PER-AGENT RETURN DISTRIBUTIONS IN VALUE DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributional value decomposition for cooperative multi-agent reinforcement learning predicts a return distribution for each agent and combines the per-agent quantiles into a team distribution and risk-sensitive variants act on the per-agent distributions. We show that training on the team return does not identify them: at every state many decompositions give the same team quantiles and any decision rule that uses only team-level data violates a per-agent risk constraint with probability at least one half in some environment. Supervising each agent with its own return removes this ambiguity, but summing quantiles at matched levels, as DFAC-style mixers do, is comonotone recomposition, which for finite-variance returns overstates the variance of the team return unless agent returns are comonotone. We propose a critic that trains per-agent quantile heads on individual returns and predicts the team distribution with a separate head. In a heterogeneous search task and in Gigastep, team-only training left per-agent coverage far from nominal in most settings and dependent on the policy and on how credit was shared. The proposed critic lowered per-agent CRPS in every setting, kept team coverage at the team-only level where supervised matched-sum critics raised it and improved per-agent risk decisions in the search task.

1 INTRODUCTION

A team of heterogeneous robots may need to limit the risk that falls on individual members, such as the chance that a quadrotor passes close to a person or that a damaged unit is sent to the most exposed task. Risk-aware multi-robot allocation already works with chance constraints and conditional value at risk (Yang & Chakraborty, 2017; Zhou & Tokekar, 2022). In cooperative multi-agent reinforcement learning, distributional value decomposition predicts a return distribution for each agent. DFAC (Sun et al., 2021) gives each agent a quantile critic and combines the per-agent quantile functions into the team's, RiskQ (Shen et al., 2023) acts on distorted per-agent quantiles and RMIX (Qiu et al., 2021) learns risk-sensitive policies from per-agent return distributions. The critics are fitted to the team return and the per-agent distributions receive gradients only through the mixer.

Such training does not determine the per-agent distributions (Section 3). At each state, a mixer that sums quantiles at matched levels returns the same team quantiles for a whole subspace of per-agent quantile assignments and the ambiguity affects decisions: under a per-agent risk constraint, any rule that uses team-level data alone selects an infeasible allocation with probability at least one half in one of two environments. Earlier analyses noted that individual utilities are fixed only up to per-agent affine maps or zero-sum residues (Son et al., 2019; Wang et al., 2021b) and set this aside because greedy joint actions do not change. Per-agent risk differs within that subspace (Figure 1).

Supervising each head with the agent's own return removes this ambiguity, but the mixer then imposes a dependence structure. The sum of quantile functions at matched levels is the quantile function of the comonotone sum (Dhaene et al., 2002), the largest in convex order among sums with those marginals. With correct marginals and finite variances the matched sum overstates the variance of the team return unless agent returns are comonotone and a matched-sum critic cannot fit both the per-agent and the team targets exactly. Without any learned critic, the matched sum of Monte Carlo marginals covered 0.995 of team returns at nominal level 0.90625 in a task whose agent returns had a mean within-state rank correlation of -0.04, at 2.24 times the width of the team interval.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Three decompositions of one team return. In each row the two agents' return distributions are summed at matched quantile levels (Σ) and give the same team distribution, so a critic trained on the team return cannot distinguish A, B and C. Their per-agent risks differ: the shaded lower tail and the tick below each baseline mark the agent's conditional value at risk at level 0.1, coral where it falls below the budget of -6.0 (dashed). Under A and B one agent crosses the budget and under C neither does (Proposition 2). Curve widths are schematic; the values are in Section 3.

The decoupled supervised critic (Section 4) trains per-agent quantile heads on individual returns and a separate team head on the team return, with no mixer between them. In a heterogeneous search task and in Gigastep (Lechner et al., 2023), team-only training left per-agent coverage between 0.26 and 0.89, a matched sum with per-agent supervision raised team coverage at every tested weight and the decoupled supervised critic brought per-agent coverage to between 0.87 and 0.96 without raising team coverage. It also lowered decision error in the search task and improved within-state ordering in both environments. Our contributions are:

- a pointwise non-identification result for per-agent return distributions under team-only training with matched-sum and mean-shape mixers, a lower bound of one half on the error of team-only rules for a per-agent risk constraint and the identification of matched quantile summation as comonotone recomposition, with an oracle measurement of its effect (Section 3);
- the decoupled supervised critic, in which per-agent supervision removes the decomposition ambiguity and a separate team head removes the recomposition constraint, inside a centralised actor-critic (Algorithm 1);
- an evaluation in two environments with two frozen policies each, a credit-split sweep, capacity and input controls, proper scores and paired decision experiments (Section 5).

2 Background

Cooperative multi-agent RL and value decomposition. We consider a decentralised partially observable Markov decision process (Bernstein et al., 2002) with n agents and a factored reward, where agent i observes $o_{i,t}$, acts with a reactive policy $\pi(a_{i,t} \mid o_{i,t})$ and receives an individual reward $r_{i,t}$ and the team reward is $r_t = \sum_i r_{i,t}$. Training is centralised and execution decentralised (Lowe et al., 2017); we train policies with PPO (Schulman et al., 2017) and a centralised critic as in MAPPO (Yu et al., 2022). VDN (Sunehag et al., 2018) writes the team value as a sum of per-agent utilities and QMIX (Rashid et al., 2020) as a monotone mixture of them; both satisfy the individual-global-max condition formalised by QTRAN (Son et al., 2019).

Distributional critics. A quantile critic represents a return distribution by K quantiles $\theta_1 \le \cdots \le \theta_K$ at the levels $\tau_k = (2k-1)/(2K)$ (Dabney et al., 2018b), trained against a distributional Bellman

target (Bellemare et al., 2017) $y \in \mathbb{R}^K$ with the quantile Huber loss

$$\rho^\kappa(\theta, y) = \sum_{k=1}^{K} \frac{1}{K} \sum_{j=1}^{K} \left|\tau_k - \mathbf{1}\{y_j < \theta_k\}\right| \frac{L_\kappa(y_j - \theta_k)}{\kappa}, \tag{1}$$

where L_κ is the Huber function with threshold κ. DFAC (Sun et al., 2021) extends VDN and QMIX to distributions: DDN sums the per-agent quantile functions at a shared level and DMIX mixes the per-agent means with QMIX and adds the summed deviations from the means. RiskQ (Shen et al., 2023) forms each team quantile as a weighted sum of per-agent quantiles and acts on distorted quantiles and RMIX (Qiu et al., 2021) learns risk-sensitive policies from per-agent return distributions. In DFAC and RiskQ each agent's distribution is trained only through the team prediction.

Risk and calibration. The conditional value at risk of a return G at level α is the mean of its lower tail, $\mathrm{CVaR}_\alpha(G) = \alpha^{-1} \int_0^\alpha F^{-1}(u)\, du$ (Rockafellar & Uryasev, 2002). A predicted distribution is calibrated at a central interval if the realised return falls inside it at the nominal rate. With $K = 32$ the grid contains the levels 0.046875 and 0.953125, so the central interval between them has nominal level 0.90625 and we report its empirical coverage. As a proper score we use the continuous ranked probability score (Gneiting & Raftery, 2007), approximated on the grid by $\mathrm{CRPS}(\theta, g) = \frac{2}{K} \sum_k (\tau_k - \mathbf{1}\{g < \theta_k\})(g - \theta_k)$ for a realised return g; unlike coverage, it penalises error in every quantile, including over-wide predictions (Gneiting et al., 2007).

3 Per-agent return distributions trained through the team return

Agent i's return is $G_i = \sum_t \gamma^t r_{i,t}$ and the team return is $G = \sum_i G_i$. At a state s, agent i's head outputs quantiles $\theta_{i,1} \le \cdots \le \theta_{i,K}$ at the levels τ_k. DDN and DMIX define two classes of mixer and VDN and QMIX applied to the per-agent means, with the deviations from the means summed, give the same two. A matched-sum mixer (DDN) computes

$$\theta_k^{\mathrm{tot}} = \sum_{i=1}^{n} \theta_{i,k}, \tag{2}$$

and a mean-shape mixer (DMIX) computes $\theta_k^{\mathrm{tot}} = f(\mu_1, \ldots, \mu_n; s) + \sum_i (\theta_{i,k} - \mu_i)$, where $\mu_i = K^{-1} \sum_k \theta_{i,k}$ and f is non-decreasing in each μ_i. Team-only training fits θ^{tot} to the team return and observes the per-agent heads only through the mixer output.

Definition 1 (Observational equivalence)**.** *Two per-agent quantile assignments θ and θ' are observationally equivalent under a mixer at state s if the mixer returns the same team quantiles for both. A loss that depends on the per-agent heads only through the team quantiles assigns equivalent assignments equal value.*

Proposition 1 (Non-identification)**.** *Fix a state s and a per-agent assignment $\theta \in \mathbb{R}^{n \times K}$ with strictly increasing rows. Every $\Delta \in \mathbb{R}^{n \times K}$ with $\sum_i \Delta_{i,k} = 0$ for every k and sufficiently small entries makes $\theta + \Delta$ observationally equivalent to θ under a matched-sum mixer; if in addition $\sum_k \Delta_{i,k} = 0$ for every i, also under a mean-shape mixer. The assignments equivalent to θ therefore contain a neighbourhood of θ in an affine subspace of dimension $(n-1)K$ for the matched sum and $(n-1)(K-1)$ for the mean-shape mixer, while the mixer output has K coordinates.*

The proof is a substitution into both mixer forms (Section B). The result is pointwise in output space. It applies to trained critics when the head class can represent the alternatives, which heads that share parameters and receive identical inputs cannot. A weighted matched sum, as in RiskQ, has a subspace of the same kind, $\sum_i \omega_i \Delta_{i,k} = 0$. In both environments agents receive different observations and in the search task each observation includes a capability vector.

The ambiguity is not confined to directions that leave decisions unchanged. Let $Z \sim \mathcal{N}(0, 1)$ and consider three decompositions of one two-agent team return, $A = (3.8Z, 0.2Z)$, $B = (0.2Z, 3.8Z)$ and $C = (2.0Z, 2.0Z)$. Their components are comonotone, so their quantile functions add exactly: all three give the team return $4.0Z$ with zero means and both mixer classes return identical team quantiles for them. The individual risks differ: at level $\alpha = 0.1$ the per-agent conditional value

at risk is $(-6.67, -0.35)$ under A, $(-0.35, -6.67)$ under B and $(-3.51, -3.51)$ under C and the separation survives the 32-level grid (agent 1: -6.65, -0.35, -3.50).

Proposition 2 (Decision lower bound). *Require every agent's conditional value at risk at level 0.1 to be at least -6.0, so that C is feasible and A and B are not. Consider two environments with two available allocations whose per-agent returns follow (C, A) in the first environment and (A, C) in the second. Any decision rule, deterministic or randomised, whose input is team-level data alone selects an infeasible allocation with probability at least one half in at least one of the two environments.*

Both environments produce identical team-level data for both allocations, so the bound applies to any team-only method, independently of training (Section B). Supervising each head against its own return removes the ambiguity where the conditional quantiles are unique, since each is the minimiser of its pinball loss (Koenker & Bassett, 1978). It does not make the matched sum consistent with the team return. Random variables $(X_1, \dots, X_n)$ are comonotone if they are distributed as $(F_1^{-1}(U), \dots, F_n^{-1}(U))$ for a single U uniform on $(0, 1)$ and $X \leq_{\mathrm{cx}} Y$ in convex order if $\mathbb{E}\,\varphi(X) \leq \mathbb{E}\,\varphi(Y)$ for every convex φ for which the expectations exist.

