Fall 2022 - Winter Break Update

Mason Smith

January 31, 2023

Agenda

Updated Worlds

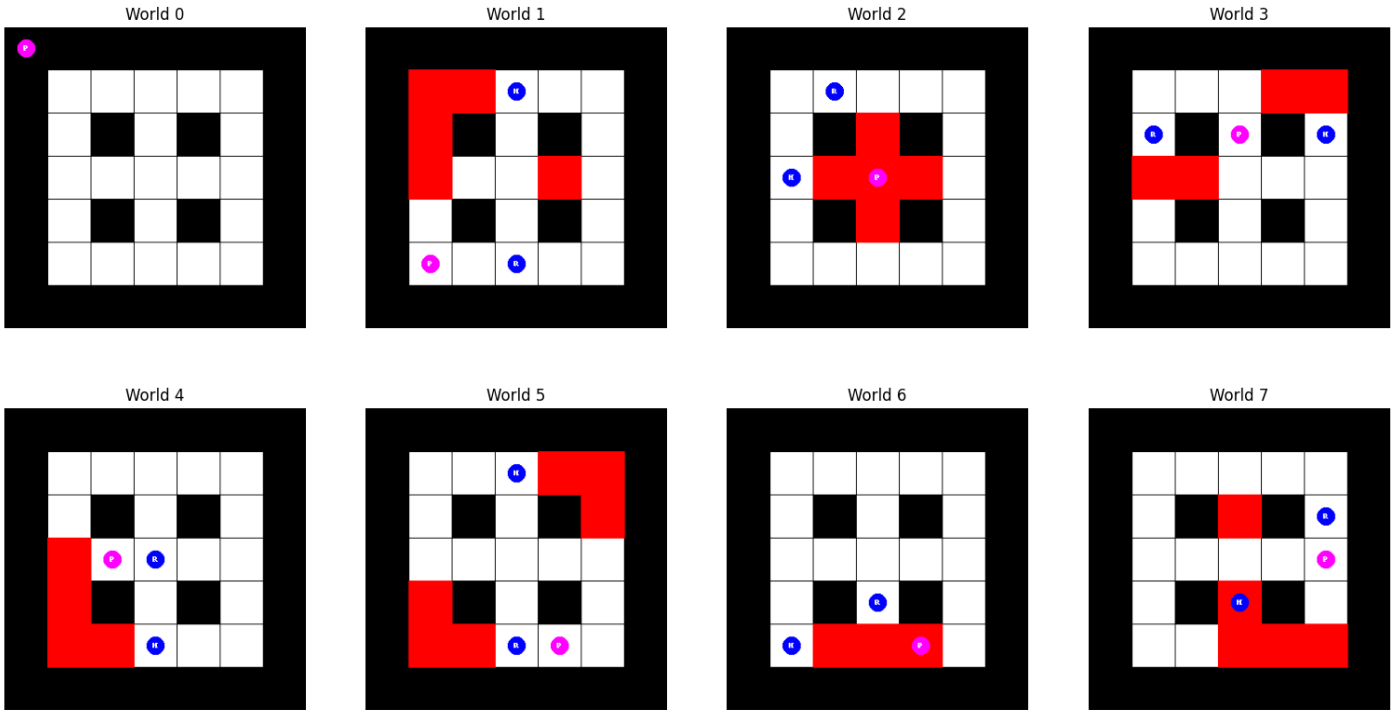


Figure 1: Updated World Designs

Introduction

Winter Break:

- Simulation of assuming bias policies revealed trivial effects
- Primarily a product of game design and training approach

Current Work:

- Improving game design to allow for consistent training (non-trivial)
- Training stable CPT value functions that represent biased policies
- Testing effects of assuming different biases for H in simulation

Challenges:

- Stuck in local optima due to sparse gains and frequent penalties
- Sensitive and fine balance for the following hyper-parameters that induce interesting policies to contrast:
  - Admissible bounds for risk-sensitivity
  - World design and initial conditions
  - Learning hyper-parameters

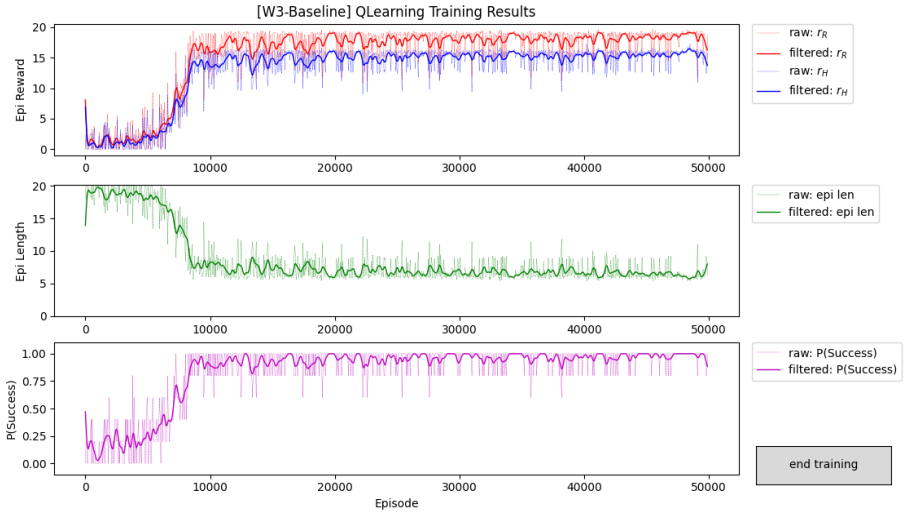
Agenda

Issues with Current Game

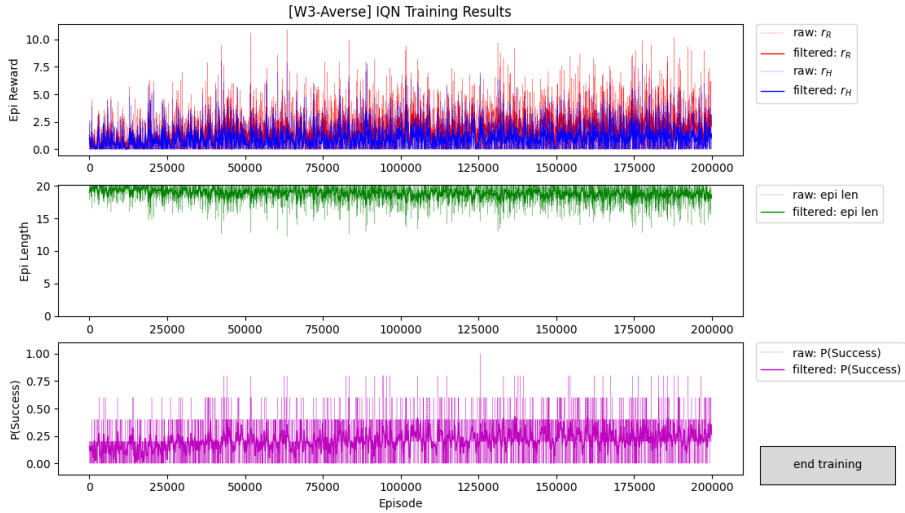
Different worlds induced differences in bias policies either

