Modeling Human Behavior Without Humans: Prospect Theoretic Multi-Agent Reinforcement Learning

Sheyan Lalmohammed [*1] **Khush Gupta** [*1] **Alok Shah** [*1] **Keshav Ramji** [*2]

Abstract

Bridging the gap between algorithmic precision and human-like risk nuance is essential for crafting multi-agent systems that learn adaptable and strategically intuitive behaviors. We introduce CPT-MADDPG, an extension of the Multi-Agent Deep Deterministic Policy Gradient algorithm, embedding Cumulative Prospect Theory (CPT) value and probability weight transforms into both actor and critic updates. By replacing expected return maximization with rank-dependent Choquet integrals over gains and losses, CPT-MADDPG endows agents with tunable risk profiles —ranging from exploratory, risk-seeking to conservative, loss-averse behaviors—without human intervention. Across competitive pursuit (Simple Tag), cooperative coverage (Simple Spread), and strategic bidding (first-price auctions), we show that risk-seeking parameterized CPT speeds early learning, extreme risk-averse parameterized CPT enforces prudence at a performance cost, transparent utility sharing preserves coordination under heterogeneity, and naive dynamic adaptation destabilizes convergence. In auction settings, learned CPT policies replicate documented overbidding phenomena, with short-term gains followed by long-term losses. Our work demonstrates a principled framework for integrating human-like risk attitudes toward strategic multi-agent deployment.

1. Introduction

Multi-agent reinforcement learning (MARL) has achieved remarkable success in domains ranging from autonomous driving to strategic game playing by training agents to maximize expected cumulative rewards (Lowe et al., 2017). Yet, such agents implicitly assume classical rationality, neglecting systematic human decision biases under risk. Decades of behavioral economics research have shown that real humans deviate from expected-utility theory in predictable ways—exhibiting loss aversion, reference dependence, and probability weighting—captured by Prospect Theory (Kahneman & Tversky, 1979) and its extension, Cumulative Prospect Theory (CPT) (Tversky & Kahneman, 1992).

Despite its promise, naively inserting CPT into multi-agent actor–critic frameworks poses several challenges. First, the non-linear Choquet integrals in CPT introduce nonconvexity into the objective, destabilizing standard gradient updates. Second, the probability-weighting step requires empirical estimation of tail probabilities over returns, demanding careful batch-based approximations to avoid bias. Third, heterogeneous risk profiles across agents can yield non-stationary dynamics, complicating convergence.

In this work, we bridge the gap between rational MARL agents and human-like risk-sensitive behavior by embedding full CPT value and probability transformations into the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework. Unlike prior risk-sensitive RL approaches that focus on single-agent settings, CPT-MADDPG applies rank-dependent weighting directly to cumulative returns in both critic and actor updates, enabling agents to exhibit calibrated risk-seeking or risk-averse behaviors without humans in the loop. We address the above issues by (1) designing a minibatch-based CPT integral approximation that is fully differentiable, (2) integrating rank-dependent weighting inside the MADDPG critic and actor updates to maintain stability, and (3) extending the approach with observability and adaptive-parameter modules that we show preserve coordination and control non-stationarity.

- **Observability Adjustment:** Allowing agents to access each other's subjective CPT-adjusted utilities, and deriving a cross-agent valuation aggregation that modifies the Bellman backup.
- **Adaptive Behavioral Parameters:** Treating CPT parameters (α, β, λ) as learnable variables, optimized alongside network weights to adapt risk profiles dy-

[*]Equal contribution [1]University of Pennsylvania [2]IBM Research AI. Correspondence to: Sheyan Lalmohammed <sheyan@upenn.edu>, Khush Gupta <khushg@upenn.edu>, Alok Shah <alokshah@upenn.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

namically during training.

We evaluate CPT-MADDPG in two multi-particle environments (MPE's) (Lowe et al., 2017): competitive Simple Tag, cooperative Simple Spread, and a first-price auction with mixed CPT and non-CPT bidders. Our experiments demonstrate that (1) moderate risk-seeking parameterized CPT values yield exploratory, risk-seeking dynamics, (2) extreme risk-averse parameters induces conservative, low-variance strategies, (3) transparency of utilities from rewards preserves coordination, and (4) adaptive behavioral parameter dynamics can destabilize learning if updated too frequently. These demonstrate how human-like risk biases can be systematically tuned to enhance exploration, enforce safety, or predictably modulate strategic behavior in richly interactive settings.

Our contributions can be summarized as follows:

1. We introduce CPT-MADDPG, integrating full CPT value and probability transforms into a multi-agent actor–critic algorithm.
2. We derive and implement observability-adjusted CPT updates, aggregating cross-agent utilities in the critic target.
3. We propose a secondary optimization of CPT hyperparameters for adaptive risk profiling during training.
4. We provide extensive empirical validation across competitive, cooperative, and auction tasks, highlighting the behavioral and performance trade-offs of agents trained to follow CPT-integrated policies.

2. Related Work

Prospect Theory and Cumulative Prospect Theory. Prospect Theory (PT) was introduced by Kahneman & Tversky (1979) to explain systematic deviations from expected-utility theory, notably loss aversion, reference dependence, and probability weighting. Cumulative Prospect Theory (CPT) extends PT to multi-outcome gambles by applying rank-dependent weighting to cumulative probabilities, which corrects several anomalies of the original formulation and enables tractable aggregation of outcomes (Tversky & Kahneman, 1992).