Proposition 3 (Matched summation is comonotone recomposition). *Let F_i be the distribution of G_i given a conditioning variable h shared by all agents, with finite mean. The function $\tau \mapsto \sum_i F_i^{-1}(\tau)$ is the quantile function of the comonotone sum $S^c = \sum_i F_i^{-1}(U)$ and $\sum_i G_i \leq_{\mathrm{cx}} S^c$ given h for every joint distribution of $(G_1, \dots, G_n)$ with these marginals (Dhaene et al., 2002, Theorems 5 and 7). If the G_i are jointly Gaussian given h with common standard deviation σ and pairwise correlation $\rho \geq -1/(n-1)$ and every head is Gaussian with the correct mean and a common scale σ', the matched sum equals the team quantile function exactly when $\sigma' = \sigma\, c(n, \rho)$ with $c(n, \rho) = \sqrt{(1 + (n-1)\rho)/n}$. Each head's central interval at nominal level $1 - \delta$ then covers $2\Phi(c\, z_{1-\delta/2}) - 1$ of the agent's returns.*

With finite variances, convex order at equal means implies that S^c has at least the variance of the team return, with equality only when the returns are comonotone (Section B). A matched sum of correct marginals therefore overstates the variance of the team return otherwise; convex order does not imply that every central interval is wider, so we measured it. In the other direction, a matched sum fitted to the team return needs heads narrower than the marginals, by the factor $c(n, \rho)$ in the Gaussian case. Unless the team's quantile function equals the matched sum of the per-agent quantile functions, a critic that forms its team quantiles by Equation (2) cannot fit both targets exactly.

We measured the first direction without any learned critic. From a frozen policy we drew $2{,}048$ complete episodes from each of 128 initial states, so that the marginals and the team return share the simulator state as h. We estimated each agent's quantiles on half of the episodes and scored central intervals at nominal level 0.90625 on the other half (Figure 2). In the search task, where the mean within-state rank correlation between agents' returns was -0.04, per-agent intervals covered 0.904 and the Monte Carlo team interval 0.906, while the matched sum of the same per-agent quantiles covered 0.995 at 2.24 times the width. In Gigastep, where returns co-moved (rank correlation 0.75), the matched sum covered 0.948 at 1.12 times the width. Changing only how Gigastep divides the reward for a defeated hider among the seekers, with the policy, the states and the team return fixed, moved the rank correlation to -0.13 and the width ratio to 5.13 (Figure 2b).

4 The decoupled supervised critic

Section 3 gives two requirements: each head needs a target of its own and the team prediction must not be tied to a recomposition of the heads. The decoupled supervised critic meets the first with per-agent supervision and the second with a separate team head and Section 5 measures the effect of each. It applies wherever individual rewards $r_{i,t}$ are observable during centralised training, as in both environments we study.

Architecture. Agent i's head maps its observation o_i (including a capability vector in the search task) to K sorted quantiles $\theta_i(o_i)$; the head's parameters are shared across agents. A separate team head maps the team observation $\bar{o} = (o_1, \dots, o_n)$, which our implementation also passes to mixers as the state s, to K sorted quantiles $\theta^{\mathrm{tot}}(\bar{o})$. No mixer connects the two, so the team quantiles

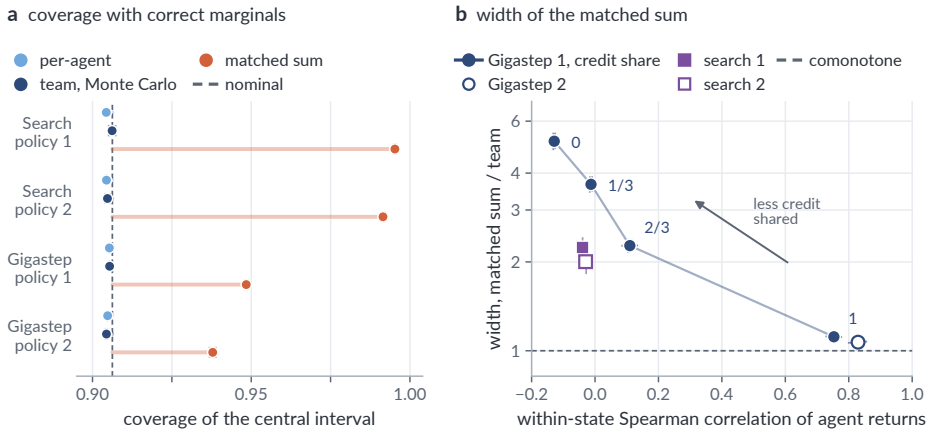


Figure 2: Recomposition with correct marginals, estimated from 2,048 complete episodes per initial state at 128 initial states; each setting has two samples of initial states. (a) Coverage of central intervals at nominal level 0.90625 (dashed): per-agent intervals, the Monte Carlo team interval and the matched sum of the per-agent quantiles. (b) Width of the matched-sum interval relative to the team interval, on a logarithmic axis, against the Spearman correlation of agent returns, computed within each initial state, averaged over agent pairs and then over states, with state-block bootstrap 95% intervals. Labels give the share of the defeat reward paid equally to all seekers; share 1 is policy 1.

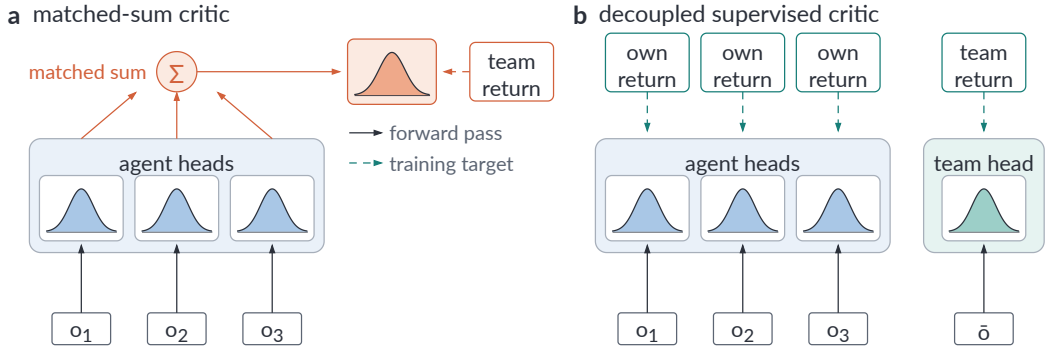


Figure 3: (a) The matched-sum critic sums the per-agent heads level by level and fits the sum to the team return, so the per-agent heads have no target of their own. (b) The decoupled supervised critic fits each agent head to the agent's own return and a separate team head, which reads the team observation $\bar{o}$, to the team return, with no connection between the two. Solid arrows are the forward pass and dashed arrows training targets.

are not constrained to equal any function of the per-agent quantiles. Figure 3 contrasts the two arrangements.

Targets. Both heads regress on multi-step distributional targets computed from the rollout and treated as constants. For a rollout of length T, the per-agent target is

$$y_{i,t} = \mathrm{sort}\big(r_{i,t} + \gamma(1 - d_t)\, y_{i,t+1}\big), \qquad y_{i,T} = \theta_i(o_{i,T}), \tag{3}$$

applied elementwise over the K quantiles, where d_t marks episode ends. The team target y_t^{tot} follows Equation (3) with the team reward $r_t = \sum_i r_{i,t}$ and the bootstrap $\theta^{\mathrm{tot}}(\bar{o}_T)$. Replacing the team target by the scalar λ-return, with the per-agent targets unchanged, leaves both heads under-covering without per-agent supervision (Section J).

Algorithm 1 One update of the decoupled supervised critic. The same loop with a mixer in place of the team head gives the mixer critics and $w = 0$ gives team-only training. In our experiments the policy is loaded and frozen, so only ψ and ω are updated and the two policy terms of $\mathcal{L}$ are inactive.

Require: policy π_ϕ, per-agent head θ_ψ, team head $\theta^{\mathrm{tot}}_\omega$, levels $\tau_{1:K}$, weights c_v, w, c_H, discount γ, trace λ

for $t = 0, \dots, T-1$ in E parallel environments **do**
  $\bar{o}_t \leftarrow (o_{1,t}, \dots, o_{n,t})$; $\quad a_{i,t} \sim \pi_\phi(\cdot \mid o_{i,t})$ for every agent i
  step all environments; record $o_{i,t}$, $a_{i,t}$, $\log \pi_\phi(a_{i,t} \mid o_{i,t})$, $r_{i,t}$, $r_t = \sum_i r_{i,t}$, d_t, $\theta^{\mathrm{tot}}_\omega(\bar{o}_t)$
end for
$V_t \leftarrow K^{-1} \sum_k \theta^{\mathrm{tot}}_{\omega,k}(\bar{o}_t)$ for $t \le T$; $\quad \hat{A}_t \leftarrow \mathrm{GAE}_{\gamma,\lambda}(r, V, d)$
$y^{\mathrm{tot}}_t \leftarrow \mathrm{sort}(r_t + \gamma(1 - d_t)\, y^{\mathrm{tot}}_{t+1})$ with $y^{\mathrm{tot}}_T = \theta^{\mathrm{tot}}_\omega(\bar{o}_T)$ treated as a constant
$y_{i,t} \leftarrow \mathrm{sort}(r_{i,t} + \gamma(1 - d_t)\, y_{i,t+1})$ with $y_{i,T} = \theta_\psi(o_{i,T})$ treated as a constant
for each epoch and minibatch **do**
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{PPO}}(\phi; \hat{A}) + c_v\, \rho^\kappa(\theta^{\mathrm{tot}}_\omega(\bar{o}), y^{\mathrm{tot}}) + \frac{w}{n} \sum_i \rho^\kappa(\theta_\psi(o_i), y_i) - c_H\, \mathcal{H}(\pi_\phi)$
  take a gradient step on (ψ, ω) and on ϕ when the policy is not frozen
end for

Objective. With the quantile Huber loss ρ^κ of Equation (1), the critic loss is

$$\mathcal{L}_{\mathrm{critic}} = c_v\, \rho^\kappa\big(\theta^{\mathrm{tot}}(\bar{o}_t), y^{\mathrm{tot}}_t\big) + \frac{w}{n} \sum_{i=1}^{n} \rho^\kappa\big(\theta_i(o_{i,t}), y_{i,t}\big), \tag{4}$$

where w weights per-agent supervision. The policy is trained with the clipped PPO objective (Schulman et al., 2017). Its advantage is computed by generalised advantage estimation (Schulman et al., 2016) on the team reward, with the mean of the team quantiles as the value and every agent in an environment receives the same advantage, the centralised-critic arrangement of MAPPO (Yu et al., 2022); the per-agent loss has no direct gradient path to the actor, so the policy is risk-neutral (Algorithm 1).

Relation to mixer critics. Replacing the team head by the matched sum of Equation (2) and keeping Equation (4) gives the critics we compare against; a mean-shape mixer fits the same template. At $w = 0$ this is team-only training, whose per-agent heads Proposition 1 leaves unidentified. At $w > 0$ both terms reach the per-agent heads and by Proposition 3 a matched-sum critic cannot fit both targets exactly unless the team's quantile function is the matched sum of the per-agent ones. The decoupled supervised critic gives up one property in exchange: its team quantiles need not be coherent with any coupling of its per-agent quantiles.

What the heads estimate. Where agent i's conditional quantiles are unique, they minimise the per-agent pinball loss (Koenker & Bassett, 1978); the quantile Huber loss with $\kappa = 1$ (Dabney et al., 2018b) is a smoothed surrogate for it. The per-agent heads estimate the distribution of G_i given o_i and the team head that of G given $\bar{o}$. Under partial observability the conditioning information differs, so the per-agent heads are not the marginals of the team head's distribution; Section 5 includes a control whose per-agent heads also receive $\bar{o}$.

5 Experiments

The experiments test three consequences of Sections 3 and 4: team-only training leaves the per-agent distributions uncontrolled; per-agent supervision through a matched sum moves team coverage away from nominal and the decoupled supervised critic does not; and the per-agent distributions change decisions that depend on per-agent risk.