- too strong
  - risk-averse had 0% catch rate while risk-seeking had 100% catch rate
  - When switching to incorrect assumption $\rightarrow$ 0% catch rate
  - impossible to evaluate coordination strategies were incompatible or had same outcome
- too weak
  - both policies either had near 100% or both had 0% performance
  - solution is so obvious that CPT does not change it
  - again produced trivial result due similarity between strategies

Training Results (World 3)



(a) Optimal



(b) Risk-Averse

Figure 2: World 3 Training Results

List of Updates

- Redesigned world initial states and penalty locations
- Reward for catching target $r(catch) = 20 \rightarrow 25$
- Single cumulative reward r_ζ at the end of the game
  - apply eligibility traces ($TD(\lambda)$) to account for increased sparsity
- Reward is now non-negative
  - the cumulative reward at the end of the game is $r_\zeta = max(r_\zeta, 0)$
  - new objective: maximize your reward **upon catching the target**
  - doing really bad and not catching the stag are now equivalent
  - eliminates trivial policy of avoiding rewards by never moving (especially relevant in averse-conditions)

Updated Worlds

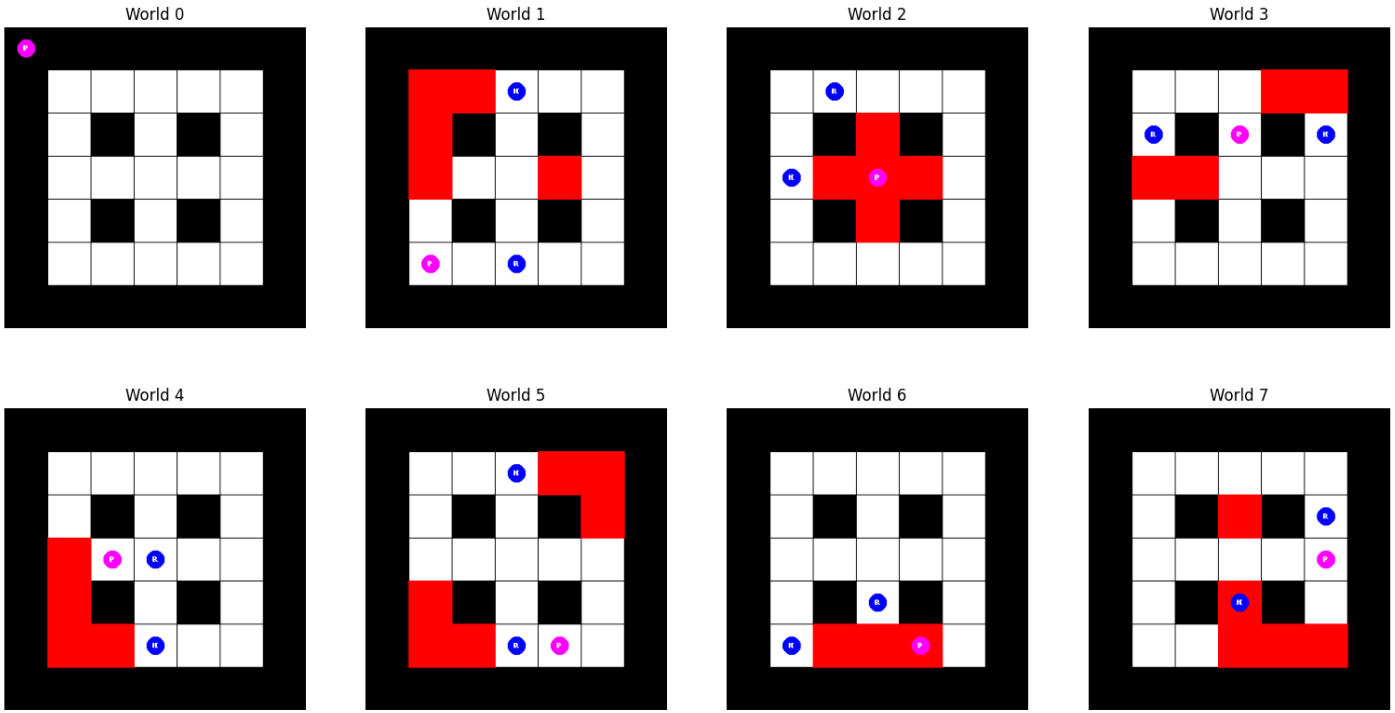


Figure 3: Updated World Designs

Agenda

Training Setup

- Implemented a independent
  - joint-Q learning algorithm
  - directed exploration
  - eligibility traces
  - Quantal Response Equilibrium (QRE) used as the equilibrium condition
- Trained 3x policies π trained during self-play:
  - baseline/optimal (π_0)
  - risk-averse (π_A)
  - risk-seeking (π_S)
- Baseline policy π_0 was used as prior for biased policies π_A and π_S trained with cumulative prospect theory (CPT) agents
- Sophistication (level of recursion) was set to 3 in the QRE

Notation

- ego agent denoted by subscript $(\cdot)_k$ where $(\cdot)_{-k}$ represents the partner and $k, -k \in K$
- joint state $s \in S$ where $S = S_k \times S_{-k}$
- joint action $a \in A$ where $A = A_k \times A_{-k}$
  - a may be written as $a = \{a_k, a_{-k}\}$ for clarity
  - alternativley, a expressed as policies $a = \{\pi_k(s), \pi_{-k}(s)\}$
- ego stage reward r_t
- let $e_k(s, a)$ denote an eligibility trace keeping track of state visitations within an episode that decays by λ each step
- let $\mathcal{P}_{CPT} = \{b, l, \gamma^+, \gamma^-, \delta^+, \delta^-\}$ denote a CPT model
- let $\mathbb{C}[(\cdot)]$ denote the expected value of rewards given true rewards and a CPT transformation

Notation

- a policy π_k
  - always denotes choosing ego action a_k in state s
  - a_k sampled from *joint state-joint action* values $Q_k(s, a)$
  - requires inferring partner $-k$ policy $\hat{\pi}_{-k}$
  - $Q_k(s, a)$ is reduced to *joint state-ego action* values $Q_k(s, a_k)$ by
    - conditioning on $\hat{\pi}_{-k}$
    - s.t. $Q_k(s, a_k) = \mathbb{E}\big[Q_k(s, a | a = \{a_k, a_{-k} = \hat{\pi}_{-k}(s)\})\big]\ \forall a_k \in A_k$
  - a_k is drawn from $Q_k(s, a_k)$ using a nominal Boltzmann dist. [1]
  - for brevity this reduction will be implied and we will write $Q_k(s, \{a_k, \hat{\pi}_{-k}(s)\})$ to denote the full expression $\mathbb{E}\big[Q_k(s, a | a = \{a_k, a_{-k} = \hat{\pi}_{-k}(s)\})\big]\ \forall a_k \in A_k$