Risk-Aware Learning and CPT Integration. Utility-based approaches, leveraging exponential or power utilities, provide an alternate route for encoding risk attitudes (García & Fernández, 2015). In single-agent RL, risk-sensitive objectives have been studied extensively (Shen et al., 2014). Conditional Value at Risk (CVaR) (Rockafellar & Uryasev, 2002) criteria have been incorporated into MDPs to control downside risk (Bäuerle & Ott, 2014), with recent works extending it to a class of policy gradients (Tamar et al., 2015). These frameworks demonstrate that modifying the reward aggregation can systematically steer agent behavior toward risk-averse or risk-seeking policies. Additionally, CVaR-based objectives have been leveraged in both cooperative (MARL) and single-agent tasks to model the distribution over Q-values in the MARL setting, thereby mitigating collective downside risk (Qiu et al., 2021). Similar risk-sensitive approaches have also been applied in entropy-regularized actor–critic methods (Nachum et al., 2017). Expanding on risk-sensitivity Ghaemi et al. (2024) analyzes network-aggregative games under risk awareness. L. A. et al. (2016) first proposed the combination of CPT and RL and since then a large amount of work has come to bridge the integration in the single (Jie et al., 2018; Borkar & Chandak, 2021; Ramasubramanian et al., 2021), and multi-agent settings (Danis et al., 2023; Lepel & Barakat, 2024). Borkar & Chandak (2021) analyzed a Q-learning algorithm for CPT policies and in the MARL environment, (Danis et al., 2023) proposed a multi-agent CPT-based Q-learning algorithm with weight sharing. Most recently, Lepel & Barakat (2024) proposed a policy gradient and theorem to solve the CPT policy optimization problem, and Ethayarajh et al. (2024) devised an approach for aligning language models driven by CPT principles.

Multi-Agent Actor–Critic Methods. There have been a plethora of works in MARL Actor-Critic Methods (Lowe et al., 2017; Iqbal & Sha, 2019; Du et al., 2019; Su et al., 2020; Pu et al., 2021; Xiao et al., 2022) Our work builds upon Multi-Agent Deep Deterministic Policy Gradient (MADDPG) (Lowe et al., 2017), an actor–critic framework tailored for mixed cooperative–competitive environments. By employing centralized critics with access to all agents' observations and decentralized actors for scalable execution, MADDPG achieves stabilized learning and effective coordination. This paradigm directly informs our CPT-MADDPG design, where CPT-driven value transformations are embedded within each agent's critic update to capture risk preferences across strategic interactions.

3. Preliminaries

3.1. Policy Gradient Algorithms

Policy Gradient and Actor-Critic. Policy Gradient methods (Sutton et al., 1999) directly optimize the policy parameters by estimating $\nabla_\theta J(\theta)$ and performing gradient ascent:

$$J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}\big[R(\tau)\big] = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Big],$$

where $\tau = (s_0, a_0, s_1, \dots)$, $p_\theta(\tau) = \rho_0(s_0) \prod_t \pi_\theta(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t)$, and $0 < \gamma < 1$ is the discount factor.

The Policy Gradient Theorem states:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d^\pi,\, a \sim \pi_\theta} \big[\nabla_\theta \log \pi_\theta(a \mid s)\, Q^\pi(s, a) \big],$$

where $d^\pi(s) \propto \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi)$ is the discounted state-visitation distribution and

$$Q^\pi(s, a) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \Big].$$

To reduce variance, one replaces $Q^\pi(s, a)$ with the advantage function

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s), \quad V^\pi(s) = \mathbb{E}_{a \sim \pi_\theta} \big[Q^\pi(s, a) \big].$$

Actor–Critic methods (Konda & Tsitsiklis, 1999) maintain the following:

- an *actor* $\pi_\theta(a \mid s)$, with an update rule of $\theta \leftarrow \theta + \alpha\, \mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s)\, A^\pi(s, a)]$,
- a *critic*, $V_w(s)$ or $Q_w(s, a)$, trained (e.g. by temporal-difference learning) to approximate V^π or Q^π.

This coupling yields low-variance, on-policy gradient estimates while retaining exploration.

Multi-Agent Deep Deterministic Policy Gradient. The Multi-Agent Deep Deterministic Policy Gradient (MADDPG) (Lowe et al., 2017) algorithm extends the Deep Deterministic Policy Gradient (DDPG) framework (Lillicrap et al., 2019) to multi-agent settings, particularly those involving mixed cooperative-competitive interactions. A core tenet of MADDPG is *centralized training with decentralized execution*. During execution, each agent i acts based on its own local observation o_i using its actor policy $\mu_i : \mathcal{O}_i \to \mathcal{A}_i$, parameterized by θ_i^μ, which outputs a deterministic action $a_i = \mu_i(o_i|\theta_i^\mu)$.

For training, MADDPG introduces a separate centralized critic $Q_i(x, a_1, \dots, a_N|\theta_i^Q)$ for each agent i. This critic is parameterized by θ_i^Q and takes as input some representation of the global state x (e.g., the concatenation of all agents' observations $(o_1, \dots, o_N)$ and potentially other state information) and the actions of *all* N agents $a_1, \dots, a_N$. It outputs an estimate of the expected return for agent i. The critic Q_i for each agent i is updated by minimizing the loss:

$$\mathcal{L}(\theta_i^Q) = \mathbb{E}_{(x, \mathbf{a}, \mathbf{r}, x') \sim \mathcal{D}} \left[\Big(Q_i(x, a_1, \dots, a_N|\theta_i^Q) - y_i \Big)^2 \right], \tag{1}$$

where $\mathbf{a} = (a_1, \dots, a_N)$, $\mathbf{r} = (r_1, \dots, r_N)$, and the target value y_i is computed as:

$$y_i = r_i + \gamma Q_i'(x', a_1', \dots, a_N'|\theta_i^{Q'})\big|_{a_j' = \mu_j'(o_j'|\theta_j^{\mu'})}. \tag{2}$$

Here, $\mathcal{D}$ is an experience replay buffer storing tuples $(x, \mathbf{a}, \mathbf{r}, x')$. Q_i' and μ_j' are target networks with parameters $\theta_i^{Q'}$ and $\theta_j^{\mu'}$, which are typically updated via soft updates (Polyak averaging) from their respective online network parameters.

The actor policy μ_i for each agent i is updated using the deterministic policy gradient, derived from the expected return $J(\theta_i^\mu) = \mathbb{E}[R_i]$:

$$\nabla_{\theta_i^\mu} J(\theta_i^\mu) = \mathbb{E}_{x, \mathbf{a} \sim \mathcal{D}} \left[\nabla_{\theta_i^\mu} \mu_i(o_i|\theta_i^\mu) \nabla_{a_i} Q_i(x, \mathbf{a}|\theta_i^Q) \right].$$

By conditioning the critic on the actions of all agents, the environment becomes stationary from the perspective of each agent's learning process, even as other agents' policies change. The use of separate critics for each agent allows MADDPG to be applied in scenarios with differing reward functions, including competitive or mixed settings. Optionally, if true policies of other agents are unknown during training, they can be inferred from observations.