5.1 Setup

Environments. In the search task, written in JAX (Bradbury et al., 2018), four quadrotors and two quadrupeds search a 120 m arena for 40 targets among 12 moving humans; each agent is rewarded for the targets it detects and penalised near humans, in collisions and for control effort. In Gigastep

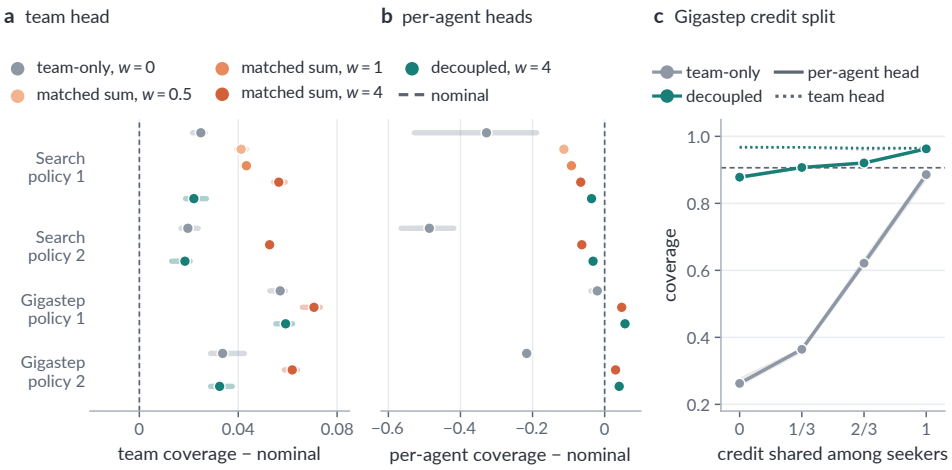


Figure 4: (a, b) Coverage minus the nominal level 0.90625 for the team head and the per-agent heads of each critic: the dot is the mean over training seeds and the bar the range over seeds. (c) Coverage of the team-only critic and the decoupled supervised critic across the Gigastep credit split under policy 1: per-agent heads solid with the range over seeds shaded, team heads dotted (the two nearly coincide), nominal level dashed.

hide-and-seek (Lechner et al., 2023), five seekers of three types chase five scripted hiders; damage is rewarded to the seeker that deals it and the defeat reward and the penalty for a disabled teammate are shared among live seekers. Agent returns are weakly rank-correlated in the search task and co-move in Gigastep (Figure 2b). Section D gives the details.

Protocol. Critics trained under different policies are trained on different data, so we fix the policy. In each environment we freeze two trained policies, one trained with a matched-sum critic (policy 1) and one with a QMIX critic (policy 2) and train each critic from scratch on their rollouts with Algorithm 1 for 90,000 updates. The arms are the matched sum at per-agent weight $w = 0$ (team-only) and $w = 4$ and the decoupled supervised critic at $w = 4$, plus the matched sum at $w \in \{0.5, 1\}$ under search policy 1. Search policy 1 has 10 training seeds per main arm and 4 for the other weights; every other cell reported in this section has 6.

Evaluation. Calibration is scored on held-out episodes. In the search task, whose training episodes start at a random step, predictions are scored at the distribution of remaining horizons seen in training; in Gigastep, whose episodes all start at step 0, they are scored at the initial state (Section E). Arms under one policy are evaluated on identical episodes, so we report matched-seed differences with pointwise 95% intervals from a two-way bootstrap that resamples training seeds and, independently, the shared episodes. Full tables and the GPU model of each cell are in Sections F and K.

5.2 Calibration

Team-only training leaves the marginals uncontrolled. Without per-agent supervision the team heads covered 0.926 to 0.963, in the range of the supervised critics, while per-agent coverage ranged from 0.420 to 0.886 (Figure 4a, b). Under search policy 1 its standard deviation across training seeds was 0.119, against 0.002 and 0.001 with per-agent supervision. The credit-split sweep keeps the policy, the states and the team return fixed and changes only how Gigastep divides the defeat reward: team-only per-agent coverage fell from 0.886 with all of it shared to 0.263 with none, while team coverage stayed between 0.961 and 0.966 (Figure 4c, Section H). Against the decoupled supervised critic, the team-only critic's team CRPS differed by at most 0.19 with every interval containing zero and its per-agent CRPS was higher by 0.42 to 6.81 with every interval excluding zero.

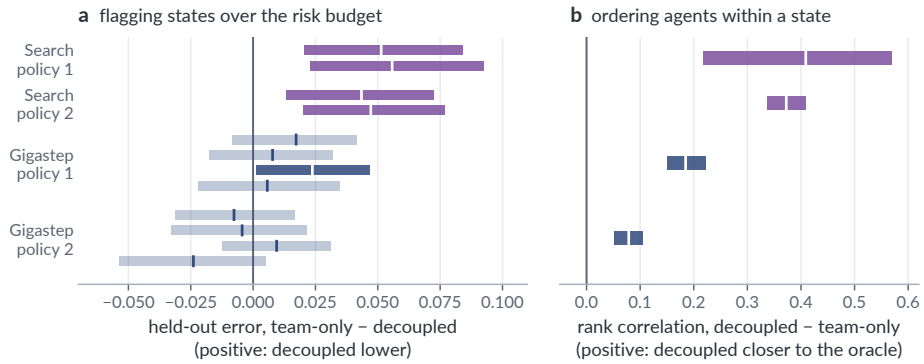


Figure 5: Team-only critic against the decoupled supervised critic, matched training seeds. Bars are pointwise 95% intervals of the mean paired difference from a two-way bootstrap over training seeds and states, ticks the means; solid bars exclude zero. A state seed is an independent draw of the initial states. (a) Held-out error of flagging states whose worst agent falls below the risk budget, one bar per state seed (13 and 14 in the search task, 11 to 14 in Gigastep). (b) Within-state Spearman correlation with the oracle's ordering of agents, state seed 12. Table 16 lists the values.

Matched sums trade team for per-agent calibration. Adding per-agent supervision to the matched sum raised per-agent coverage and raised team coverage above that of team-only training at every tested weight and in every setting. Under search policy 1, weights 0.5, 1 and 4 gave team coverage 0.947, 0.950 and 0.963 against 0.931 at $w = 0$ and per-agent coverage 0.793, 0.814 and 0.840. Even with correct marginals the matched sum imposes the comonotone coupling and the oracle experiment measured how much this coupling widens the team interval (Figure 2).

The decoupled supervised critic keeps team coverage at the team-only level. At the same weight, its team coverage stayed at the team-only level and its team CRPS was lower than that of the matched sum at $w = 4$ in all four settings, by 1.36, 1.20, 0.31 and 0.53, with every interval excluding zero. Per-agent CRPS was within 0.013 of the matched sum's; per-agent coverage was closer to nominal in the search task and further above it in Gigastep, where both supervised critics over-covered. A wider team-only critic and a team-only critic whose per-agent heads also read the team observation left per-agent coverage at 0.614 and 0.598, against 0.562 for the default critic on the same seeds (Section G). With that input the matched sum at $w = 4$ had a per-agent CRPS lower than the decoupled supervised critic's by 0.29 and a team CRPS lower by 0.16 with an interval containing zero, while further from nominal coverage on both heads; part of the team-CRPS advantage above may come from the team observation as input.

5.3 Decisions

Two decisions depend on per-agent risk. Both are scored against a Monte Carlo oracle of 512 continuations per state at $1{,}024$ initial states, with the conditional value at risk at level 0.1 (Section I). Across states, a critic flags the states whose worst agent falls below a budget set at the oracle's median, with its threshold fitted on half of the states and scored on the other half. Within a state, a critic orders the agents by risk, scored by the Spearman correlation with the oracle's ordering averaged over states. Decision states are initial states, which share one remaining horizon (Section E).

The decoupled supervised critic changes decisions about agents. In the search task the team-only critic's held-out error was near the chance level of 0.5 (0.449 to 0.488) and the decoupled supervised critic lowered it at both state seeds under both policies, with every interval excluding zero (Figure 5a). Within a state, the team-only critic's ordering of agents was anti-correlated with the oracle's (-0.11 and -0.10) and that of the decoupled supervised critic positively correlated (0.30 and 0.27). In Gigastep the across-state differences were small and of mixed sign, while the within-state ordering improved under both policies (Figure 5b). As an exploratory, post hoc reading, the across-state decision in Gigastep follows a team-level quantity, how exposed the hiders are, which

the team head also learns. The matched sum at $w = 4$ had no consistent advantage: its held-out error minus the decoupled supervised critic's ranged from -0.018 to $+0.008$ with mixed signs.

6 Related work

Value decomposition and credit. VDN, QMIX and QPLEX (Wang et al., 2021c) factorise the team value under the individual-global-max condition and QFIX (Baisero et al., 2025) extends VDN and QMIX mixers to the full class that condition allows. Convergence analyses concern the joint value and the greedy policy (Dou et al., 2022). Non-uniqueness of the individual utilities has been noted in two forms: the condition is invariant to per-agent positive affine maps (Son et al., 2019) and linear factorisation fixes individual Q-values only up to a zero-sum residue (Wang et al., 2021b). Both are treated as harmless for greedy action selection. Credit assignment shapes what each agent learns from a shared reward, through counterfactual baselines (Foerster et al., 2018), difference utilities (Wolpert & Tumer, 2001) or Shapley values (Wang et al., 2022). COSAC (Deshmukh et al., 2026) finds per-agent credit under-determined for agents acting in a fixed order with one team reward and CAST (Zhang et al., 2026) finds individual rewards identifiable in a causal return decomposition. These results concern rewards and greedy values; we study per-agent return distributions.

Distributional and risk-sensitive multi-agent RL. DFAC (Sun et al., 2021) introduced DDN and DMIX and its extension (Sun et al., 2023) notes that adding quantile functions combines the per-agent variables with a specific dependence structure. RiskQ (Shen et al., 2023) forms team quantiles as learned non-negative combinations of per-agent quantiles and notes that risk measures add only for comonotone variables; neither measures per-agent calibration. DRIMA (Son et al., 2022) separates the risk from teammates' actions from environmental risk, which it models jointly; RMIX (Qiu et al., 2021) learns CVaR policies from per-agent return distributions; and Geng et al. (2025) decompose a global reward distribution into individual ones and add penalties because many decompositions fit. Ghaemi et al. (2024) survey risk sensitivity in Markov games. In single-agent RL, multivariate distributional methods learn the joint distribution of vector-valued returns (Zhang et al., 2021; Wiltzer et al., 2024) and copulas separate dependence from marginals (Nelsen, 2006), as in multi-agent imitation learning (Wang et al., 2021a). The decoupled team head is simpler than a copula and gives up coherence between the two.

Identification and calibration. The non-identification result belongs to the identification literature (Rothenberg, 1971; Manski, 2003); Skalse et al. (2023) treat the analogous problem for reward learning. Section A compares the closest work and covers calibration and decision-focused evaluation.

7 Limitations and conclusion

The decoupled supervised critic needs individual rewards in training, which many cooperative benchmarks do not expose. We froze the policies and compared matched-sum critics only, so the results do not show how the critic changes policy learning or how mean-shape critics behave. The evidence comes from two environments with two policies each, Gigastep calibration was scored at initial states and the non-identification result is pointwise in output space. In Gigastep the decoupled supervised critic improved the within-state ordering and not the across-state decision. The team head is not tied to the per-agent heads, so its predicted mean need not equal the sum of the per-agent means. Same-seed reruns and a single-commit replication moved team-only per-agent coverage; supervised cells moved by at most 0.002 (Section G). The decision experiments do not train risk-sensitive policies, whose usual CVaR-greedy form has known limits (Lim & Malik, 2022); quantile critics need not represent true quantiles (Rowland et al., 2019), so we measured calibration directly.

Team-only training does not identify per-agent return distributions (Proposition 1). In the tested settings their coverage ranged from 0.26 to 0.89 and moved across seeds and reruns while team scores did not. In every tested matched-sum setting per-agent supervision raised team coverage and the oracle experiment supports comonotone coupling. Supervising each agent with its own return and predicting the team distribution with a separate head lowered per-agent CRPS in every setting, kept team coverage at the team-only level and improved within-state ordering in both environments and across-state decisions in the search task.

REPRODUCIBILITY STATEMENT

Section B contains the proofs, Section C the algorithm details and hyperparameters, Section D the environments and Section E the evaluation protocol. Every reported result is generated by a script from result files, design constants are taken from the run configurations and Section K lists the result files and the training runs behind them. The supplementary material contains the environments, critics, training, evaluation and decision scripts, the result files and run configurations behind every reported number and the scripts that regenerate the tables, macros and figures from them.