[1] nominal implies that rationality =1

Algorithm

IQL-QRE

Initialize $Q_k(s, a)$ arbitrarily for all s, a
foreach *episode* **do**
  Initialize s and $e_k(s, a) = \mathbf{0}$
  foreach *step of episode* **do**
    $a_k \leftarrow$ ego action given by $\pi_k(s|\hat{\pi}_{-k}(s))$
    Take action a_k, observe joint action a, rewards r_k, and next state s'
    $\delta \leftarrow r_k + \gamma \max_{a'_k} Q_k(s', \{a'_k, \hat{\pi}_{-k}(s')\}) - Q_k(s, a)$
    $e_k(s, a) \leftarrow e_k(s, a) + 1$
    foreach $s \times a$ **do**
      $Q_k(s, a) \leftarrow Q_k(s, a) + \alpha\delta e_k(s, a)$
      $e_k(s, a) \leftarrow \gamma\lambda e_k(s, a)$
    end foreach
    $s, a \leftarrow s', a'$
    Until s is terminal ;
  end foreach
end foreach

Area Under the Indifference Curve (AUIC)

area under the indifference curve (AUIC) is an expression of preference for accepting or rejecting a gamble over actions with certain outcomes in terms of probabilities $p(accept)$

- AUIC is evaluated over space of feasible rewards $\mathbf{R}$ in the game
- Define binomial-choices (a_1, a_2)
  - with outcomes sampled from $\mathbf{R}$ s.t. $\mathbf{R}_1, \mathbf{R}_2 = \mathbf{R}$
- The outcomes of each choice are then:
  - a_1 containing one certain outcome
    - with possible rewards $\mathbf{R}_1 = \{r_1 - 0.5 * r_\rho \; \forall \; r_1 \in \mathbf{R}_1\}$
  - a_2 containing two uncertain outcomes (with/without penalty r_ρ)
    - with possible rewards $\mathbf{R}_2 = \{[r_2, (r_2 - r_\rho)] \; \forall \; r_2 \in \mathbf{R}_2\}$
    - with probabilities $p = [(1 - p_\rho), p_\rho]$ for each outcome occurring

Area Under the Indifference Curve (AUIC)

- *Indifference Curve*:[2]
  - a continuous curve through the 2D reward space ($\mathbf{R}_1 \times \mathbf{R}_2$)
  - occurs when no preference is expressed s.t. $p(accept) = 1 - p(accept)$
- $p(accept) = p(a_2)$ then implies risk-sensitivity where
  - An optimal agent expresses no preference[3] given $r_1 = r_2 \; \forall \; r_1, r_2 \in \mathbf{R}$
  - Preferences become more complex as we apply CPT $\mathbb{C}[(\cdot)]$
- AUIC[4] will be calculated as follows:
  - Expresses the cumulative (mean) probability of $p(accept)$ across a symmetrical space of rewards transformed by CPT
  - Centered around 0 for legibility s.t. AUIC$\in (-0.5, 0.5)$
  - AUIC $= \frac{1}{|\mathbf{R}_1 \times \mathbf{R}_2|} \sum_{r_1, r_2 \in \mathbf{R}_1, \mathbf{R}_2} p(accept | \mathbb{C}[r_1, r_2]) - 0.5$

[2] white line in the following figures

[3] is indifferent when presented choices

[4] also be described as a preference anomaly but a modified-AUIC is more consistent with the literature

Interpretting AUIC

- AUIC $< p_\epsilon$: the agent cumulatively prefers rejecting the gamble and is risk-averse
- AUIC $> p_\epsilon$: the agent cumulatively prefers accepting the gamble and is risk-seeking
- $|\text{AUIC}| < p_\epsilon$: the agent agent has week cumulative preferences and is risk-insensitive
- AUIC $= 0$: the agent has no cumulative preferences and is optimal
- where $p_\epsilon = 0.1$ is a threshold defining what we consider risk-sensitive

Area Under the Indifference Curve (AUIC)

