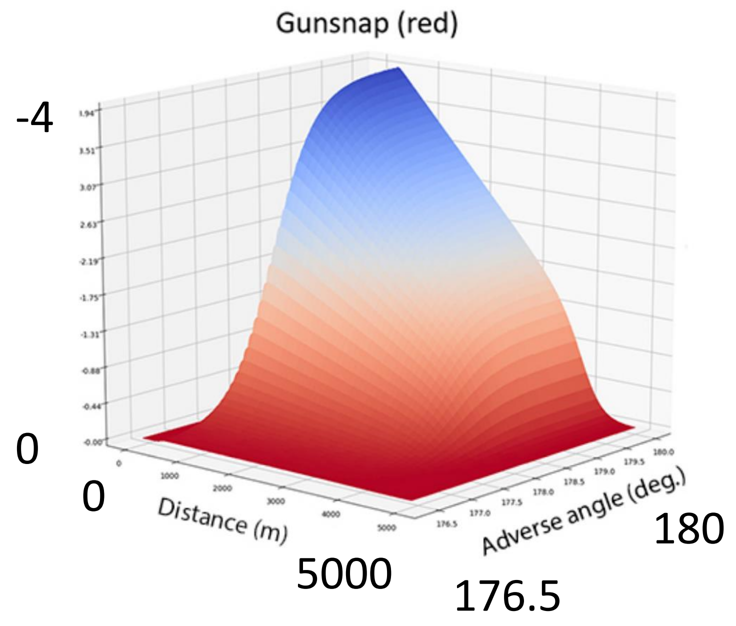
$$R_{total} = R_{track\ \theta} + R_{gunsnap(blue)} + R_{gunsnap(red)} + R_{deck}$$

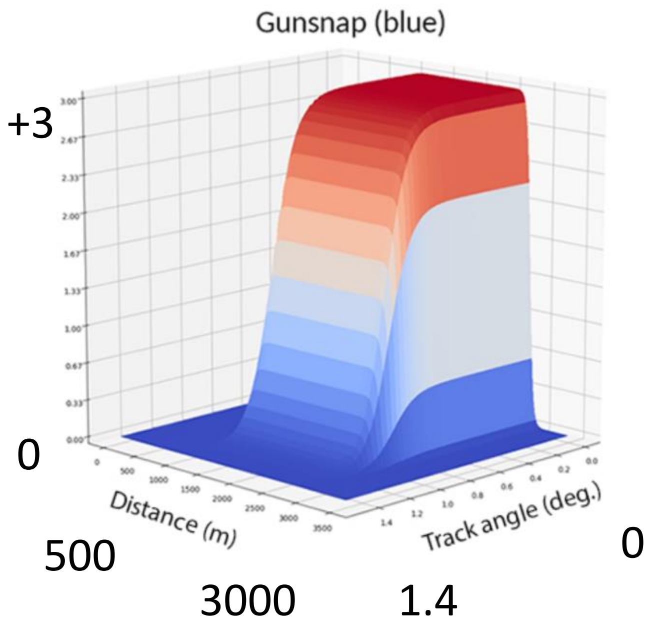




Conservative Shooter (CS)

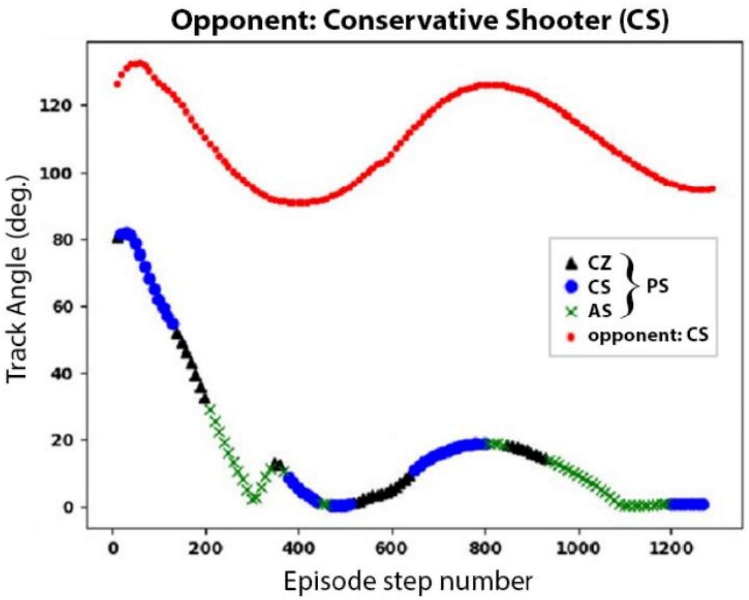
$$R_{total} = R_{track\ \theta} + R_{gunsnap(blue)} + R_{gunsnap(red)} + R_{deck}$$