USE OF LARGE LANGUAGE MODELS

Large language models were used to debug code and to edit the text. The authors take full responsibility for the content.

REFERENCES

Andrea Baisero, Rupali Bhati, Shuo Liu, Aathira Pillai, and Christopher Amato. Fixing Incomplete Value Function Decomposition for Multi-Agent Reinforcement Learning, May 2025. URL https://arxiv.org/abs/2505.10484v1.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458, Sydney, NSW, Australia, August 2017. JMLR.org. URL https://dl.acm.org/doi/10.5555/3305381.3305428.

Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4): 819–840, 2002. doi: 10.1287/moor.27.4.819.297.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1522–1530, Cambridge, MA, USA, December 2015. MIT Press.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond Pinball Loss: Quantile Methods for Calibrated Uncertainty Quantification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10971–10984. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/5b168fdba5ee5ea262cc2d4c0b457697-Abstract.html.

Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit Quantile Networks for Distributional Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1096–1105. PMLR, July 2018a. URL https://proceedings.mlr.press/v80/dabney18a.html.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pp. 2892–2901, New Orleans, Louisiana, USA, February 2018b. AAAI Press. ISBN 978-1-57735-800-8. doi: 10.1609/aaai.v32i1.11791. URL https://dl.acm.org/doi/10.5555/3504035.3504388.

Shripad Deshmukh, Jayakumar Subramanian, Raghavendra Addanki, and Nikos Vlassis. COSAC: Counterfactual Credit Assignment in Sequential Cooperative Teams, 2026. URL https://arxiv.org/abs/2604.17693v2. arXiv:2604.17693v2.

J. Dhaene, M. Denuit, M. J. Goovaerts, R. Kaas, and D. Vyncke. The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31(1):3–33, August 2002. ISSN 0167-6687. doi: 10.1016/S0167-6687(02)00134-8. URL https://www.sciencedirect.com/science/article/pii/S0167668702001348.

Zehao Dou, Jakub Grudzien Kuba, and Yaodong Yang. Understanding Value Decomposition Algorithms in Deep Cooperative Multi-Agent Reinforcement Learning, 2022. URL https://arxiv.org/abs/2202.04868.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pp. 2974–2982, New Orleans, Louisiana, USA, February 2018. AAAI Press. ISBN 978-1-57735-800-8. URL https://dl.acm.org/doi/10.5555/3504035.3504398.

Wei Geng, Baidi Xiao, Rongpeng Li, Ning Wei, Dong Wang, and Zhifeng Zhao. Noise Distribution Decomposition Based Multi-Agent Distributional Reinforcement Learning. *IEEE Transactions on Mobile Computing*, 24(3):2301–2314, March 2025. ISSN 1558-0660. doi: 10.1109/TMC.2024.3492272. URL https://ieeexplore.ieee.org/document/10746312.

Hafez Ghaemi, Shirin Jamshidi, Mohammad Mashreghi, Majid Nili Ahmadabadi, and Hamed Kebriaei. Risk Sensitivity in Markov Games and Multi-Agent Reinforcement Learning: A Systematic Review, June 2024. URL https://arxiv.org/abs/2406.06041.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, April 2007. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2007.00587.x. URL https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Roger Koenker and Gilbert Bassett. Regression Quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 0012-9682. doi: 10.2307/1913643. URL https://www.jstor.org/stable/1913643.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2796–2804. PMLR, July 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Mathias Lechner, Lianhao Yin, Tim Seyde, Tsun-Hsuan Johnson Wang, Wei Xiao, Ramin Hasani, Joshua Rountree, and Daniela Rus. Gigastep - One Billion Steps per Second Multi-agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 155–170, December 2023. doi: 10.52202/075280-0009. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/00ba06ba5c324efdfb068865ca44cf0b-Abstract-Datasets_and_Benchmarks.html.

Shiau Hong Lim and Ilyas Malik. Distributional Reinforcement Learning for Risk-Sensitive Policies. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30977–30989. Curran Associates, Inc., 2022. doi: 10.52202/068431-2246. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c88a2bd0e793550d0e885aa6e31ca277-Abstract-Conference.html.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6382–6393, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL https://dl.acm.org/doi/10.5555/3295222.3295385.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated Model-Based Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4314–4323. PMLR, May 2019. URL https://proceedings.mlr.press/v97/malik19a.html.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer-Verlag, New York, 2003. ISBN 978-0-387-00454-9. doi: 10.1007/b97478. URL http://link.springer.com/10.1007/b97478.

Lorenz Meier, Dominik Honegger, and Marc Pollefeys. PX4: A node-based multithreaded open source robotics framework for deeply embedded platforms. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6235–6240, 2015. doi: 10.1109/ICRA.2015.7140074.

Allan H. Murphy. The Value of Climatological, Categorical and Probabilistic Forecasts in the Cost-Loss Ratio Situation. *Monthly Weather Review*, 105(7):803–816, July 1977. ISSN 1520-0493, 0027-0644. URL https://journals.ametsoc.org/view/journals/mwre/105/7/1520-0493_1977_105_0803_tvocca_2_0_co_2.xml.

Roger B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, New York, NY, second edition, 2006. ISBN 978-0-387-28659-4. doi: 10.1007/0-387-28678-0. URL http://link.springer.com/10.1007/0-387-28678-0.

Wei Qiu, Xinrun Wang, Runsheng Yu, Rundong Wang, Xu He, Bo An, Svetlana Obraztsova, and Zinovi Rabinovich. RMIX: Learning Risk-Sensitive Policies for Cooperative Reinforcement Learning Agents. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23049–23062, 2021. URL https://openreview.net/forum?id=83SeeJals7j.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(178):1–51, January 2020. ISSN 1532-4435. URL https://dl.acm.org/doi/10.5555/3455716.3455894.

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

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553, Red Hook, NY, USA, December 2019. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3454287.3454605.

Thomas J. Rothenberg. Identification in Parametric Models. *Econometrica*, 39(3):577–591, 1971. ISSN 0012-9682. doi: 10.2307/1913267. URL https://www.jstor.org/stable/1913267.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Remi Munos, Marc G. Bellemare, and Will Dabney. Statistics and Samples in Distributional Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5528–5536. PMLR, May 2019. URL https://proceedings.mlr.press/v97/rowland19a.html.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016. URL http://arxiv.org/abs/1506.02438. arXiv:1506.02438 [cs.LG].

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347 [cs.LG].

Siqi Shen, Chennan Ma, Chao Li, Weiquan Liu, Yongquan Fu, Songzhu Mei, Xinwang Liu, and Cheng Wang. RiskQ: Risk-sensitive Multi-Agent Reinforcement Learning Value Factorization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34791–34825. Curran Associates, Inc., 2023. doi: 10.52202/075280-1512. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6d3040941a2d57ead4043556a70dd728-Abstract-Conference.html.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pp. 32033–32058, Honolulu, Hawaii, USA, July 2023. JMLR.org.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896, 2019. URL https://arxiv.org/abs/1905.05408v1.

Kyunghwan Son, Junsu Kim, Sungsoo Ahn, Roben D. Delos Reyes, Yung Yi, and Jinwoo Shin. Disentangling Sources of Risk for Distributional Multi-Agent Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20347–20368. PMLR, June 2022. URL https://proceedings.mlr.press/v162/son22a.html.

Wei-Fang Sun, Cheng-Kuang Lee, and Chun-Yi Lee. DFAC Framework: Factorizing the Value Function via Quantile Mixture for Multi-Agent Distributional Q-Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9945–9954. PMLR, July 2021. URL https://proceedings.mlr.press/v139/sun21c.html.

Wei-Fang Sun, Cheng-Kuang Lee, Simon See, and Chun-Yi Lee. A Unified Framework for Factorizing Distributional Value Functions for Multi-Agent Reinforcement Learning. *Journal of Machine Learning Research*, 24(220), 2023. arXiv:2306.02430.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 2085–2087, Richland, SC, July 2018. International Foundation for Autonomous Agents and Multiagent Systems. URL https://dl.acm.org/doi/10.5555/3237383.3238080.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Hongwei Wang, Lantao Yu, Zhangjie Cao, and Stefano Ermon. Multi-agent Imitation Learning with Copulas. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 139–156, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-86486-6. doi: 10.1007/978-3-030-86486-6_9.

Jianhao Wang, Zhizhou Ren, Beining Han, Jianing Ye, and Chongjie Zhang. Towards Understanding Cooperative Multi-Agent Q-Learning with Value Factorization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29142–29155, 2021b. URL https://arxiv.org/abs/2006.00587.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=Rcmk0xxIQV.

Jianhong Wang, Yunjie Gu, Yuan Zhang, and Tae-Kyun Kim. SHAQ: incorporating shapley value theory into multi-agent Q-Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5941–5954, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of Multivariate Distributional Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 101297–101336, 2024. doi: 10.52202/079017-3212. URL https://openreview.net/forum?id=aq3I5B6GLG.

David H. Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(2–3):265–279, June 2001. ISSN 0219-5259. doi: 10.1142/S0219525901000188. URL https://www.worldscientific.com/doi/abs/10.1142/S0219525901000188.

Fan Yang and Nilanjan Chakraborty. Algorithm for optimal chance constrained linear assignment. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 801–808, May 2017. doi: 10.1109/ICRA.2017.7989099. URL https://ieeexplore.ieee.org/document/7989099.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 35, pp. 24611–24624, November 2022. doi: 10.52202/068431-1787. URL http://arxiv.org/abs/2103.01955. arXiv:2103.01955 [cs.LG].

Kevin Zakka, Yuval Tassa, and MuJoCo Menagerie Contributors. MuJoCo Menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL https://github.com/google-deepmind/mujoco_menagerie.

Pushi Zhang, Xiaoyu Chen, Li Zhao, Wei Xiong, Tao Qin, and Tie-Yan Liu. Distributional Reinforcement Learning for Multi-Dimensional Reward Functions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1519–1529. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/0b9e57c46de934cee33b0e8d1839bfc2-Abstract.html.

Yudi Zhang, Yali Du, Biwei Huang, Mykola Pechenizkiy, and Meng Fang. A Causality-Inspired Spatial-Temporal Return Decomposition Approach for Multi-Agent Reinforcement Learning. *Reinforcement Learning Journal*, 7, 2026. URL https://rlj.cs.umass.edu/2026/papers/Paper83.html.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibration. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22313–22324, 2021. URL http://arxiv.org/abs/2107.05719. arXiv:2107.05719 [stat.ML].

Lifeng Zhou and Pratap Tokekar. Risk-Aware Submodular Optimization for Multi-Robot Coordination. *IEEE Transactions on Robotics*, 38(5):3064–3084, 2022. doi: 10.1109/TRO.2022.3158227. URL http://arxiv.org/abs/2003.10492. arXiv:2003.10492 [cs.RO].

A Extended related work

Table 1 lists the prior results closest to ours, what each finds to be non-unique or decomposed and how this paper differs.