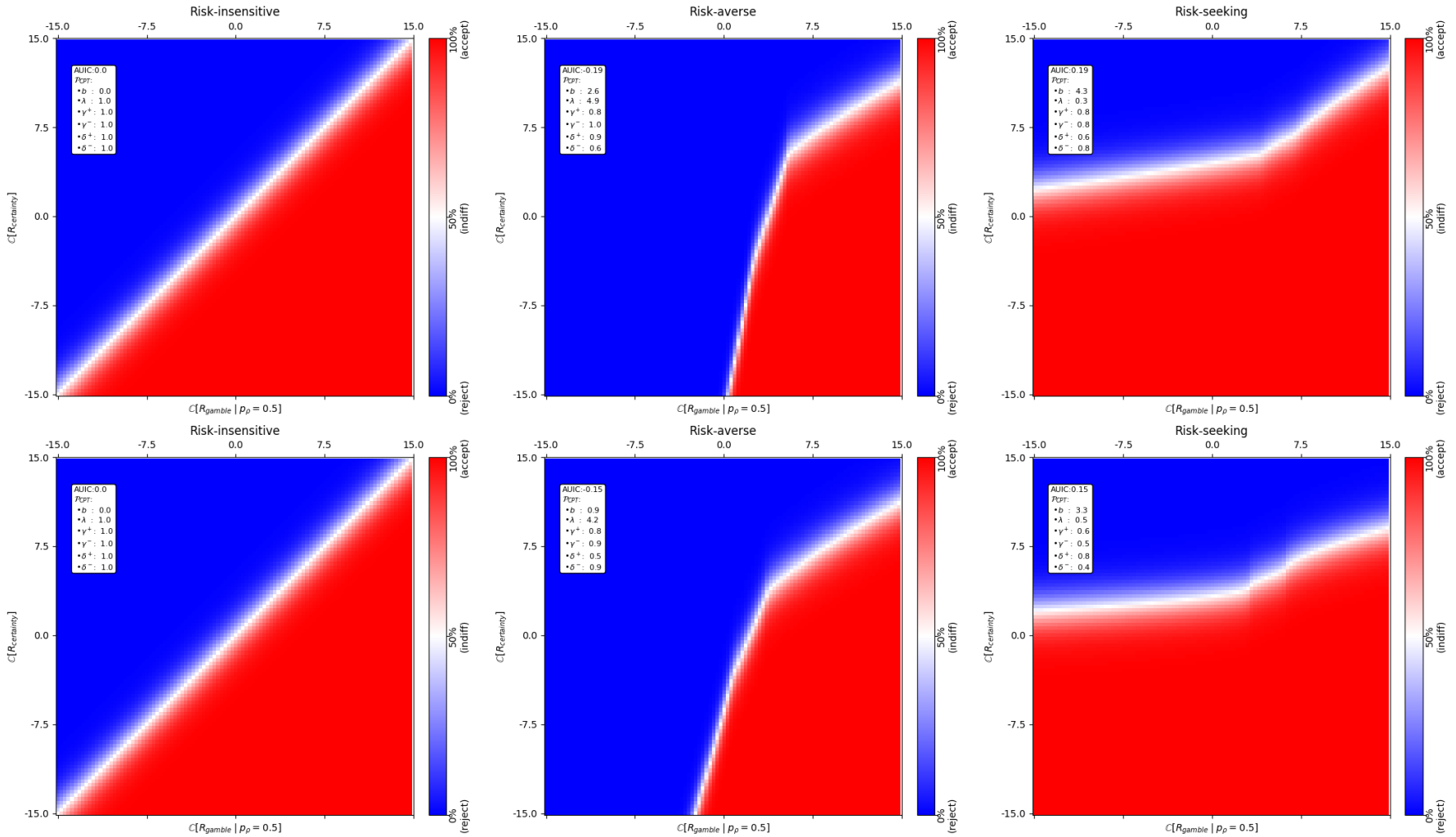
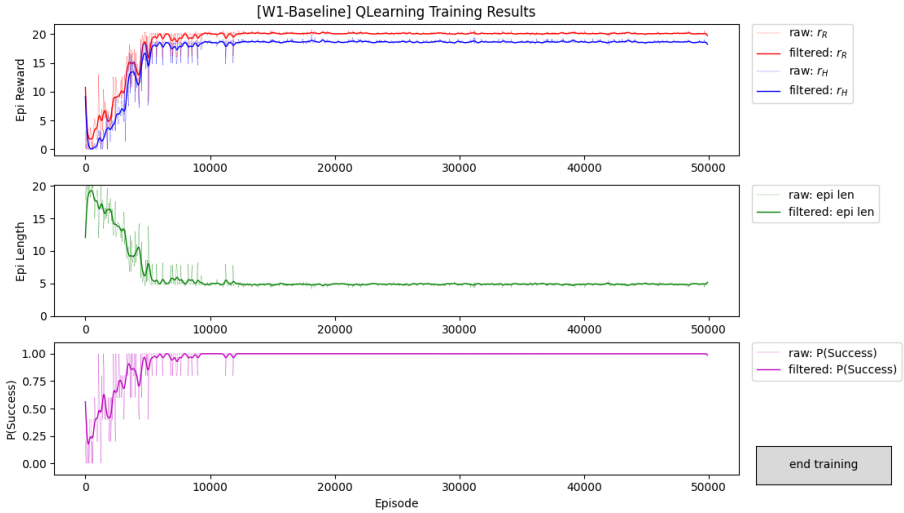


Figure 4: AUIC Samples

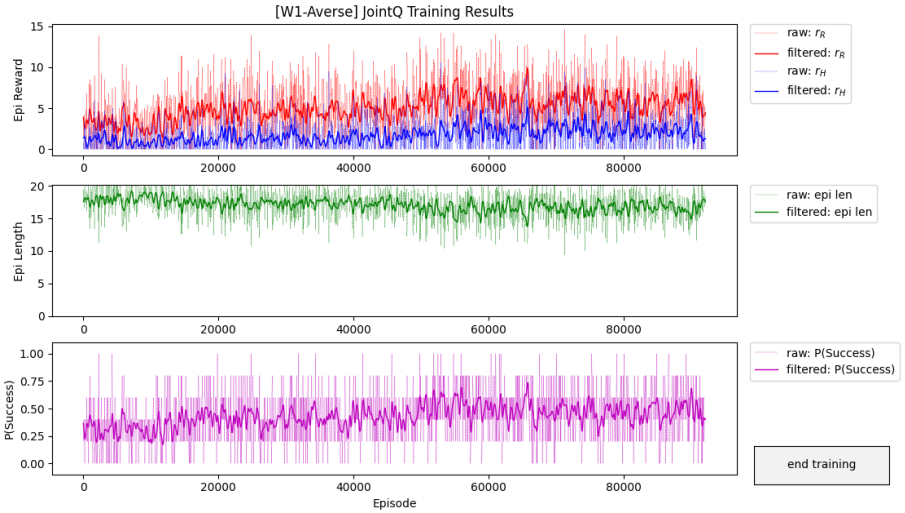
Creating Biased Policies

- CPT parameters $\mathcal{P}_{CPT}$ were stochastically perturbed while training biased policies
  - $\mathcal{P}_{CPT}$ were sampled in batches every 200 episodes
  - $\mathcal{P}_{CPT}$ were sampled from feasible bounds found in previous studies
  - $\mathcal{P}_{CPT}$ attributed to averse or seeking behavior based on AUIC
  - $\mathcal{P}_{CPT}$ is continuously sampled until intended risk-sensitivity (AUIC) is met

Training Results (World 1)