3.2. Cooperative-Competitive Environments

Simple Tag (Competitive MPE) In this environment, $N_p = 1$ predator agent (adversary) attempts to capture a single prey agent within a bounded two-dimensional arena with stationary obstacles. The reward is defined as follows:

- Predator: $r_t^i = \begin{cases} +10, & \text{if predator } i \text{ tags prey at time } t \\ 0, & \text{otherwise} \end{cases}$
- Prey: $r_t^{\text{prey}} = \begin{cases} -10, & \text{if the prey is tagged at time } t \\ 0, & \text{otherwise} \end{cases}$

To discourage escape from a bounded area, the prey receives a penalty defined by:

$$\text{bound}(x) = \begin{cases} 0, & x < 0.9, \\ 10(x - 0.9), & 0.9 \le x < 1.0, \\ \min(\exp(2x - 2), 10), & x \ge 1.0. \end{cases}$$

This environment is implemented practically through PettingZoo[1].

Simple Spread (Cooperative MPE) N agents must cover M fixed landmarks. Each agent's observation o_t^i includes its own position and those of landmarks and other agents. At each step, agent i receives

$$r_t^i = \underbrace{\sum_{m=1}^{M} \mathbb{I}[\|x_t^i - \ell_m\| < d_{\text{cov}}]}_{\text{covered landmarks}} - \underbrace{\sum_{j \neq i} \mathbb{I}[\|x_t^i - x_t^j\| < d_{\text{coll}}]}_{\text{collision penalty}}.$$

[1] https://pettingzoo.farama.org/index.html

Here d_{cov} is the coverage radius and d_{coll} the collision threshold. All agents share identical reward functions.

First-Price Auction N agents receives a private valuation $v_i \sim \text{Uniform}(0, 100)$. Agents simultaneously submit bids $b_i \in [0, 100]$. The highest bidder wins and pays their bid; all others pay nothing. Agent i's reward is

$$r_i = \begin{cases} v_i - b_i, & \text{if } b_i = \max_j b_j \text{ (tie-broken uniformly)}, \\ 0, & \text{otherwise.} \end{cases}$$

In competitive mode, each agent maximizes its own payoff; in cooperative mode, the group reward is $\sum_i r_i$ and is shared equally.

4. Methods

4.1. Problem Formulation

We study an N-agent Markov game represented by the tuple $(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{r_i\}_{i=1}^N, \gamma)$, where $\mathcal{S}$ denotes the global state space and $\mathcal{A}_i$ is the action space of agent i. The state transition probability $P(s' \mid s, \mathbf{a})$ defines dynamics from state $s \in \mathcal{S}$ under joint action $\mathbf{a} = (a_1, \ldots, a_N)$. Each agent i receives reward $r_i(s, \mathbf{a})$, and future rewards are discounted by $\gamma \in [0, 1)$. Agent i employs a stochastic policy $\pi_{\theta_i}(a_i \mid o_i)$ parameterized by θ_i, mapping its local observation o_i to action probabilities. The objective in Multi-Agent Reinforcement Learning (MARL) for each agent i is to maximize the expected discounted cumulative return:

$$J(\theta_i) = \mathbb{E}_{\tau \sim \Pi_\theta} \left[\sum_{t=0}^{T} \gamma^t r_i(s_t, \mathbf{a}_t) \right], \tag{3}$$

where τ is a trajectory $(s_0, \mathbf{a}_0, r_0, \ldots, s_T, \mathbf{a}_T, r_T)$ and $\Pi_\theta = \prod_{j=1}^N \pi_{\theta_j}$ is the joint policy.

4.2. Cumulative Prospect Theory Adjustments

To capture human-like decision-making biases and risk attitudes, we integrate Cumulative Prospect Theory (CPT) (Tversky & Kahneman, 1992) into the agents' objectives. Rather than maximizing the standard expected return $R_i = \sum_{t=0}^{T} \gamma^t r_i(s_t, \mathbf{a}_t)$ directly, agents maximize its CPT value $C(R_i)$, defined via Choquet integrals with gains and losses relative to a reference point.

$$C(R_i) = \int_0^\infty w^+\big(P(u(R_i) > z)\big) dz + \int_{-\infty}^0 w^-\big(P(u(R_i) > z) - 1\big) dz \tag{4}$$

where $u(x)$ denotes a subjective utility function reflecting diminishing sensitivity and loss aversion, typically modeled by a power-law form:

$$u(x) = \begin{cases} x^\alpha, & x \geq 0, \\ -\lambda(-x)^\beta, & x < 0, \end{cases} \tag{5}$$

with parameters $\alpha, \beta \in (0, 1]$ controlling curvature for gains and losses, respectively, and $\lambda \geq 1$ quantifying loss aversion. The functions $w^\pm(p)$ are non-linear probability weighting transformations, typically inverse-S-shaped, overweighting small probabilities and underweighting large ones. We approximate these functions using a piecewise-linear form for efficient gradient estimation during training.

Consequently, the CPT-adjusted objective for agent i is:

$$J_{\text{CPT}}(\theta_i) = \mathbb{E}_{\tau \sim \Pi_\theta}[C(R_i)]. \tag{6}$$

4.3. Approximation of the CPT Integral

Direct computation of the CPT value $C(R_i)$ is intractable in RL, as the return distribution $P(u(R_i) > z)$ depends on the agent's evolving policy and complex environment dynamics, and is rarely available in closed form. Moreover, estimating these probabilities and evaluating the integrals for each state-action pair or trajectory during training is computationally prohibitive.

To make CPT tractable in our RL framework, we approximate $C(R_i)$ using empirical estimates from batches of sampled trajectory returns. Given a batch of B trajectories from the replay buffer, with returns $\{R_i^{(k)}\}_{k=1}^B$ for agent i, we estimate the CPT value $\hat{C}(R_i)$ using the following procedure, implemented as the `compute_cpt_integral` function:

1. **Utility Transformation:** Transform each sampled return $R_i^{(k)}$ into its subjective utility $u_k = u(R_i^{(k)})$ (see Eq. 5).