Identification asks whether observations determine a parameter (Rothenberg, 1971) and partial identification describes the set of values consistent with the data (Manski, 2003). In reward learning, Skalse et al. (2023) characterise which reward ambiguities change downstream policies; Proposition 2 is a statement of the same kind for per-agent risk. Quantile calibration is read from reliability curves (Kuleshov et al., 2018), pinball training alone does not achieve it (Chung et al., 2021) and conformal methods correct intervals after training under exchangeability (Romano et al., 2019),

Table 1: Prior work on non-unique or decomposed per-agent quantities.

work	non-unique or decomposed quantity	difference from this paper
QTRAN (Son et al., 2019)	individual utilities, up to per-agent positive affine maps under the individual-global-max condition	treated as harmless for greedy actions; we study per-agent return distributions, where the ambiguity changes risk decisions
Wang et al. (2021b)	individual Q-values under linear factorisation, up to a zero-sum residue	fitted Q-iteration on expected values; we study quantile heads and their calibration
DFAC (Sun et al., 2021; 2023)	team distribution as a level-wise sum of per-agent quantile functions, with an implied dependence structure	we identify the coupling as comonotone, measure its effect and remove it
RiskQ (Shen et al., 2023)	team quantiles as learned non-negative combinations of per-agent quantiles; risk measures add only for comonotone variables	no measurement of per-agent calibration or of decisions that read the per-agent heads
DRIMA (Son et al., 2022)	risk from teammates' actions and environmental risk, the latter modelled jointly	we model per-agent environmental return distributions
Geng et al. (2025)	decompositions of a global reward distribution into individual ones, reduced by penalties	we supervise with observed individual rewards and study value decomposition
CAST (Zhang et al., 2026)	individual rewards, identifiable in a causal return decomposition	concerns rewards; we study per-agent value distributions
COSAC (Deshmukh et al., 2026)	per-agent credit for teams acting in a fixed order with one team reward	reward decomposition in a critic-free policy gradient
Zhang et al. (2021); Wiltzer et al. (2024)	joint distribution of a vector-valued return	single-agent; no decomposition into agents or team head

which serially correlated states do not satisfy. Calibrated uncertainty matters for model-based RL (Malik et al., 2019) and calibration defined relative to decisions (Murphy, 1977; Zhao et al., 2021) motivates the decision experiments of Section 5.3. Risk-sensitive sequential decisions with the conditional value at risk (Chow et al., 2015) and distorted quantiles (Dabney et al., 2018a) use the per-agent tails whose calibration we measure.

B PROOFS

Proof of Proposition 1. Under a matched-sum mixer the team quantiles of $\theta + \Delta$ are $\sum_i(\theta_{i,k} + \Delta_{i,k}) = \sum_i \theta_{i,k} + \sum_i \Delta_{i,k} = \sum_i \theta_{i,k}$ for every k. Under a mean-shape mixer the per-agent means of $\theta + \Delta$ are $\mu_i + K^{-1}\sum_k \Delta_{i,k} = \mu_i$, so f receives the same arguments and the shape term is $\sum_i(\theta_{i,k} + \Delta_{i,k} - \mu_i) = \sum_i(\theta_{i,k} - \mu_i)$. In both cases the team quantiles coincide and the implemented mixers sort them, which preserves equality. The matched-sum condition is K independent linear constraints on nK entries, which leaves an $(n-1)K$-dimensional subspace. The mean-shape conditions add n constraints of which $n-1$ are independent of the first K, because summing $\sum_i \Delta_{i,k} = 0$ over k already gives $\sum_i \sum_k \Delta_{i,k} = 0$; this leaves $nK - K - (n-1) = (n-1)(K-1)$ dimensions. Because the rows of θ are strictly increasing, every Δ whose entries are smaller in absolute value than half the smallest gap between consecutive entries of a row keeps the rows of $\theta + \Delta$ increasing, so the equivalent assignments contain a neighbourhood of θ in each subspace. For the mean-shape mixer the subspace is a subset of the equivalent assignments: moving the means along a level set of f can add further directions. With ties the neighbourhood can be lower-dimensional; for $n = K = 2$ and both rows equal to $(0, 0)$, monotonicity forces $\Delta_{1,1} = \Delta_{1,2}$. □

The three decompositions. For $X = \sigma Z$ with $Z \sim \mathcal{N}(0,1)$ and $\sigma > 0$, the quantile function is $\sigma\Phi^{-1}(\tau)$. Under A, B and C the per-agent quantile functions sum to $(3.8 + 0.2)\Phi^{-1}(\tau) = 4.0\,\Phi^{-1}(\tau)$ at every level, all means are zero and the deviations equal the quantiles, so both mixer classes return identical team quantiles. The conditional value at risk of a zero-mean Gaussian at level α is $-\sigma\,\phi(\Phi^{-1}(\alpha))/\alpha$, which gives the values in Section 3. The grid values average the lowest $\lfloor \alpha K \rfloor$ of the K midpoint quantiles.

Proof of Proposition 2. In both environments the team return is $4.0Z$ for both allocations, so every quantity computed from team-level data has the same distribution in both. A decision rule using team-level data alone therefore selects the first allocation with a probability p that is the same in both environments. The first allocation is feasible in the first environment and infeasible in the second; the second allocation has the reverse status. The probabilities of an infeasible choice are $1 - p$ and p, whose maximum is at least one half. □

Proof of Proposition 3. Conditionally on h, the first statement is Theorem 5 of Dhaene et al. (2002) with $\alpha = 1$, for which their α-inverse is the usual left-continuous quantile function and the convex-order statement is their Theorem 7; both assume finite means. For the Gaussian case, let $G_1, \ldots, G_n$ be jointly Gaussian given h with means m_i, common standard deviation σ and pairwise correlation ρ; the covariance matrix is positive semi-definite exactly when $\rho \geq -1/(n-1)$. The team return has standard deviation $\sigma\sqrt{n(1+(n-1)\rho)}$. Heads of the form $m_i + \sigma'\Phi^{-1}(\tau)$ sum at matched levels to $\sum_i m_i + n\sigma'\Phi^{-1}(\tau)$, which equals the team quantile function exactly when $n\sigma' = \sigma\sqrt{n(1+(n-1)\rho)}$, that is $\sigma' = \sigma\, c(n,\rho)$. The central interval $m_i \pm \sigma' z_{1-\delta/2}$ then contains G_i with probability $\Pr(|Z| \leq c\, z_{1-\delta/2}) = 2\Phi(c\, z_{1-\delta/2}) - 1$. Without the common scale, matching fixes only $\sum_i \sigma'_i$, which is the non-identification of Proposition 1 again. □

Variance of the comonotone sum. With finite variances, $\mathrm{Var}(\sum_i G_i) = \sum_i \mathrm{Var}(G_i) + 2\sum_{i<j} \mathrm{Cov}(G_i, G_j)$ and for fixed marginals each covariance is at most its comonotone value, with equality only when the pair is comonotone (Hoeffding's covariance identity and the upper Fréchet bound). Pairwise comonotone vectors are comonotone (Dhaene et al., 2002, Theorem 3), so $\mathrm{Var}(S^c) \geq \mathrm{Var}(\sum_i G_i)$ with equality only for comonotone returns.

C Algorithm and implementation

Networks. The actor maps an agent's own observation through two tanh layers of width 256 to the mean of a diagonal Gaussian policy whose log standard deviation is a state-independent parameter clipped to $[-2.5, 0]$; its parameters are shared across agents. The per-agent critic head is a separate network of the same shape, two tanh layers and a linear output of K values sorted into quantiles. The team head has the same shape and reads the team observation, the concatenation of all agents' observations (726 dimensions in the search task, where each agent observes 121; 450 in Gigastep, where each of the five seekers observes 90 values, nine for each of the ten agents). The concatenation fixes an agent order. Matched-sum arms compute the team quantiles as the sum of the per-agent means plus the summed deviations from the means, which equals Equation (2). The decoupled supervised critic has 365,376 critic parameters in the search task. The wider matched-sum control has a trunk of width 532 (365,516 parameters) and the team-input control, whose per-agent heads read their own and the team observation, has width 307 (364,748 parameters).

Optimisation. All critics use Adam with $\epsilon = 10^{-5}$, a learning rate decayed linearly to zero over all gradient steps and global gradient-norm clipping (Table 2). Advantages are normalised within each minibatch. The distributional targets of Equation (3) are computed once per update from the rollout, with the bootstrap from the current heads and fixed through the epochs. In critic training the actor parameters are loaded from the frozen policy and their gradients are masked to zero, so the rollouts of every arm come from the same policy. Each run performs 11.8 billion environment steps on one GPU. Table 3 lists the arms and Table 4 the frozen policies.

Table 2: Critic training hyperparameters shared by every arm; the per-agent weight and critic width of each arm are in Table 3.

hyperparameter	value
parallel environments E	2,048
rollout length T	64
updates	90,000
epochs per update	4
minibatches per epoch	4
learning rate (Adam, linear decay to zero)	0.0003
gradient norm clip	0.5
discount γ	0.99
GAE trace λ	0.95
PPO clip range	0.2
team-loss weight c_v	0.5
entropy weight c_H	0.001
quantiles K	32
quantile Huber threshold κ	1
hidden width (actor, heads)	256
minimum policy log standard deviation	-2.5
actor update weight in critic training	0 (policy frozen)

Table 3: Critic arms. Per-agent input "own + team" means each head reads the agent's observation and the team observation.

arm	w	team head	critic width	per-agent input	target
matched sum, team-only	0	matched sum	256	own	distributional
matched sum, $w = 0.5$	0.5	matched sum	256	own	distributional
matched sum, $w = 1$	1	matched sum	256	own	distributional
matched sum, $w = 4$	4	matched sum	256	own	distributional
decoupled supervised critic	4	separate	256	own	distributional
separate heads, team-only	0	separate	256	own	distributional
scalar target, team-only	0	matched sum	256	own	scalar (team only)
scalar target, $w = 4$	4	matched sum	256	own	scalar (team only)
wider matched sum, team-only	0	matched sum	532	own	distributional
wider matched sum, $w = 4$	4	matched sum	532	own	distributional
team-input matched sum, team-only	0	matched sum	307	own + team	distributional
team-input matched sum, $w = 4$	4	matched sum	307	own + team	distributional

Table 4: Frozen policies. Each was trained with the listed critic, with its actor updated.

frozen policy	critic it was trained with	environments	rollout	updates	width	quantiles
Search, policy 1	VDN	2,048	64	90,000	256	32
Search, policy 2	QMIX	640	200	90,000	256	32
Gigastep, policy 1	VDN	2,048	64	30,000	256	32
Gigastep, policy 2	QMIX	2,048	64	30,000	256	32

D ENVIRONMENTS

Figure 6 shows one recorded episode of each environment and Tables 5 and 6 list their parameters. Figure 7 replays the search-task episode in MuJoCo (Todorov et al., 2012), with the Go2 model from MuJoCo Menagerie (Zakka et al., 2022) and the X500 model from the PX4 simulation models (Meier et al., 2015).

Search task. Four quadrotors and two quadrupeds, with platform parameters taken from the Holybro X500 and Unitree Go2 specifications, search an arena for static targets while humans walk through it. Each agent observes its own state, a capability vector that identifies its platform class, up to eight neighbours, the four nearest visible humans and the four nearest visible targets not yet detected. Rewards are individual: every agent within detection range of a target when it is first

detected receives the target reward and each agent pays penalties for approaching humans, for collisions and for the norm of its normalised translational action, plus a small shaping term for closing on the nearest undetected target. The team reward is the sum. Training episodes start from a fresh state with the step counter set to a random value, so the remaining horizon varies; the observation does not include time.

Gigastep hide-and-seek. Five seekers controlled by the policy chase five scripted hiders in Gigastep (Lechner et al., 2023). The seeker types differ in thrust, sensing range and health. Gigastep pays each live seeker a share of the defeat reward when a hider is removed, pays damage to the seeker that deals it, pays a detection reward for hiders in view, charges every live seeker when a teammate is disabled and penalises and disables an agent that leaves the arena. Episodes run to a fixed length whether or not all hiders are caught. Each seeker observes a vector of nine fields for each of the ten agents, with Gigastep's stochastic detection.

Credit split. The share parameter β moves the defeat reward between the pool and the seekers that caused the defeat. For each defeat, a fraction β of the team's defeat reward is paid equally to the live seekers and the rest is divided among the seekers that dealt damage to that hider on that step (Gigastep deals damage once a seeker has tracked a hider for more than three steps). Defeats without an identifiable dealer, such as a hider leaving the arena, stay shared. The team reward at every step is unchanged, so the team return, the policy and the states are the same at every level; $\beta = 1$ is the unmodified environment.