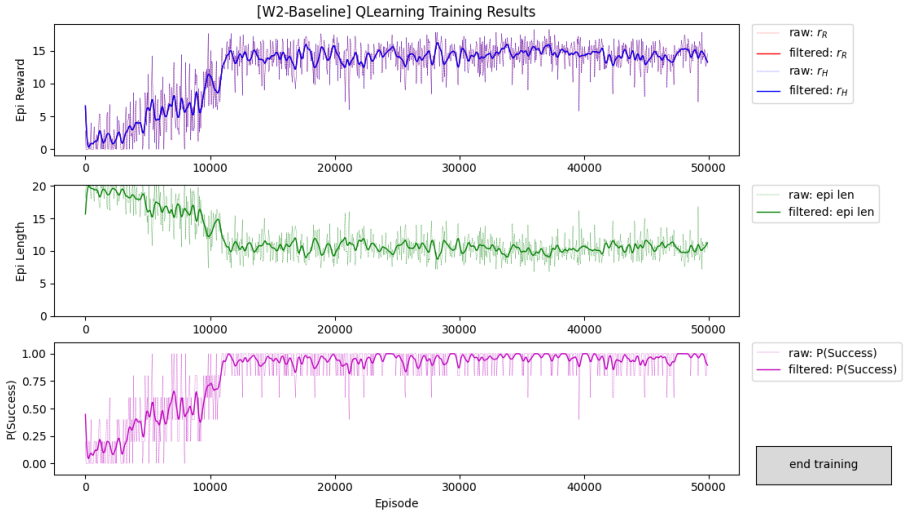
(a) Optimal



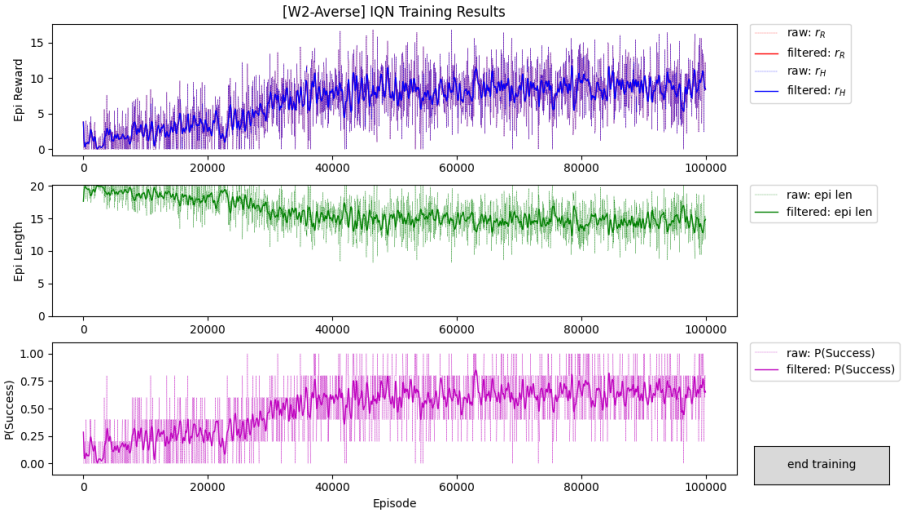
(b) Risk-Averse

Figure 5: World 1 Training Results

Training Results (World 2)



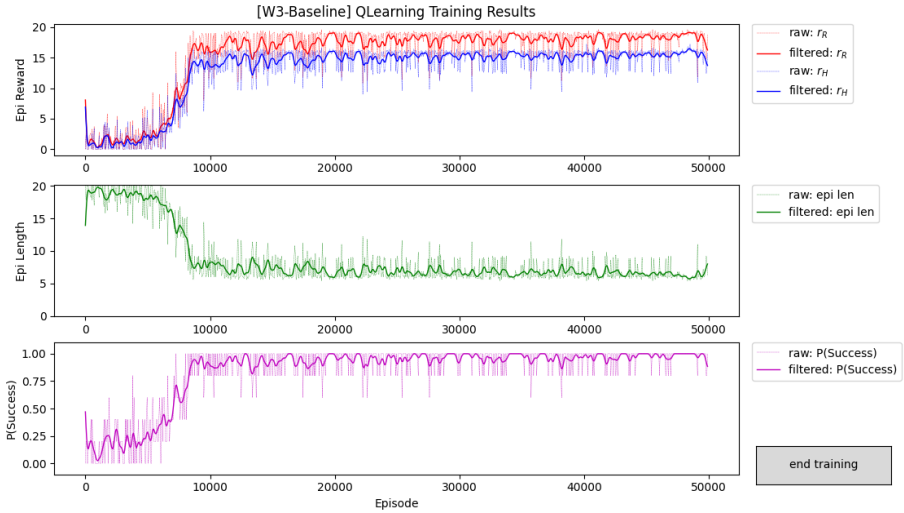
(a) Optimal



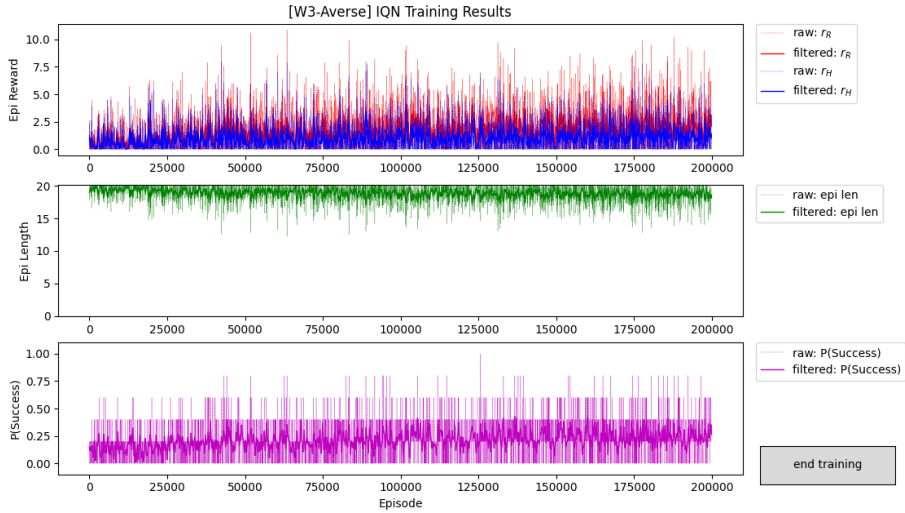
(b) Risk-Averse

Figure 6: World 2 Training Results

Training Results (World 3)