2. **Empirical Probability Estimation:** For any threshold z, estimate $P(u(R_i) > z)$ for gains and $P(u(R_i) \leq z)$ for losses empirically from $\{u_k\}_{k=1}^B$. For gains, the empirical tail probability is $\hat{P}(u(R_i) > z) \approx \frac{1}{B} \sum_{k=1}^B \mathbb{I}(u_k > z)$.

3. **Probability Weighting:** Apply the piecewise linear weighting functions $w^+(p)$ and $w^-(p)$ via `w_approx(L, p)`.

4. **Numerical Integration:** Approximate the Choquet integrals by summing over sorted unique utility values (including 0) to define integration segments. For gains ($z \geq 0$), compute $\sum_j w^+(\hat{P}(u(R_i) > z_j))(z_{j+1} - z_j)$, and similarly for losses ($z < 0$) using $w^-(\hat{P}(u(R_i) \leq z_j))$.

This empirical, batch-based approach yields a tractable estimate $\hat{C}(R_i)$ and its gradient with respect to input returns, enabling integration with gradient-based RL algorithms. While the quality of the estimate depends on batch representativeness, this method allows practical incorporation of CPT-based risk preferences in RL.

5. CPT-MADDPG Algorithm

We propose CPT-MADDPG, which extends Multi-Agent Deep Deterministic Policy Gradient (MADDPG; (Lowe et al., 2017)) to incorporate risk preferences from Cumulative Prospect Theory (CPT). MADDPG uses centralized training with decentralized execution: each agent i learns a deterministic policy $\mu_{\theta_i}(o_i)$ from its local observation o_i and a centralized critic $Q_{\phi_i}(s, \mathbf{a})$ that evaluates joint actions $\mathbf{a}$ in state s.

In CPT-MADDPG, the critic is trained with the standard temporal-difference loss:

$$L(\phi_i) = \mathbb{E}_{(s,\mathbf{a},\mathbf{r},s')\sim\mathcal{D}}\left[\left(Q_{\phi_i}(s,\mathbf{a}) - y_i\right)^2\right], \qquad (7)$$

for replay buffer $\mathcal{D}$, and target value

$$y_i = r_i + \gamma Q'_{\bar{\phi}_i}(s', \mathbf{a}')\Big|_{\mathbf{a}'_j=\mu'_{\bar{\theta}_j}(o'_j)}$$

using target networks $Q'_{\bar{\phi}_i}$ and $\mu'_{\bar{\theta}_j}$ as in MADDPG.

The actor update incorporates a CPT-based scaling factor Φ_i, computed from terminal returns in the batch using our `compute_cpt_integral` function. Specifically, the actor is updated with:

$$L(\theta_i) = -\mathbb{E}\left[\exp(\nu\Phi_i) Q_{\phi_i}(s, \mu_{\theta_i}(o_i), \mathbf{a}_{-i})\right], \qquad (8)$$

where ν controls the influence of CPT. The corresponding policy gradient is:

$$\nabla\theta_i J_{\mathrm{CPT}}(\theta_i) \approx \mathbb{E}\left[\exp(\nu\Phi_i)\nabla_{\theta_i}\mu_{\theta_i}(o_i)\nabla_{a_i}Q_{\phi_i}(s,\mathbf{a})\right]. \qquad (9)$$

This formulation implies that actions leading to higher (CPT-valued) terminal returns are more strongly reinforced if $\Phi_i > 0$, and actions leading to lower CPT-valued terminal returns are less reinforced or more strongly penalized if $\Phi_i < 0$. If insufficient terminal returns are available in a batch to reliably compute Φ_i, it defaults to a neutral value (e.g., $\Phi_i = 0$, resulting in $\exp(0) = 1$). The overall algorithm is presented in 5.

.

Algorithm 1 CPT-MADDPG Algorithm

```
Initialize actors μ_{θ_i}, critics Q_{φ_i}, and target networks
Initialize replay buffer D
for episode = 1, ..., M do
    Collect trajectory using policies {μ_{θ_i}} and store in D
    for update step = 1, ..., K do
        Sample minibatch from D
        Compute CPT values C(R_i) using compute_cpt_integral
        Update critics φ_i ← φ_i − η_φ ∇_{φ_i} L(φ_i)
        Update actors θ_i ← θ_i − η_θ ∇_{θ_i} J_CPT(θ_i)
        Soft-update target networks
    end for
end for
```

5.1. Observability-Adjusted CPT Transformation

When an agent is granted access to other agents' utility functions, we replace the single-agent CPT transform $C(R)$ with a cross-agent aggregation. Let $\mathcal{A}$ be the set of agents whose parameters are visible, and for each agent $j \in \mathcal{A}$ let

$$u_j^+(x) = x^{\alpha_j}, \qquad u_j^-(x) = \lambda_j\,(-x)^{\alpha_j},$$

and define constant weights $w'_{j,+}, w'_{j,-}$. For a (flattened) return R, we compute

$$\phi_{\mathrm{cross}}(R) = \frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}} \begin{cases} w'_{j,+}\,u_j^+(R), & R \geq 0, \\ -\,w'_{j,-}\,u_j^-(R), & R < 0. \end{cases}$$

This replaces the usual CPT wrapper in the critic target:

$$y_i \;=\; u_i(r_i) \;+\; \gamma\,\mathbb{E}_{\mathbf{a}'\sim\mu_{\bar{\theta}}}\left[\phi_{\mathrm{cross}}(R'_i)\right].$$

By averaging each visible agent's subjective valuation, the update incorporates observed risk biases directly into the Bellman backup.