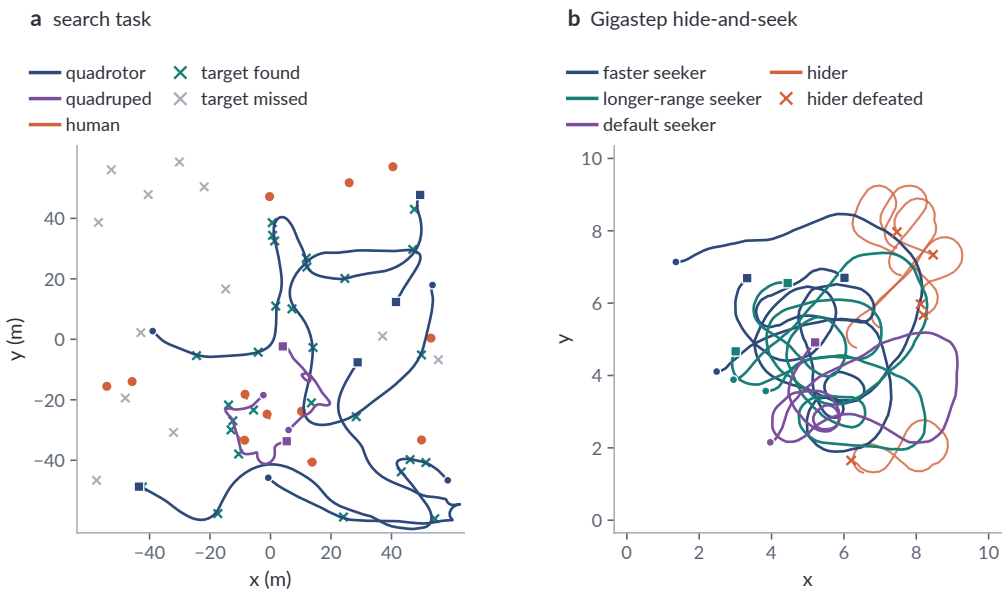


Figure 6: One recorded episode of each environment under policy 1, seen from above. Circles mark the start and squares the end of each robot and seeker trajectory. (a) Search task: dots mark the humans' final positions. (b) Gigastep: crosses mark where hiders were defeated and dots where surviving hiders ended.

E Evaluation protocol

Frozen policies. Search policy 1 is the reference policy. Policy 2 came from a predeclared screen of 37 candidate checkpoints trained with different critics, rollout lengths and trace parameters. A candidate had to find at least 0.8 times the targets found by the reference policy (11.62 of 40 on average) and to change a block of within-state Spearman correlations between agent returns by at least 0.05, with a state-block bootstrap interval (1,000 resamples, 128 states, 512 continuations)

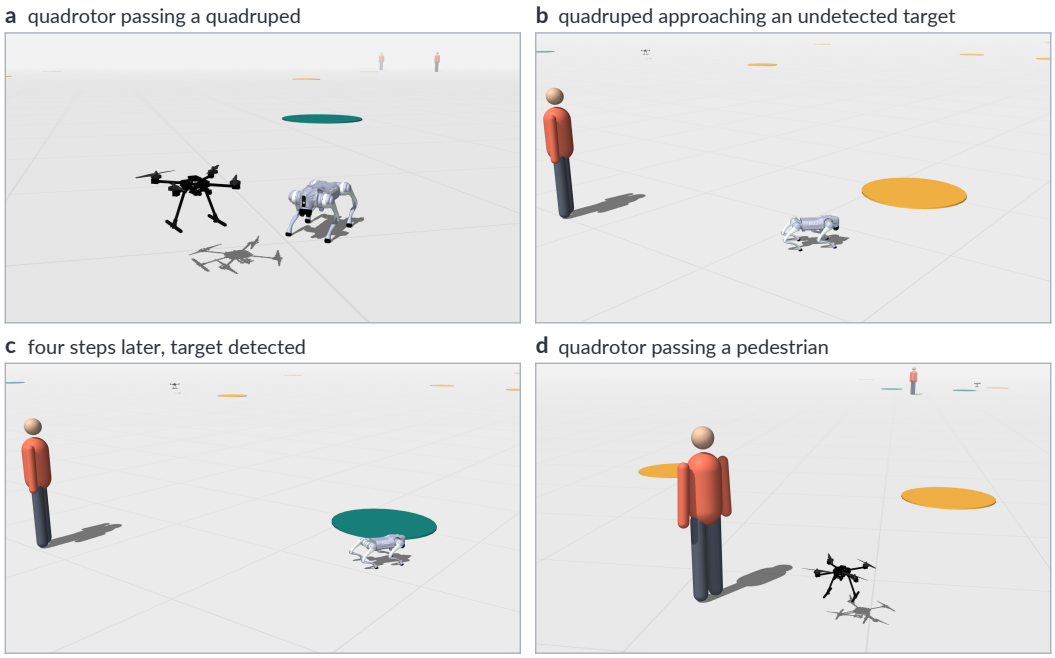


Figure 7: The search-task episode of Figure 6a replayed in MuJoCo for illustration; training and evaluation use the JAX simulator. Every robot base and every human is placed on its recorded track, so the distances shown are those of the recorded episode; the quadruped legs trot under MuJoCo physics. Target discs are amber until detected and teal after. (a) A quadrotor and a quadruped 1.1 m apart horizontally. (b, c) A quadruped approaching a target and, four steps later, the target detected; no other robot was within 20 m of the target. (d) A quadrotor 1.4 m horizontally from a pedestrian.

Table 5: Search task parameters.

property	value
agents	4 quadrotors (X500) and 2 quadrupeds (Go2)
arena	120 m square
static targets	40
moving humans	12 at 1.3 m/s
episode	200 steps of 0.1 s
detection radius	2 m
reward per newly detected target	10, to each detecting agent
human-proximity penalty	up to 5 per step, linear inside 3 m
collision penalty	2 per step per neighbour within 1.5 m
control penalty	0.01 times the norm of the normalised translational action
shaping	0.05 per metre closed on the nearest undetected target
observation mode	ranged

excluding zero. Of 29 competent candidates, none passed. We took the competent candidate with the largest change whose interval excluded zero ($+0.045$ [$+0.029$, $+0.063$] among the aerial agents, 11.17 targets), so policy 2 repeats the experiments under a second policy with nearly the same dependence structure. In Gigastep a policy qualified if it caught at least two of five hiders per episode on average; the VDN-trained policy caught 4.73 and the QMIX-trained policy 3.87.

Calibration. Every critic is scored at its last checkpoint on episodes drawn with an evaluator seed not used in training. In the search task we draw 4,096 episodes the way training does, with a random starting step, keep each with probability proportional to its length, choose a uniform step within each kept episode and compare the critic's prediction on the observation the actor acted on with the

Table 6: Gigastep parameters, with rewards given as the parameters of the Gigastep scenario.

property	value
seekers (controlled)	2 faster, 2 longer-range, 1 default
hiders (scripted)	5, flee the nearest live seeker at full throttle
hider heading noise	uniform in $[-0.3, 0.3]$
hider wall margin	turn to the centre within 1.5 of a wall
arena	10×10, no obstacles
episode	200 steps, fixed length
defeat reward parameter	50
damage reward parameter	10
detection reward parameter	0.5
disabled-teammate penalty parameter	10
arena-exit penalty parameter	10
Gigastep revision	`6288d650`

discounted return from that step to the end. This reproduces the distribution of remaining horizons seen in training. In Gigastep, where training episodes start at step 0, we predict at the first step of 2,048 episodes. The central interval is formed by the grid levels closest to 0.05 and 0.95, which are 0.046875 and 0.953125, so its nominal level is 0.90625. The coverage at each quantile level (Figure 8) is the fraction of realised returns below the predicted quantile. CRPS uses the 32-point approximation of Section 2.

Paired intervals. Arms under one policy share the frozen actor and the evaluator seed, so their evaluation episodes are identical and per-episode scores pair across arms. For a matched-seed difference we resample training seeds with replacement and, independently, the shared episodes with replacement, 2,000 times and report the 2.5 and 97.5 percentiles of the mean difference.

Across-state decision. We draw 1,024 initial states of fresh episodes, with the step counter at zero, and, from each, 512 continuations of the frozen policy with fresh random keys; the critic and the first action of every continuation share one observation per state. Per-agent conditional value at risk at level 0.1 is computed on both sides with the boundary sample or quantile weighted fractionally so that the level is exact. The budget is the median of the oracle's worst-agent conditional value at risk, so half the states are feasible. For each critic, a threshold on its worst-agent value is chosen on one half of the states to match the feasible fraction there and scored on the other half, in both directions. We report held-out error, AUROC, an exact sign-flip permutation test over paired training seeds, a two-way bootstrap over training seeds and states and the fraction of states whose oracle label flips in more than a tenth of 200 continuation resamples.

Within-state ordering. In each state the agents' conditional values at risk are demeaned across agents for the oracle and for the critic. We report the Spearman correlation between the two orderings (average ranks for ties) averaged over states, the rate at which the critic's worst agent is the oracle's worst agent and the regret, the oracle value of the critic's choice minus the oracle minimum. Differences are team-only minus decoupled for regret and decoupled minus team-only otherwise, so positive values favour the decoupled supervised critic.

Remaining horizon. The search-task observation contains no time, so each head predicts a return averaged over remaining horizons. Table 7 shows the resulting bias of the team head's mean by remaining horizon at the evaluation distribution: short horizons are over-predicted and long ones under-predicted, with the same pattern in every arm. Scoring the search task at the first step of full-length episodes, which training rarely visits, is a distribution shift and we do not report it as calibration there. The decision experiments do use initial states. Every decision state then has the same remaining horizon, which removes the variation due to horizon length alone. The ranking of states is unchanged only by a bias that is the same at every state and we did not test whether the remaining bias has that form; the within-state ordering is unchanged by a bias common to the agents of a state and a bias that differs between agent classes, which we have not measured, would affect it.

Table 7: Team head, predicted mean minus realised return by remaining steps, search policy 1, mean over training seeds.

remaining steps	1 to 40 (n 752)	41 to 80 (n 580)	81 to 120 (n 446)	121 to 160 (n 263)	161 to 200 (n 90)
team-only	+36.2	+1.1	−22.8	−35.3	−47.3
$w = 4$	+41.6	+4.0	−23.0	−37.3	−49.9
decoupled	+35.9	+0.7	−23.2	−34.4	−47.0

F FULL RESULTS

Tables 8 and 9 report every arm with per-class coverage and the standard deviation of per-agent coverage across training seeds, Table 10 the paired CRPS differences and Figure 8 the coverage at every quantile level. The oracle recomposition behind Figure 2 is in Table 11.