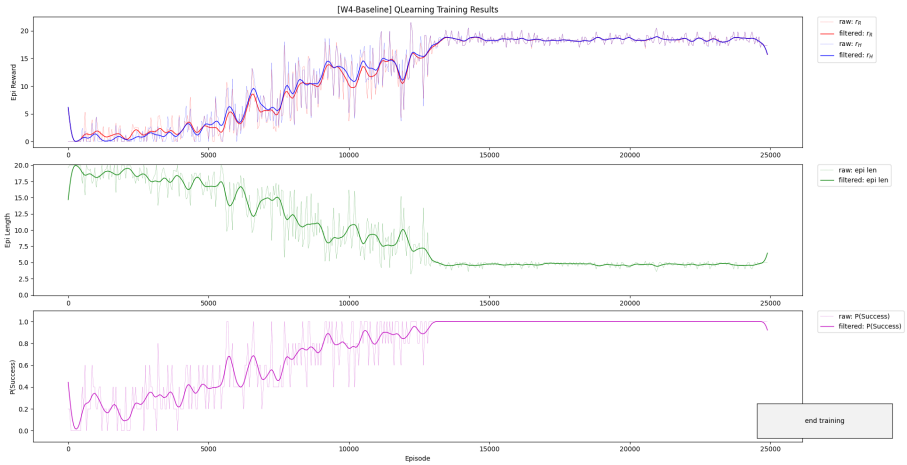
(a) Optimal



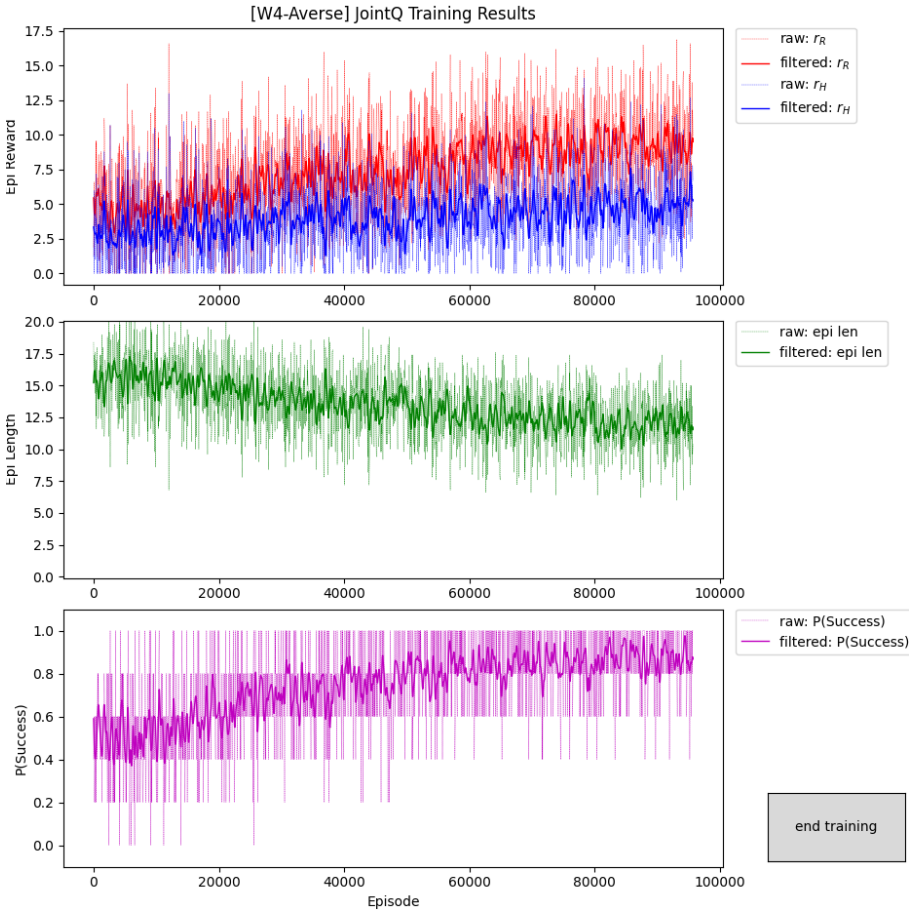
(b) Risk-Averse

Figure 7: World 3 Training Results

Training Results (World 4)



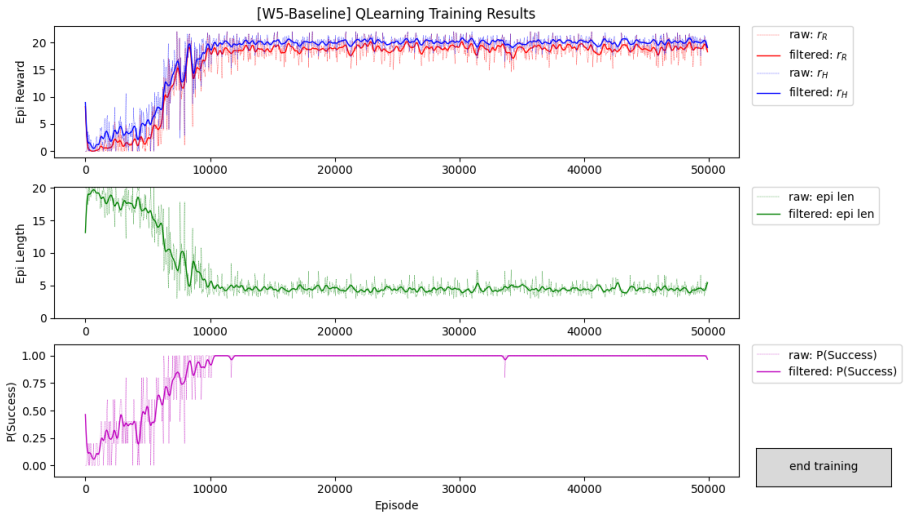
(a) Optimal



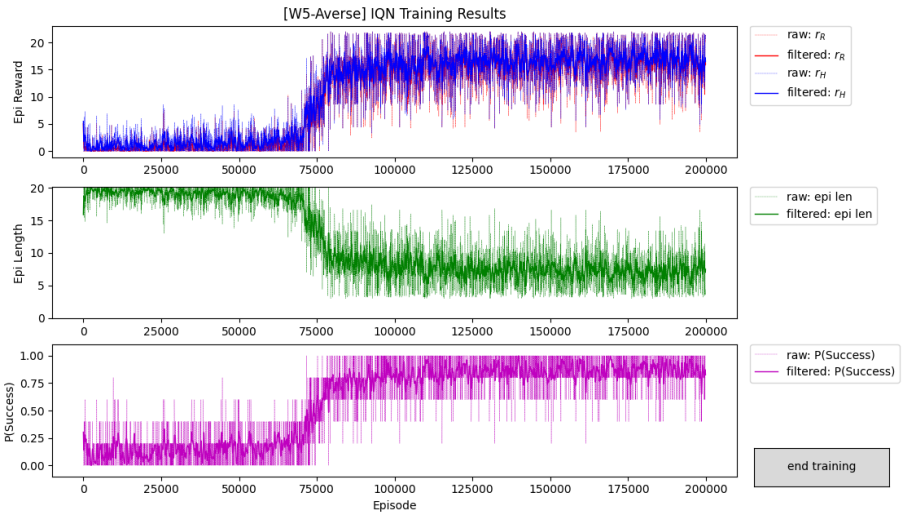
(b) Risk-Averse

Figure 8: World 4 Training Results

Training Results (World 5)