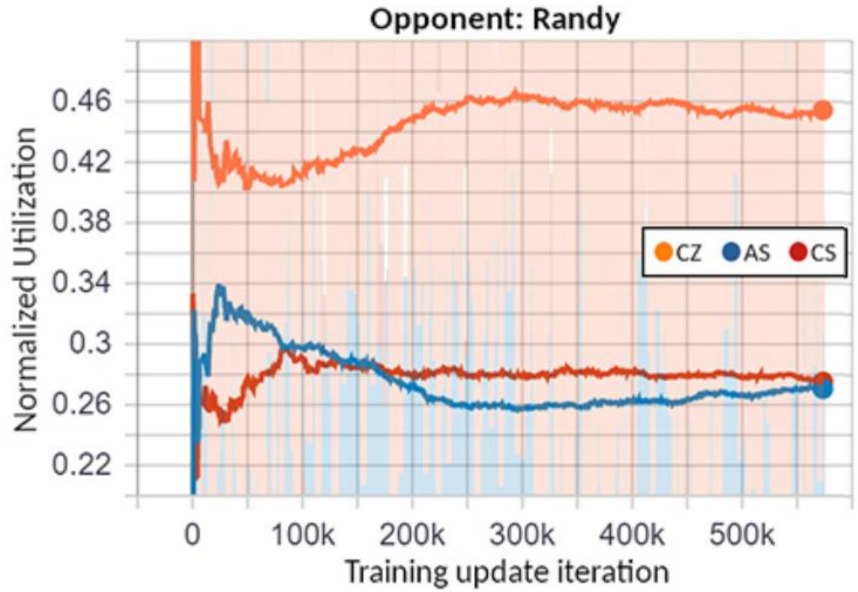


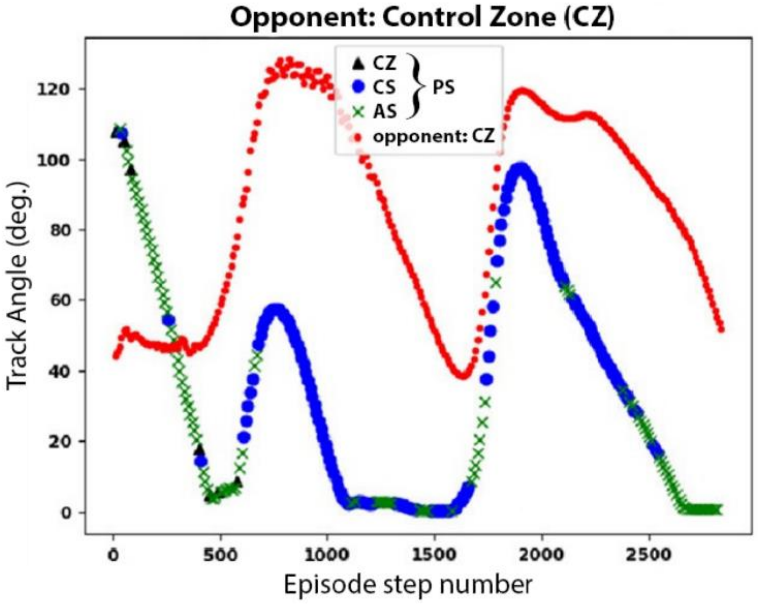


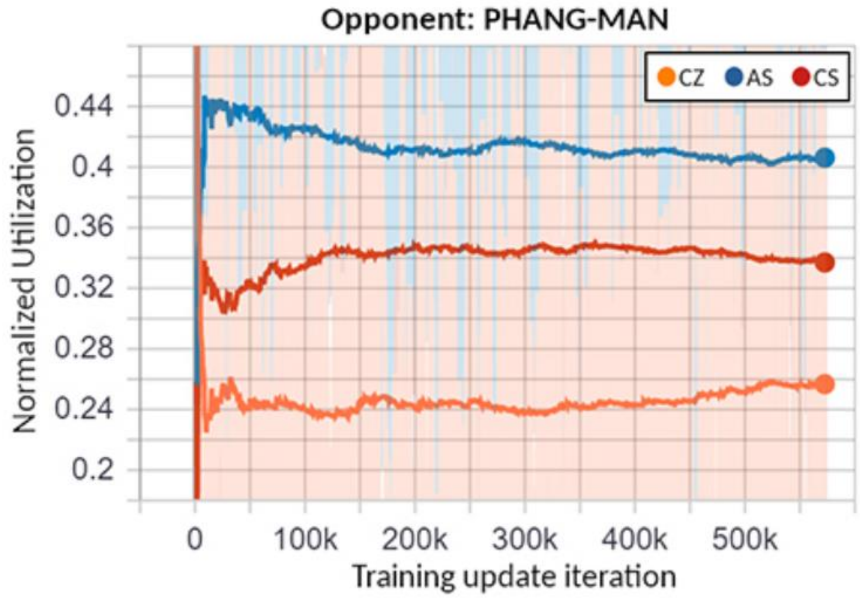
Policy Selector

- Similar to Option Learning
- A new selection of low-level policy is made periodically at a frequency of 10Hz.
- All three low-level policies are pre-trained and the parameters are frozen during the policy selector training.
- Policy selector trained via SAC.
- A reward proportional to track angle was included as well as the WEZ damage function
  - $r_t = r_{WEZ} + r_{track\ \theta}$


Opponent: Conservative Shooter (CS)
Track Angle (deg.)
Episode step number
CZ
CS
AS
PS
opponent: CS


Opponent: Randy
Normalized Utilization
Training update iteration
CZ
AS
CS


Opponent: Control Zone (CZ)
Track Angle (deg.)
Episode step number
CZ
CS
AS
PS
opponent: CZ


Opponent: PHANG-MAN
Normalized Utilization
Training update iteration
CZ
AS
CS

PHANG-MAN VS Heron

Result:

- PHANG-MAN did not survive most of initial exchange.
- 7% more total shots against Heron
- Average shots were further away → low average damage
- Agent disengaged its offence inside of 800ft, for better positioning for the next exchange
- Heron continued to aggressively pursue head on
- When survived in the initial exchange, PHANG-MAN attained commanding position.

Analysis:

- Artificially inflated agent's health by a factor of 10.
- Reward regarding opponents remaining health not provided.

Video of the dogfight: https://www.youtube.com/watch?v=NzdhIA2S35w watch from 2:40:50

Questions?