Table 8: Search task, every arm (coverage at nominal level 0.90625). Per-class columns give per-agent coverage for the aerial and legged agents. In the separate-heads team-only arm no loss term reaches the per-agent heads, which therefore stay at their initialisation; the row shows the output of untrained heads. The scalar-target arms use the scalar λ-return for the team target only. Bold: within each block and among the matched-sum and decoupled arms, the coverage closest to the nominal level and the lowest CRPS.

critic	seeds	team cov.	per-agent cov. (sd)	aerial	legged	team CRPS	agent CRPS
Search, policy 1							
matched sum, team-only	10	0.931	0.579 (0.119)	0.630	0.475	20.29	8.19
matched sum, $w = 0.5$	4	0.947	0.793 (0.002)	0.800	0.779	20.71	4.86
matched sum, $w = 1$	4	0.950	0.814 (0.002)	0.822	0.799	20.98	4.75
matched sum, $w = 4$	10	0.963	0.840 (0.002)	0.851	0.818	21.51	**4.69**
decoupled	10	**0.928**	**0.870 (0.001)**	**0.882**	**0.846**	**20.15**	**4.69**
separate heads, team-only	4	0.927	0.276 (0.001)	0.212	0.405	20.13	12.15
scalar target, team-only	4	0.575	0.319 (0.114)	0.358	0.240	23.64	9.86
scalar target, $w = 4$	4	0.929	0.760 (0.002)	0.776	0.730	21.95	4.93
Search, policy 2							
matched sum, team-only	6	0.926	0.420 (0.051)	0.570	0.120	19.93	11.54
matched sum, $w = 4$	6	0.959	0.843 (0.002)	0.853	0.823	20.94	**4.73**
decoupled	6	**0.925**	**0.874 (0.001)**	**0.886**	**0.850**	**19.74**	**4.73**

Table 9: Gigastep, every arm, including the credit-split levels. Per-class columns give per-agent coverage for the three seeker types. Bold: within each block, the coverage closest to the nominal level and the lowest CRPS.

critic	seeds	team cov.	per-agent cov. (sd)	default	faster	longer-range	team CRPS	agent CRPS
Gigastep, policy 1 (share 1)								
matched sum, team-only	6	**0.963**	**0.886 (0.010)**	**0.900**	0.849	**0.915**	13.00	3.36
matched sum, $w = 4$	6	0.977	0.954 (0.003)	0.954	**0.942**	0.965	13.22	**2.93**
decoupled	6	0.965	0.963 (0.001)	0.964	0.951	0.973	**12.91**	**2.93**
Gigastep, policy 2								
matched sum, team-only	6	0.940	0.690 (0.004)	0.709	0.518	0.853	17.45	6.34
matched sum, $w = 4$	6	0.968	**0.936 (0.004)**	**0.932**	**0.941**	**0.934**	17.92	3.90
decoupled	6	**0.939**	0.947 (0.002)	0.940	0.950	0.946	**17.39**	**3.88**
Gigastep, policy 1, share 0								
matched sum, team-only	6	**0.965**	0.263 (0.010)	0.266	0.251	0.272	12.95	19.66
decoupled	6	0.968	**0.878 (0.005)**	**0.884**	**0.860**	**0.893**	**12.94**	**12.97**
Gigastep, policy 1, share 1/3								
matched sum, team-only	6	**0.966**	0.364 (0.006)	0.374	0.339	0.385	12.98	13.04
decoupled	6	0.968	**0.907 (0.002)**	**0.903**	**0.902**	**0.914**	**12.95**	**9.23**
Gigastep, policy 1, share 2/3								
matched sum, team-only	6	**0.961**	0.621 (0.012)	0.702	0.520	0.683	13.04	7.10
decoupled	6	0.966	**0.921 (0.001)**	**0.917**	**0.913**	**0.929**	**12.98**	**5.70**

Table 10: Paired CRPS differences, decoupled supervised critic minus the named arm (negative favours the decoupled supervised critic), with pointwise two-way bootstrap 95% intervals that resample matched training seeds and, independently, the shared episodes.

decoupled minus	seeds	team CRPS [95%]	per-agent CRPS [95%]
Search 1, $w = 4$	10	−1.36 [−1.62, −1.10]	+0.004 [−0.004, +0.013]
Search 1, $w = 0$	10	−0.14 [−0.35, +0.07]	−3.498 [−5.030, −2.126]
Search 1, $w = 1$	4	−0.84 [−1.10, −0.58]	−0.056 [−0.072, −0.039]
Search 1, $w = 0.5$	4	−0.56 [−0.81, −0.32]	−0.165 [−0.185, −0.145]
Search 2, $w = 4$	6	−1.20 [−1.52, −0.91]	+0.000 [−0.010, +0.011]
Search 2, $w = 0$	6	−0.19 [−0.45, +0.07]	−6.809 [−8.004, −5.688]
Gigastep 1, $w = 4$	6	−0.31 [−0.43, −0.19]	−0.001 [−0.007, +0.007]
Gigastep 1, $w = 0$	6	−0.09 [−0.20, +0.03]	−0.423 [−0.482, −0.379]
Gigastep 2, $w = 4$	6	−0.53 [−0.70, −0.36]	−0.013 [−0.024, −0.002]
Gigastep 2, $w = 0$	6	−0.06 [−0.21, +0.09]	−2.458 [−2.609, −2.306]
Search 1, team-input matched sum, $w = 4$	6	+0.16 [−0.01, +0.33]	+0.289 [+0.244, +0.335]
Search 1, team-input matched sum, $w = 0$	6	+0.46 [+0.30, +0.62]	−2.589 [−4.161, −1.619]

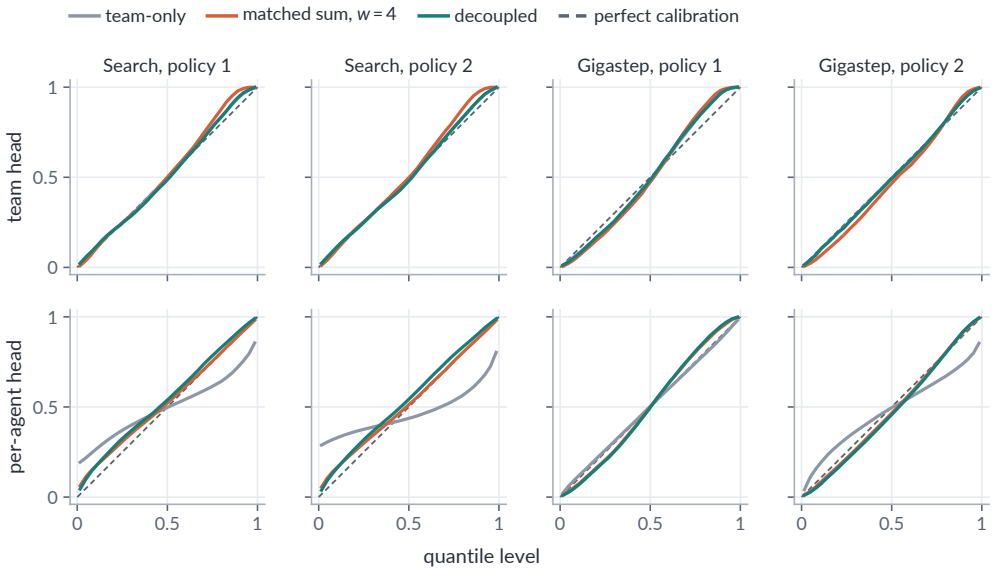


Figure 8: Reliability of the team head (top) and the per-agent heads (bottom): fraction of realised returns below the predicted quantile at each of the 32 levels, mean over training seeds. The team-only critic's per-agent heads are too narrow in the search task and in Gigastep policy 2.

G CONTROLS

Capacity and input. Table 12 compares, under search policy 1 and seeds 4 to 9, the matched sum at its default width, at a width that matches the decoupled supervised critic's parameter count and with the team observation added to every per-agent head. Neither change repaired team-only per-agent coverage; with per-agent supervision, the team input lowered the proper scores and left coverage calibration unchanged. The decision results for the two team-only controls are in Table 17.

Same-seed rerun. We reran three team-only critics of search policy 1 with the same seeds on the same GPU model one day later (Table 13). Per seed, team coverage changed by at most 0.003 and team CRPS by at most 0.03, while per-agent coverage changed by 0.09 to 0.29, which is what Propo-

Table 11: Oracle recomposition, 128 initial states and 2,048 continuations per state at each state seed, with state-block bootstrap 95% intervals for the width ratio and the within-state Spearman correlation.

		coverage			width			
setting	state seed	per-agent	team	matched sum	team	matched sum	ratio [95%]	Spearman [95%]
Search 1	7	0.904	0.905	0.995	33.1	74.7	2.26 [2.12, 2.41]	−0.039 [−0.047, −0.032]
Search 1	8	0.904	0.907	0.996	33.8	74.9	2.22 [2.10, 2.35]	−0.040 [−0.047, −0.032]
Search 2	7	0.905	0.904	0.992	37.3	77.6	2.08 [1.97, 2.21]	−0.031 [−0.039, −0.023]
Search 2	8	0.904	0.905	0.991	44.1	84.9	1.93 [1.83, 2.03]	−0.028 [−0.037, −0.019]
Gigastep 1, share 1	7	0.905	0.905	0.948	59.1	65.8	1.11 [1.10, 1.12]	+0.757 [+0.737, +0.776]
Gigastep 1, share 1	8	0.906	0.906	0.949	58.6	65.5	1.12 [1.11, 1.13]	+0.749 [+0.729, +0.769]
Gigastep 2	7	0.905	0.905	0.937	80.1	85.2	1.06 [1.05, 1.08]	+0.822 [+0.800, +0.841]
Gigastep 2	8	0.905	0.904	0.939	84.7	91.0	1.07 [1.06, 1.09]	+0.837 [+0.819, +0.855]
Gigastep 1, share 0	7	0.906	0.905	1.000	59.1	303.4	5.13 [4.86, 5.45]	−0.126 [−0.136, −0.115]
Gigastep 1, share 0	8	0.906	0.906	1.000	58.6	300.6	5.13 [4.82, 5.45]	−0.132 [−0.142, −0.120]
Gigastep 1, share 1/3	7	0.904	0.905	0.998	59.1	216.7	3.67 [3.48, 3.88]	−0.009 [−0.021, +0.002]
Gigastep 1, share 1/3	8	0.904	0.906	0.998	58.7	215.0	3.66 [3.46, 3.87]	−0.016 [−0.029, −0.003]
Gigastep 1, share 2/3	7	0.905	0.905	0.991	59.0	134.2	2.27 [2.18, 2.38]	+0.115 [+0.097, +0.133]
Gigastep 1, share 2/3	8	0.904	0.906	0.991	58.6	133.0	2.27 [2.16, 2.38]	+0.104 [+0.085, +0.124]

Table 12: Capacity and input controls, search policy 1, seeds 4 to 9.

critic (search policy 1, seeds 4 to 9)	seeds	team cov.	per-agent cov. (sd)	aerial	legged	team CRPS	agent CRPS
matched sum, team-only, width 256	6	0.931	0.562 (0.102)	0.649	0.387	20.29	8.50
matched sum, team-only, width 532	6	0.932	0.614 (0.106)	0.659	0.526	20.27	7.52
matched sum, team-only, own and team input	6	0.931	0.598 (0.083)	0.681	0.431	19.70	7.28
matched sum, $w = 4$, width 256	6	0.963	0.840 (0.001)	0.851	0.818	21.51	4.69
matched sum, $w = 4$, width 532	6	0.961	0.838 (0.001)	0.850	0.816	21.47	4.67
matched sum, $w = 4$, own and team input	6	0.962	0.833 (0.002)	0.844	0.813	20.00	4.40
decoupled, width 256	6	0.929	0.870 (0.001)	0.882	0.845	20.16	4.69

sition 1 allows: the team loss constrains only what the mixer returns. Run-to-run nondeterminism and a source change between the two runs are not separable here.

Table 13: Same-seed reruns of the team-only critic, search policy 1.

	team coverage		team CRPS		per-agent coverage	
seed	original	rerun	original	rerun	original	rerun
4	0.933	0.936	20.28	20.29	0.393	0.682
5	0.931	0.929	20.32	20.31	0.599	0.688
6	0.932	0.929	20.29	20.32	0.521	0.714

Replication on one GPU model. The search-task critics of both policies were retrained from one commit on one GPU model, with the same frozen policies, seeds and arguments and scored with the same protocol (Table 14). Every mean coverage of the supervised critics moved by at most 0.002 and every mean CRPS by at most 0.03; the team-only critic's team coverage and team CRPS moved by at most 0.001 and 0.01, while its per-agent coverage moved from 0.579 to 0.655 under policy 1 and from 0.420 to 0.502 under policy 2 and its per-agent CRPS from 8.19 to 6.76 and from 11.54 to 9.38. The across-state decision advantage of the decoupled supervised critic over the team-only critic was $+0.031$ to $+0.047$ over the eight pairs of policy and state seed, every interval excluding zero, the narrowest being $[+0.002, +0.061]$ with exact $p = 0.062$; the within-state Spearman difference was $+0.34$ to $+0.37$, with the team-only critic's own correlation between -0.10 and -0.04. The main tables keep the originally analysed runs, against which every earlier analysis and review was checked. The Gigastep set of the same replication did not complete within the monthly GPU allocation.