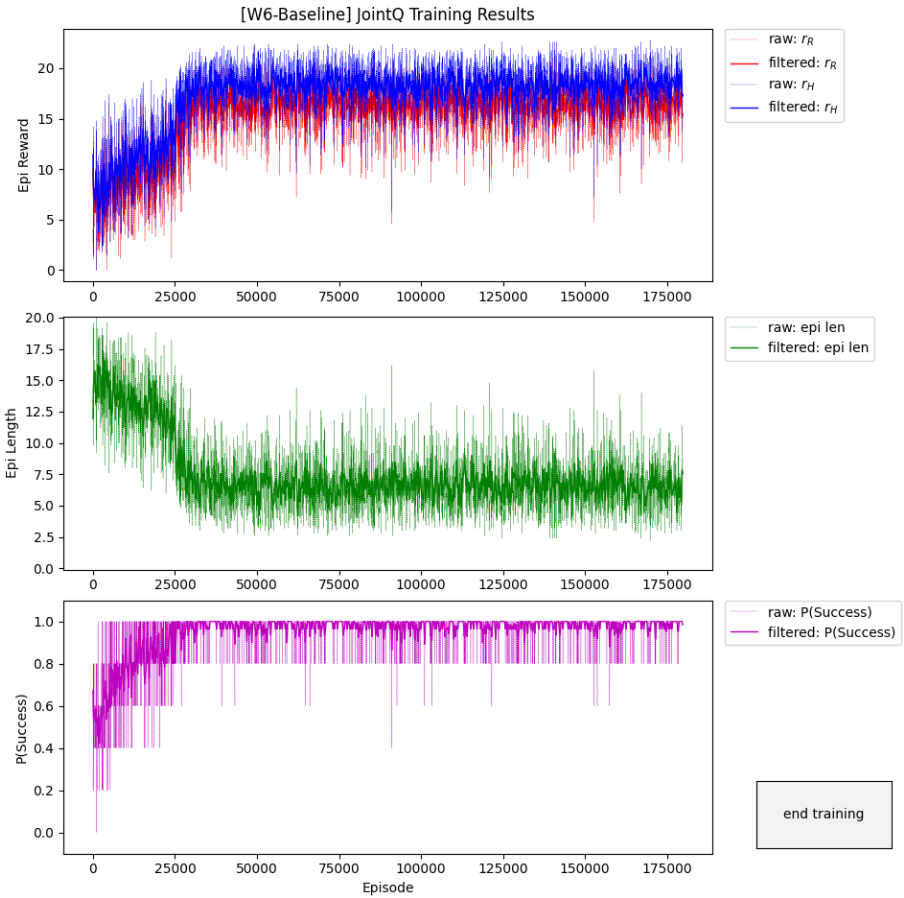
(a) Optimal



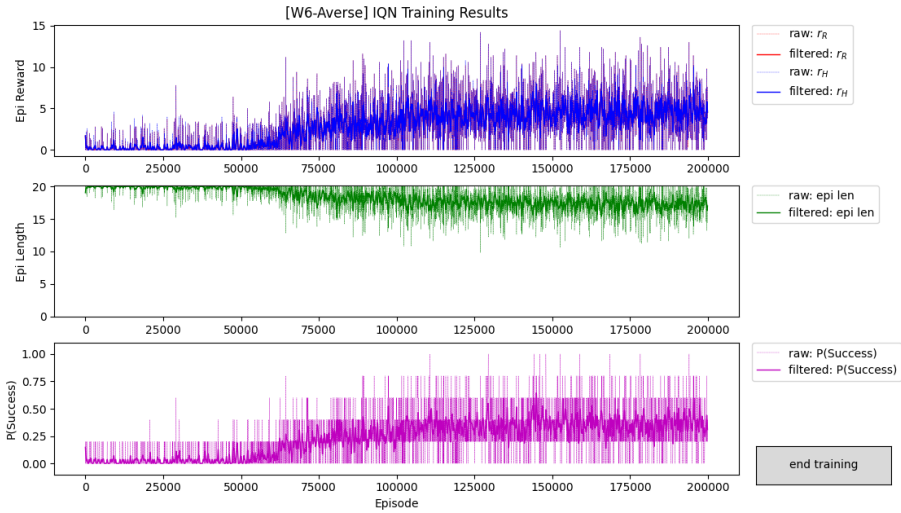
(b) Risk-Averse

Figure 9: World 5 Training Results

Training Results (World 6)



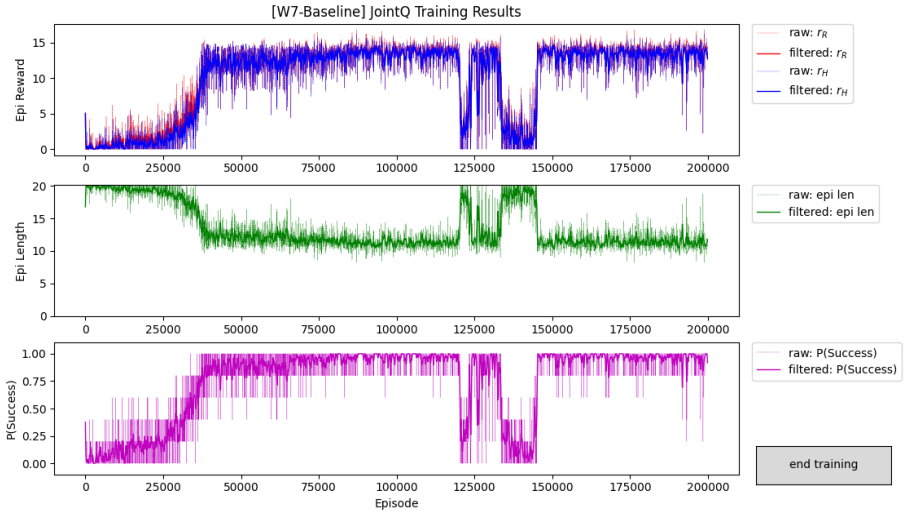
(a) Optimal



(b) Risk-Averse

Figure 10: World 6 Training Results

Training Results (World 7)



(a) Optimal

(b) Risk-Averse

Figure 11: World 7 Training Results

Discussion

- Policies are generally noisy due to
  - non-stationarity
  - rationality constant = 1
- Equilibrium would be less stochastic with higher rationalities
- Noise is sufficient for contrasting strategy e.i. there exists dis-coordination and vulnerability to partner uncertainty
- Baseline (optimal) π_0 policies are often similar to Risk-Seeking π_S policies since the game is designed for the agents to succeed
  - Rushing through penalties is often a good strategy
  - π_S still more susceptible to partner and target stochasticity
- **may make minor attempts to improve policies in future but this is good for now*

Agenda

Simulation Setup

Goal:

- How assumptions of H's risk-sensitivity effect team performance
- Validate conflicting optimal policies based on risk-sensitivity

Experimental Conditions:

- We manipulate
  - what R assumes H's policy to be ($\hat{\pi}_H$)
  - what H's policy actually is (π_H)
- The experimental condition is then written as $\mathcal{C}_{\pi_H}^{\hat{\pi}_H}$
  - superscript is R's assumption of H policy
  - subscript is H's actual policy
  - no script denotes all possible combinations of the missing index
    - e.i. $\mathcal{C}_{\pi_A} = \mathcal{C}_{\pi_A}^{\hat{\pi}_A} \cup \mathcal{C}_{\pi_A}^{\hat{\pi}_S}$

Simulation Setup

Experimental Conditions:

- Therefore, we get four experimental conditions $\mathcal{C}$:
  - $\mathcal{C}_{\pi_A}^{\hat{\pi}_A}$: Assume-Averse + Is-Averse (Correct Assumption)
  - $\mathcal{C}_{\pi_A}^{\hat{\pi}_S}$: Assume-Seeking + Is-Averse (Incorrect Assumption)
  - $\mathcal{C}_{\pi_S}^{\hat{\pi}_A}$: Assume-Averse + Is-Seeking (Incorrect Assumption)
  - $\mathcal{C}_{\pi_S}^{\hat{\pi}_S}$: Assume-Seeking + Is-Seeking (Correct Assumption)
  - *two baseline policies are also used* $\mathcal{C}_{\pi_0}^{\hat{\pi}_0}$ *for reference*

Simulation Analysis

Approach:

- We **only compare between R's assumption** and within H's actually policy
- H's actually policy directly effects game performance and invalidates some evaluation metrics
- R's policy will be $\pi_R = \pi_0$ conditioned on $\hat{\pi}_H$ using QRE
- H will assume R uses H's true policy $\hat{\pi}_R = \pi_H$
- Run simulated game 1000x for each of the 4 conditions

Simulation Results (Mean)

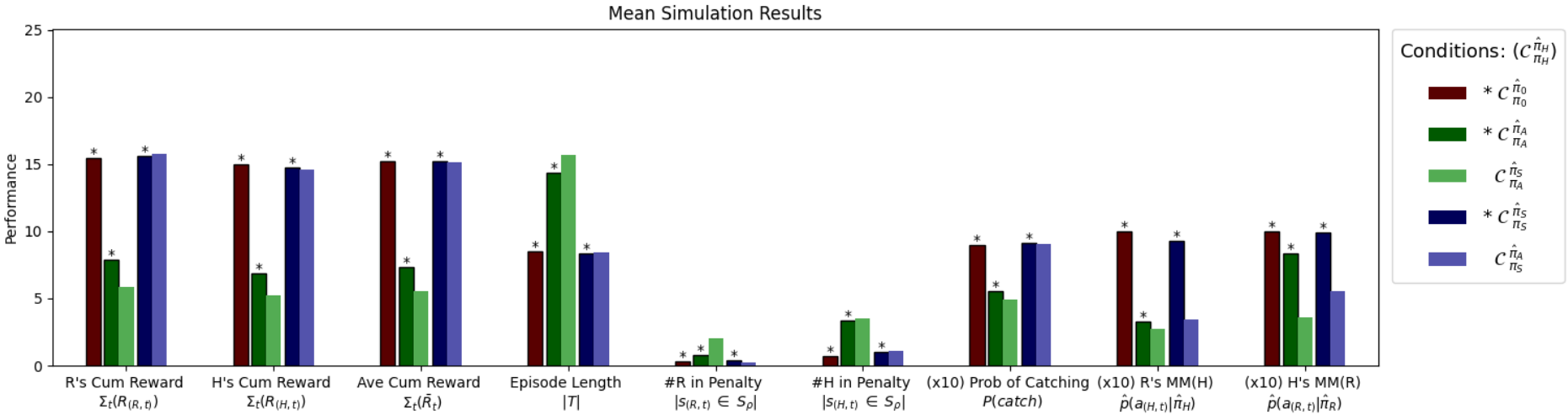


Figure 12: Evaluation of simulated conditions summary

Observations

Observation 1: Similar performances in baseline $\pi_H = \pi_0$ and risk-seeking $\pi_H = \pi_S$ human conditions:

- worlds designed with simple, achievable catches in mind (best strategy = rush target)
- realized trajectories ζ_t are then similar $\zeta(\pi_0)_t \approx \zeta(\pi_S)_t$
- H's perceived reward $\sum \tilde{r}(\pi_0)_t < \sum \tilde{r}(\pi_S)_t$ due to the discounted penalties
- Coordination effects diminish as we get closer to the target

Observations

Observation 2: Agent Rewards, Team Rewards, and Episode Length

- Sizable performance losses when $\mathcal{C}_{\pi_A}^{\hat{\pi}_A} \rightarrow \mathcal{C}_{\pi_A}^{\hat{\pi}_S}$
- Negligible difference when $\mathcal{C}_{\pi_S}^{\hat{\pi}_S} \rightarrow \mathcal{C}_{\pi_S}^{\hat{\pi}_A}$ as described by discussion in Observation 1
- Higher cumulative rewards for R is product of game design, not policy assumption

Observation 3: # of Penalty States Entered

- High number of penalty states entered by R when $\mathcal{C}_{\pi_A}^{\hat{\pi}_S}$
- R assumes H will rush straight towards the target but H does not
- R therefore has to idle or move about in penalty states to wait for H to approach

Observations

Observation 4: Accuracy of Agent's Mental Model (MM(Partner))

- Both MM's of their partner saw significant losses in all $\mathcal{C}$
- H's MM(R) more significant losses than when $\mathcal{C}_{\pi_A}^{\hat{\pi}_A} \rightarrow \mathcal{C}_{\pi_A}^{\hat{\pi}_S}$
- R's MM(H) was more accurate in the $\mathcal{C}_{\pi_0}^{\hat{\pi}_0}$ than when $\mathcal{C}_{\pi_S}^{\hat{\pi}_S}$
- Baseline π_0 had best prediction accuracy in MM
- Generally, agents found partner's actions harder to predict [5]
- Has implications for trust degradation (process information)

[5] This supports Observation 1 in that effective coordination in MM's can significantly decrease with little loss in task performance by mere virtue of being closer to the target

Discussion

Summary:

- Expected effects of correct vs incorrect assumption for $\mathcal{C}_{\pi_A}$ in all categories
- Expected effects of correct vs incorrect in MM of partner actions for all $\mathcal{C}$
- Unexpected negligible losses in reward, episode length and $p(catch)$ for $\mathcal{C}_{\pi_S}$
- Unexpected similarity between $\mathcal{C}^{\hat{\pi}_0}_{\pi_0}$, $\mathcal{C}^{\hat{\pi}_S}_{\pi_S}$, and $\mathcal{C}^{\hat{\pi}_A}_{\pi_S}$ policies

Implication:

- Trust as process information likely affected in all $\mathcal{C}$
- Trust as performance information likely affected when $\mathcal{C}_{\pi_A}$ [6]
- Sufficient effect in coordination to evaluate trajectories when $\mathcal{C}_{\pi_A}$

[6] Would like to see more coordination effect when $\mathcal{C}_{\pi_S}$ but this is exceedingly challenging with a approxitaley rational target

Agenda

Ongoing Work

Status:

- Game is functioning with all policy conditions
- Hosting multiple concurrent users available
- Data collection system is validated

Open Items:

- Need a long-term web hosting solution (security and ISP compatibility)
- Likely need to purchase a web host service if ISP reaches out
- Might be a good idea anyway since I am bootlegging a non-http port for http
- Final pretrial validation (4x participants)