5.2. Adaptive Behavioral Parameter Dynamics

To allow each agent's CPT parameters to evolve during training, we parameterize α_i, λ_i, γ_i^+, and γ_i^- as learnable variables and optimize them via a secondary loss. Let $\Theta_i = \{\alpha_i, \lambda_i, \gamma_i^+, \gamma_i^-\}$ and write the usual CPT-adjusted target for agent i as

$$y_i^{\mathrm{CPT}} \;=\; u_i(r_i) \;+\; \gamma\,\mathbb{E}_{\mathbf{a}'\sim\mu_{\bar{\theta}}}\left[C_{\Theta_i}(R'_i)\right],$$

and the standard Bellman target as y_i. We define the *base loss* for the adaptive parameters as

$$L_b^{(t)} = \mathbb{E}_{(s,\mathbf{a},r,s')\sim\mathcal{D}}\left[\left(y_i^{\mathrm{CPT}} - y_i\right)^2\right].$$

To make the parameter updates sensitive to recent changes in this loss, we introduce a *dynamic scaling factor*

$$d^{(t)} = 1 + \left|\,L_b^{(t)} - L_b^{(t-1)}\right|.$$

We also include an ℓ_2 *regularization* term that penalizes deviation from the initial parameter values $\Theta_{i,0}$:

$$L_r = \sum_{p \in \Theta_i} (p - p_{i,0})^2.$$

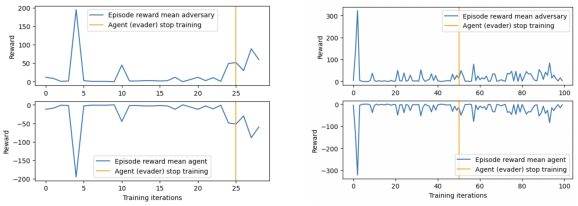
Putting these together, the total loss for the adaptive parameters at iteration t is

$$L_{\text{adapt}}^{(t)} = d^{(t)}\, \kappa\, L_b^{(t)} \;+\; \rho\, L_r,$$

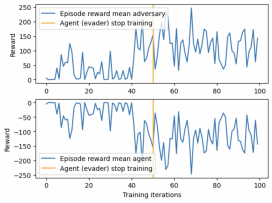
where $\kappa = 10^{-3}$ (scale_factor) and $\rho = 10^{-3}$ (reg_lambda). We then update Θ_i by gradient descent:

$$\Theta_i \;\leftarrow\; \Theta_i - \eta_{\text{adapt}}\, \nabla_{\Theta_i} L_{\text{adapt}}^{(t)}.$$

In practice, we freeze these updates for the first 20 iterations and then apply them every 10 iterations thereafter. This scheme allows the CPT parameters to adapt to the evolving reward landscape while avoiding instability from overly rapid changes.



(a) Baseline MADDPG (b) Moderate CPT



(c) Extreme CPT

Figure 1. Predator average episodic rewards in Simple Tag under baseline, moderate, and extreme CPT.

6. Results

6.1. Experimental Setup

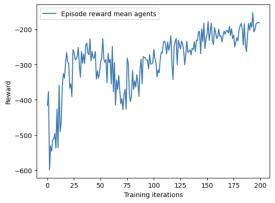
We explore the performance of our CPT-MADDPG algorithm on the environments described in Section 3.2 – namely, the Simple Tag competitive MPE, the Simple Spread cooperative MPE, and a first-price auction scenario. We implement CPT-MADDPG in PyTorch (Paszke et al., 2019) using the TorchRL (Bou et al., 2023) and Vmas (Bettini et al., 2022) libraries. Each actor and critic is a 3-layer MLP (hidden sizes 128–128), with ReLU activations and tanh outputs on actions. Key hyperparameters are $\eta_\theta = 1 \times 10^{-4}$, $\eta_\phi = 1 \times 10^{-3}$, $\gamma = 0.99$, $\tau = 0.01$, $\alpha = \beta = 0.88$, $\lambda = 2.25$. CPT components (`u_plus`, `u_minus`, `w_approx`) and the `compute_cpt_integral` routine are implemented as differentiable PyTorch modules, enabling end-to-end training with Adam.

6.2. Competitive Environment: Simple Tag

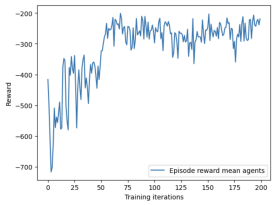
In the Simple Tag predator–prey scenario, our goal is to evaluate how wrapping cumulative returns in CPT transforms influences pursuit strategies in a continuous 2D action space; by comparing moderate (risk-seeking) and extreme (risk-averse) CPT settings against the risk-neutral baseline, we gain insight into how risk preferences trade off exploration versus safety. Figure 1 shows that moderate CPT induces intermittent spikes in episodic predator reward—reflecting willingness to risk zero payoff for potential large gains—whereas extreme CPT dramatically suppresses variance, delays convergence, and lowers mean reward relative to baseline, illustrating pronounced loss aversion and diminished risk tolerance.

6.3. Cooperative Environment: Simple Spread

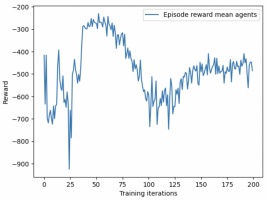
In the Simple Spread cooperative landmark-coverage task, we investigate whether CPT-based risk sensitivity can accelerate coordination without sacrificing stability; by applying moderate and extreme CPT to joint rewards, we gain an understanding of how agents hedge collision risk against coverage gains. As depicted in Figure 2, moderate (risk-seeking) CPT hyperparameter choices accelerate early convergence—agents explore varied positions to balance coverage and collision avoidance—yet stabilizes at similar asymptotic coverage to the baseline, while extreme CPT's strong loss-aversion leads to overly cautious movements and a large reduction in final coverage, highlighting the performance cost of excessive loss sensitivity. Visual comparisons of the MPE environment of the baseline and extreme cases can be found in Figure 7 (Appendix B). In the baseline, agents conduct exploration with limited fear of the negative rewards from losses realized from the distance between them, resulting in a positioning of each agent close to a unique landmark. In the risk-averse case, the agents start exploration but quickly struggle to find a risk-reward tradeoff for optimal positioning near a landmark. The risk-averse agents find themselves unable to take the risk of moving forward closer to a landmark due to the loss-aversion of a change in position from another agent, leading to both agents being unable to position themselves as close to a landmark as in the base case.



(a) Baseline



(b) Moderate CPT

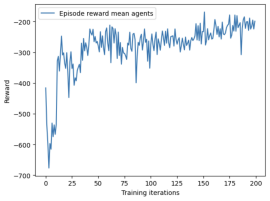


(c) Extreme CPT

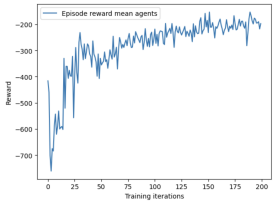
Figure 2. Landmark mean coverage rewards in Simple Spread under baseline, moderate, and extreme CPT.

6.4. Transparent Utility Sharing

Extending Simple Spread with full visibility of peers' CPT utilities, we ask whether transparency of subjective evaluations aligns expectations and preserves cooperative equilibria; this allows us to gain insight into the interplay between heterogeneous risk profiles under shared information. Figure 3 shows that both purely risk-averse pairs and mixed risk-averse/risk-seeking teams converge to the same landmark-coverage trajectory as in opaque training, indicating that observing each other's utility functions mitigates strategic uncertainty without disrupting coordination.



(a) Risk-Averse Seeing

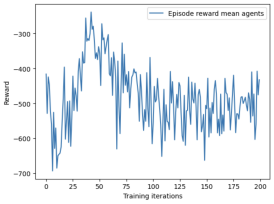


(b) Mixed Risk Profiles Seeing

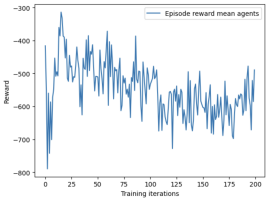
Figure 3. Coverage rewards when agents observe each other's CPT utilities.

6.5. Dynamic CPT Parameters

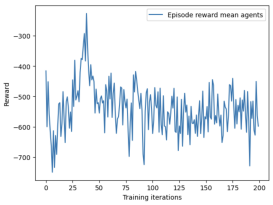
As an extension to the ability of beeing able to observe their cooperative agents utility, we enable agents to update their CPT hyperparameters every ten episodes in the Simple Spread task, aiming to learn whether dynamic profiling can improve performance or introduce instability. Figure 4 illustrates that dynamic risk-seeking, moderate, and high-aversion schedules all produce large oscillations in coverage reward and fail to converge—reward variance exceeds the baseline by over 50%—demonstrating that rapidly shifting risk parameters destabilize multi-agent learning.



(a) Dynamic Risk Seeking



(b) Dynamic Moderate CPT



(c) Dynamic Extreme CPT

Figure 4. Mean Reward trajectories under dynamic CPT hyperparameters.

6.6. First-Price Auction

Building on the empirical findings of Josheski & Apostolov (2023), who demonstrate that CPT-modeled bidders systematically overbid in first-price auctions, we evaluate whether our CPT-MADDPG agents replicate this behavior and its payoff consequences. Each agent's private valuation v_i is drawn uniformly from $[0, 100]$, and we compare three CPT-trained bidders against three risk-neutral agents.

Figure 5 overlays the empirical bid distributions after convergence. Consistent with Josheski & Apostolov's results, the CPT histogram is clearly shifted to the right: modal bids for CPT agents lie around the maximum possible, whereas non-CPT bids cluster either between the lower end and the upper end of the possibel bid spectrum. This shift visually confirms the overbidding effect driven by loss aversion and probability weighting under CPT.

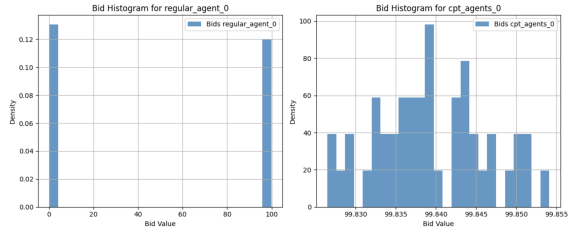


Figure 5. Empirical bid distributions for non-CPT vs. CPT agents (valuations uniform in [0,100]).

Figure 6 shows the average reward over iterations. CPT agents begin with higher payoffs—reflecting frequent wins

from aggressive bids—but rewards decline below zero over time as the cost of overbidding outweighs gains. This turnaround closely matches the long-run loss effects described in Josheski & Apostolov, illustrating the CPT trade-off between short-term advantage and eventual negative returns.

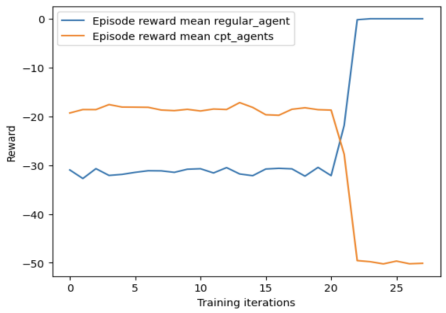


Figure 6. Average episodic reward trajectories for CPT vs. non-CPT agents in the first-price auction.

6.7. Summary of Findings

Across competitive, cooperative, transparency, dynamic, and auction settings, CPT-MADDPG produces rich, parameter-dependent behaviors: moderate (risk-seeking) CPT enhances exploratory learning, extreme (risk-averse) CPT enforces prudence at a performance cost, shared utilities preserve coordination despite risk heterogeneity, and adaptive parameter updates destabilize convergence. Auction results validate that CPT imparts human-like risk biases, granting strategic advantage through overbidding. Detailed hyperparameters for each variant are listed in Table 1 (Appendix A).

7. Discussion

Our empirical evaluation of CPT-MADDPG across competitive, cooperative, and auction domains demonstrates that embedding human-like risk preferences into multi-agent learning yields rich and interpretable behavioral variations. In Simple Tag, risk-seeking CPT drives predators to adopt more aggressive, exploratory tactics—risk-seeking "spikes" in reward—whereas risk-averse driven CPT produces conservative strategies marked by delayed convergence and lower overall payoff. In Simple Spread, moderate risk sensitivity accelerates early coordination at the cost of increased fluctuation, while extreme loss aversion impairs final coverage due to overly cautious movement. Allowing agents to observe each other's CPT utility rewards, assuming that they would adjust their own outcomes in the context of the rewards, shows that transparency of subjective evaluations preserves cooperative equilibria even under heterogeneous risk profiles. Conversely, dynamically adapting CPT parameters on the fly destabilizes learning, suggesting that introducing non-stationarity in risk attitudes undermines policy convergence. Finally, in first-price auctions, CPT-trained bidders systematically overbid—right-shifted bid distributions and an initial reward advantage followed by long-term losses— replicating the expected overbidding phenomenon documented by Josheski & Apostolov (2023).

These results highlight several key insights. First, rank-dependent probability weighting and loss aversion can be effectively integrated into actor–critic updates, endowing agents with tunable risk profiles that mirror human decision biases. Second, while moderate levels of CPT hyperparameters in a risk-seeking setting can enhance exploration and initial learning speed, extreme aversion or overly frequent adaptation of risk parameters imposes tangible performance penalties. Third, transparency of risk preferences between agents need not harm coordination; indeed, shared utility information can align expectations and stabilize behavior. Fourth, CPT-induced overbidding confers short-term auction success but risks eventual negative returns, illustrating the classic "prospect" trade-off in learned policies.

However, our approach has limitations. The empirical approximation of the CPT integral relies on batch-based estimates of tail probabilities, which may introduce bias when returns are sparse or highly skewed. Dynamic hyperparameter adaptation, while conceptually appealing, proved difficult to stabilize and would benefit from more principled schedules or meta-learning frameworks. There may also be an alternative form of providing this hyperparameter adaptation in settings with asymmetric or perfect information. Computational overhead from computing CPT integrals in large multi-agent systems remains non-trivial, suggesting the need for more efficient approximation or hierarchical risk modeling.

Looking forward, several avenues for future work arise. Extending CPT-MADDPG to high-dimensional, continuous control tasks and real-world domains (e.g., autonomous driving or energy management) could validate scalability and practical utility. Incorporating theory-guided meta-learning to tune CPT parameters online may overcome the instability of naïve dynamic schedules. Finally, integrating CPT-based agents with large language models or other human-interactive systems offers a promising path toward more psychologically realistic and interpretable AI agents, capable of anticipating and adapting to human risk behavior in complex multi-agent environments.

In sum, CPT-MADDPG offers a principled framework for endowing reinforcement learning agents with human-aligned risk attitudes, opening new opportunities for interpretable, risk-aware multi-agent systems that bridge the gap between rational optimization and realistic decision-making under uncertainty.

Impact Statement

This work develops CPT-MADDPG, a framework for embedding human-like risk attitudes into multi-agent reinforcement learning. On the positive side, CPT-MADDPG can produce agents whose exploration–exploitation trade-offs more closely mirror human decision patterns, improving safety and interpretability in applications such as autonomous driving, robotic coordination in disaster response, and adaptive traffic management.

However, risk-tuned agents could also be repurposed for adversarial or manipulative ends: for example, algorithmic trading systems that exploit human loss aversion or risk-seeking biases in high-frequency markets. We therefore recommend incorporating human-in-the-loop oversight, transparent reporting of CPT parameters and safeguards to prevent misuse.

8. Acknowledgments

We thank Professor Damek Davis for his guidance and feedback throughout this project. This work was developed as part of his course on optimization and reflects many of the ideas and insights explored therein.

References

Bettini, M., Kortvelesy, R., Blumenkamp, J., and Prorok, A. Vmas: A vectorized multi-agent simulator for collective robot learning, 2022. URL https://arxiv.org/abs/2207.03530.

Borkar, V. S. and Chandak, S. Prospect-theoretic q-learning. *Systems amp; Control Letters*, 156, October 2021. ISSN 0167-6911. doi: 10.1016/j.sysconle.2021.105009. URL http://dx.doi.org/10.1016/j.sysconle.2021.105009.

Bou, A., Bettini, M., Dittert, S., Kumar, V., Sodhani, S., Yang, X., Fabritiis, G. D., and Moens, V. Torchrl: A data-driven decision-making library for pytorch, 2023. URL https://arxiv.org/abs/2306.00577.

Bäuerle, N. and Ott, J. Markov decision processes with average value-at-risk criteria. *Mathematical Methods of Operations Research*, 80(2):281–298, 2014.

Danis, D., Parmacek, P., Dunajsky, D., and Ramasubramanian, B. Multi-agent reinforcement learning with prospect theory. In *2023 Proceedings of the Conference on Control and its Applications (CT)*, pp. 9–16, 2023. doi: 10.1137/1.9781611977745.2. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977745.2.

Du, Y., Han, L., Fang, M., Liu, J., Dai, T., and Tao, D. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/07a9d3fed4c5ea6b17e80258dee231fa-Paper.pdf.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization, 2024. URL https://arxiv.org/abs/2402.01306.

García, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

Ghaemi, H., Kebriaei, H., Moghaddam, A. R., and Ahamdabadi, M. N. Risk-sensitive multi-agent reinforcement learning in network aggregative markov games, 2024. URL https://arxiv.org/abs/2402.05906.

Iqbal, S. and Sha, F. Actor-attention-critic for multi-agent reinforcement learning, 2019. URL https://arxiv.org/abs/1810.02912.

Jie, C., L.A., P., Fu, M., Marcus, S., and Szepesvári, C. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63 (9):2867–2882, 2018. doi: 10.1109/TAC.2018.2822658.

Josheski, D. and Apostolov, M. The cumulative prospect theory and first price auctions: an explanation of overbidding. *Econometrics*, 27:33–74, 2023. doi: 10.15611/eada.2023.1.03.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

L. A., P., Jie, C., Fu, M., Marcus, S., and Szepesvári, C. Cumulative prospect theory meets reinforcement learning: Prediction and control, 2016. URL https://arxiv.org/abs/1506.02632.

Lepel, O. and Barakat, A. Beyond expected returns: A policy gradient algorithm for cumulative prospect theoretic reinforcement learning, 2024. URL https://arxiv.org/abs/2410.02605.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning, 2019. URL https://arxiv.org/abs/1509.02971.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017. URL https://arxiv.org/abs/1706.02275.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Bridging the gap between value and policy based reinforcement learning, 2017. URL https://arxiv.org/abs/1702.08892.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Pu, Y., Wang, S., Yang, R., Yao, X., and Li, B. Decomposed soft actor-critic method for cooperative multi-agent reinforcement learning, 2021. URL https://arxiv.org/abs/2104.06655.

Qiu, W., Wang, X., Yu, R., He, X., Wang, R., An, B., Obraztsova, S., and Rabinovich, Z. Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents, 2021. URL https://arxiv.org/abs/2102.08159.

Ramasubramanian, B., Niu, L., Clark, A., and Poovendran, R. Reinforcement learning beyond expectation, 2021. URL https://arxiv.org/abs/2104.00540.

Rockafellar, R. and Uryasev, S. Conditional value-at-risk for general loss distributions. *Journal of Banking Finance*, 26(7):1443–1471, 2002. ISSN 0378-4266. doi: https://doi.org/10.1016/S0378-4266(02)00271-6. URL https://www.sciencedirect.com/science/article/pii/S0378426602002716.

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, July 2014. ISSN 1530-888X. doi: 10.1162/neco_a_00600. URL http://dx.doi.org/10.1162/NECO_a_00600.

Su, J., Adams, S., and Beling, P. A. Value-decomposition multi-agent actor-critics, 2020. URL https://arxiv.org/abs/2007.12306.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. Policy gradient for coherent risk measures, 2015. URL https://arxiv.org/abs/1502.03919.

Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, October 1992. URL http://ideas.repec.org/a/kap/jrisku/v5y1992i4p297-323.html.

Xiao, Y., Tan, W., and Amato, C. Asynchronous actor-critic for multi-agent reinforcement learning, 2022. URL https://arxiv.org/abs/2209.10113.

A. CPT Function Hyperparameters

Table 1. Hyperparameters for the CPT value and probability-weighting functions across model variants.

Environment	Variant	α	β	λ	γ	δ	$(w^+)'$	$(w^-)'$
Simple Tag	Baseline	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Moderate CPT (risk-seeking)	0.9	0.6	1.5	0.69	0.61	0.8	0.2
	Extreme CPT (risk-averse)	0.88	0.88	2.25	0.61	0.69	0.2	0.8
Simple Spread	Baseline	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Moderate CPT (risk-averse)	0.88	0.88	2.25	0.61	0.69	0.2	0.8
	Extreme CPT (risk-averse)	0.7	0.95	2.5	0.61	0.69	0.2	0.8
	Observability CPT (Seeing - RS Agent)	0.7	0.7	0.8	0.61	0.69	0.8	0.2
	Observability CPT (Seeing - RA Agent)	0.65	0.65	2.8	0.61	0.69	0.25	0.75
	Dynamic (Agent 1)	0.7	0.7	2.5	0.61	0.69	0.8	0.2
	Dynamic (Agent 2)	0.65	0.65	2.8	0.61	0.69	0.8	0.2
	Dynamic Moderate (Agent 1)	0.6	0.6	1	0.5	0.55	0.2	0.8
	Dynamic Moderate (Agent 2)	0.3	0.3	1.5	0.5	0.55	0.2	0.8
	Dynamic Extreme (Agent 1)	1.2	1.2	1.2	0.5	0.69	0.2	0.8
	Dynamic Extreme (Agent 2)	0.3	0.3	1.5	0.5	0.69	0.2	0.8
Auction	CPT Agents	0.88	0.88	2.25	0.61	0.69	N/A	N/A
	Non-CPT Agents	N/A	N/A	N/A	N/A	N/A	N/A	N/A

B. MPE Visualizations in Cooperative Setting

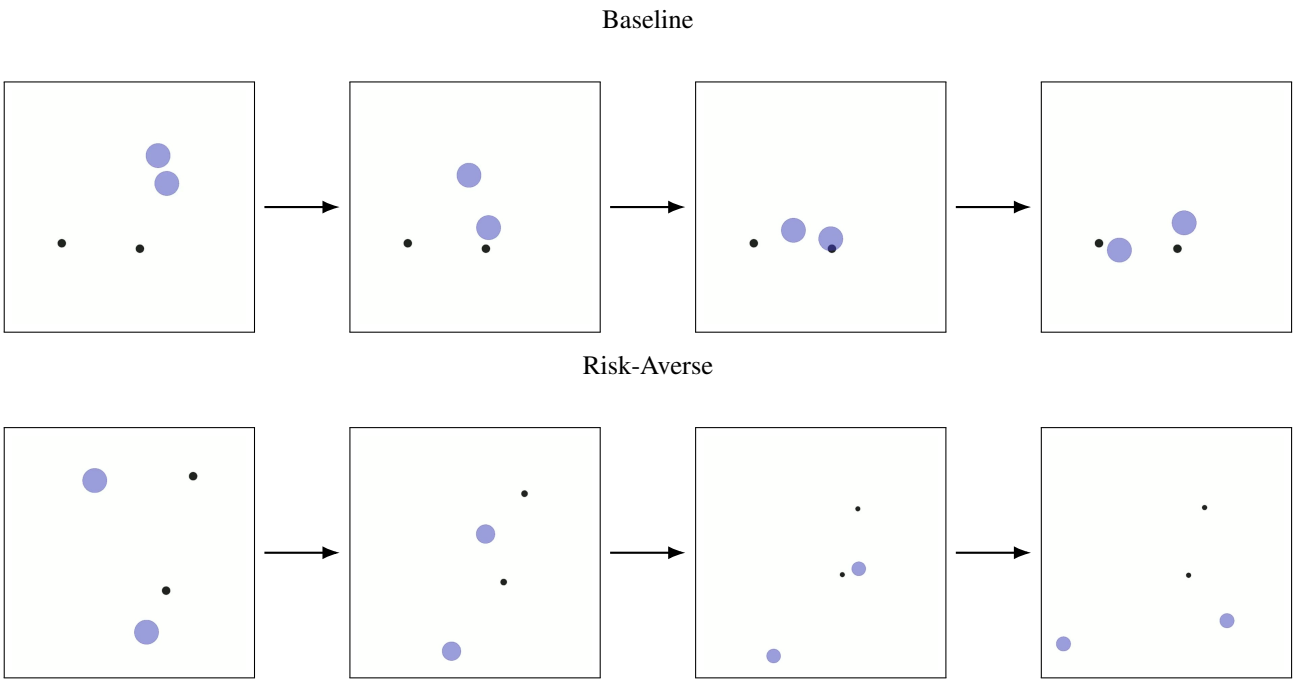


Figure 7. Simple Spread MPE steady state approach observed in Baseline (no risk profile) and Risk Averse (Extreme CPT) cases.