H CREDIT-SPLIT SWEEP

The sweep was specified before any of its results. Under Gigastep policy 1, with the dynamics and the per-step team reward fixed, the defeat-reward share took the values 0, 1/3, 2/3 and 1, with six training seeds per level for the team-only critic and the decoupled supervised critic. The primary

Table 14: Replication of the search-task critics from one commit on one GPU model: means over the same training seeds, frozen runs against the replication. Lower block: team-only minus decoupled held-out error of the across-state decision at 1,024 states, with two-way bootstrap 95% intervals.

		team coverage		per-agent coverage		team CRPS		per-agent CRPS	
critic	seeds	frozen	replication	frozen	replication	frozen	replication	frozen	replication
search policy 1									
team-only	10	0.931	0.931	0.579	0.655	20.29	20.30	8.19	6.76
matched sum, $w = 4$	10	0.963	0.962	0.840	0.840	21.51	21.51	4.69	4.69
decoupled	10	0.928	0.928	0.870	0.869	20.15	20.12	4.69	4.69
search policy 2									
team-only	6	0.926	0.927	0.420	0.502	19.93	19.94	11.54	9.38
matched sum, $w = 4$	6	0.959	0.958	0.843	0.843	20.94	20.94	4.73	4.73
decoupled	6	0.925	0.926	0.874	0.876	19.74	19.76	4.73	4.74

		across-state decision: team-only minus decoupled	
setting	state seed	frozen	replication
Search, policy 1	11	+0.058 [+0.026, +0.088]	+0.045 [+0.013, +0.076]
	12	+0.057 [+0.026, +0.088]	+0.043 [+0.014, +0.073]
	13	+0.051 [+0.020, +0.084]	+0.047 [+0.011, +0.080]
	14	+0.056 [+0.023, +0.092]	+0.036 [+0.002, +0.072]
Search, policy 2	11	+0.060 [+0.027, +0.090]	+0.045 [+0.017, +0.074]
	12	+0.038 [+0.007, +0.068]	+0.031 [+0.002, +0.061]
	13	+0.043 [+0.013, +0.072]	+0.035 [+0.009, +0.062]
	14	+0.047 [+0.020, +0.077]	+0.042 [+0.011, +0.073]

quantities were the oracle's within-state Spearman correlation and width ratio, the team-only critic's per-agent coverage error and the difference in across-state held-out error; the claim was an ordered trend across levels and the contrast between share 0 and share 1. The four predictions were that, as the share falls, the Spearman correlation falls, the width ratio rises, the team-only per-agent coverage error rises and the decision advantage of the decoupled supervised critic rises. The first three predictions were met as ordered trends and at the endpoint contrast (Table 15). The fourth did not: the across-state differences stayed small at every level (Table 17). The within-state ordering, measured afterwards, favoured the decoupled supervised critic at shares 0, 2/3 and 1 (Table 18).

Table 15: Credit-split sweep under Gigastep policy 1. Oracle columns average state seeds 7 and 8; the other columns average six training seeds.

	oracle			per-agent coverage (sd)		team coverage	team CRPS
share	Spearman	width ratio	matched sum cov.	team-only	decoupled	team-only / decoupled	team-only / decoupled
0	−0.13	5.13	1.000	0.263 (0.010)	0.878 (0.005)	0.965 / 0.968	12.95 / 12.94
1/3	−0.01	3.66	0.998	0.364 (0.006)	0.907 (0.002)	0.966 / 0.968	12.98 / 12.95
2/3	+0.11	2.27	0.991	0.621 (0.012)	0.921 (0.001)	0.961 / 0.966	13.04 / 12.98
1	+0.75	1.12	0.948	0.886 (0.010)	0.963 (0.001)	0.963 / 0.965	13.00 / 12.91

I DECISION EXPERIMENTS

Table 16 lists the values in Figure 5. Table 17 gives every across-state comparison at 1,024 states, including the credit-split levels, the capacity and input controls and the matched sum at $w = 4$ against the decoupled supervised critic; Table 18 gives the within-state ordering with the worst-agent hit rate and regret. In each row the first value belongs to the first arm named in the group heading.

J SCALAR AGAINST DISTRIBUTIONAL TARGETS

The distributional target of Equation (3) is a precondition for all of the above. In the two scalar-target arms the team head's quantiles all regress on the scalar λ-return used for the advantage, while the

Table 16: Values behind Figure 5: team-only critic against the decoupled supervised critic. Across states, held-out error at every plotted state seed and the paired difference (team-only minus decoupled). Within a state, the mean within-state Spearman correlation with the oracle's ordering of agents at state seed 12 and the paired difference (decoupled minus team-only). Brackets give pointwise two-way bootstrap 95% intervals.

across states: held-out error

setting	state seed	team-only	decoupled	difference
Search, policy 1	13	0.470	0.419	+0.051 [+0.020, +0.084]
	14	0.449	0.393	+0.056 [+0.023, +0.092]
Search, policy 2	13	0.475	0.432	+0.043 [+0.013, +0.072]
	14	0.488	0.441	+0.047 [+0.020, +0.077]
Gigastep, policy 1	11	0.344	0.327	+0.017 [−0.008, +0.042]
	12	0.316	0.309	+0.008 [−0.018, +0.032]
	13	0.337	0.313	+0.024 [+0.001, +0.047]
	14	0.331	0.325	+0.006 [−0.022, +0.035]
Gigastep, policy 2	11	0.336	0.344	−0.008 [−0.031, +0.017]
	12	0.335	0.339	−0.004 [−0.033, +0.022]
	13	0.325	0.315	+0.009 [−0.012, +0.031]
	14	0.304	0.328	−0.024 [−0.054, +0.005]

within a state, state seed 12: Spearman correlation

setting	team-only	decoupled	difference
Search, policy 1	−0.11	+0.30	+0.41 [+0.22, +0.57]
Search, policy 2	−0.10	+0.27	+0.37 [+0.34, +0.41]
Gigastep, policy 1	+0.25	+0.43	+0.18 [+0.15, +0.22]
Gigastep, policy 2	+0.48	+0.56	+0.08 [+0.05, +0.10]

per-agent targets keep the distributional recursion. Without per-agent supervision this critic undercovers on both heads; with it, team coverage reaches nominal while per-agent coverage stays below that of the distributional target (the two scalar rows of Table 8). The comparisons in the main text therefore keep the distributional target and vary only supervision and recomposition.

K Reproducibility and provenance

Numbers. Every result in the text, tables and figures is generated by scripts from result files: macros for the text, table files for the appendix and plotting scripts for the figures. Design constants in the tables come from the run configurations.

Runs. Every critic run used one GPU. Under search policy 1, the team-only and decoupled cells ran seeds 0 to 3 on NVIDIA B200 and seeds 4 to 9 on NVIDIA H200 and the matched sum at $w = 4$ ran all ten seeds on B200; the $w = 0.5$, $w = 1$ and separate-heads team-only cells ran seeds 0 to 3 on B200; the capacity, input and rerun controls and the scalar-target cells ran on H200, except one team-input seed on B200; search policy 2 and every Gigastep cell ran on B200. Comparisons of the team-only and decoupled cells are therefore matched in GPU model within seed and comparisons of either with the $w = 4$ cell are not for seeds 4 to 9. The runs span more than one revision of the training code between the first and the last run. Training uses JAX and Gigastep is used at the revision in Table 6.

Table 17: Across-state decisions at 1,024 states and 512 continuations per state, at each state seed with n matched training seeds. Held-out error and AUROC are given for the first and second arm of each group; the difference is the first minus the second, with its exact paired permutation p and two-way bootstrap interval. With six seed pairs the smallest attainable p is $2/64$, reached whenever every seed difference has the same sign; the permutation test ignores state sampling, which the interval includes, so the two can disagree. The last column is the fraction of states whose oracle label is unstable under continuation resampling.

setting	seed	n	error	AUROC	difference	exact p	interval	unstable
team-only against decoupled								
Search 1	13	10	0.470 / 0.419	0.543 / 0.624	+0.051	0.006	[+0.020, +0.084]	0.066
Search 1	14	10	0.449 / 0.393	0.573 / 0.662	+0.056	0.008	[+0.023, +0.092]	0.047
Search 2	13	6	0.475 / 0.432	0.542 / 0.604	+0.043	0.031	[+0.013, +0.072]	0.068
Search 2	14	6	0.488 / 0.441	0.525 / 0.594	+0.047	0.031	[+0.020, +0.077]	0.066
Gigastep 1	11	6	0.344 / 0.327	0.733 / 0.761	+0.017	0.031	[−0.008, +0.042]	0.122
Gigastep 1	12	6	0.316 / 0.309	0.760 / 0.771	+0.008	0.031	[−0.018, +0.032]	0.118
Gigastep 1	13	6	0.337 / 0.313	0.743 / 0.768	+0.024	0.031	[+0.001, +0.047]	0.129
Gigastep 1	14	6	0.331 / 0.325	0.742 / 0.754	+0.006	0.500	[−0.022, +0.035]	0.103
Gigastep 2	11	6	0.336 / 0.344	0.744 / 0.733	−0.008	0.281	[−0.031, +0.017]	0.104
Gigastep 2	12	6	0.335 / 0.339	0.747 / 0.739	−0.004	0.625	[−0.033, +0.022]	0.117
Gigastep 2	13	6	0.325 / 0.315	0.755 / 0.766	+0.009	0.062	[−0.012, +0.031]	0.115
Gigastep 2	14	6	0.304 / 0.328	0.776 / 0.753	−0.024	0.125	[−0.054, +0.005]	0.113
Gigastep 1, share 0	11	6	0.361 / 0.346	0.721 / 0.715	+0.015	0.031	[−0.011, +0.043]	0.177
Gigastep 1, share 0	12	6	0.358 / 0.341	0.720 / 0.722	+0.016	0.094	[−0.011, +0.044]	0.204
Gigastep 1, share 0	13	6	0.358 / 0.341	0.719 / 0.720	+0.017	0.094	[−0.011, +0.045]	0.213
Gigastep 1, share 0	14	6	0.361 / 0.360	0.714 / 0.701	+0.001	0.844	[−0.028, +0.029]	0.210
Gigastep 1, share 1/3	11	6	0.352 / 0.336	0.728 / 0.739	+0.016	0.031	[−0.010, +0.041]	0.147
Gigastep 1, share 1/3	12	6	0.325 / 0.321	0.753 / 0.755	+0.003	0.625	[−0.020, +0.031]	0.146
Gigastep 1, share 2/3	11	6	0.343 / 0.326	0.739 / 0.755	+0.017	0.062	[−0.010, +0.044]	0.140
Gigastep 1, share 2/3	12	6	0.315 / 0.320	0.765 / 0.764	−0.005	0.312	[−0.032, +0.021]	0.125
controls against decoupled, search policy 1								
wider team-only	11	6	0.462 / 0.405	0.557 / 0.638	+0.058	0.031	[+0.019, +0.096]	0.064
wider team-only	12	6	0.452 / 0.402	0.573 / 0.659	+0.050	0.031	[+0.016, +0.085]	0.057
team-input team-only	11	6	0.478 / 0.405	0.532 / 0.638	+0.074	0.031	[+0.038, +0.106]	0.064
team-input team-only	12	6	0.484 / 0.402	0.531 / 0.659	+0.082	0.031	[+0.049, +0.116]	0.057
matched sum at $w = 4$ against decoupled								
Search 1	11	6	0.400 / 0.405	0.642 / 0.638	−0.005	0.219	[−0.023, +0.013]	0.064
Search 1	12	6	0.384 / 0.402	0.661 / 0.659	−0.018	0.031	[−0.036, −0.000]	0.057
Search 2	11	6	0.429 / 0.423	0.608 / 0.605	+0.007	0.375	[−0.012, +0.025]	0.059
Search 2	12	6	0.439 / 0.432	0.599 / 0.601	+0.008	0.125	[−0.011, +0.025]	0.065
Gigastep 1	11	6	0.316 / 0.324	0.768 / 0.765	−0.008	0.094	[−0.023, +0.007]	0.117
Gigastep 1	12	6	0.308 / 0.308	0.772 / 0.772	+0.000	1.000	[−0.018, +0.017]	0.113

Table 18: Within-state ordering of agents at 1,024 states, at each state seed with n matched training seeds. Upper block: Spearman correlation and worst-agent hit rate for the first and second arm of each group, with the chance hit rate. Lower block: paired differences with two-way bootstrap 95% intervals, positive favouring the second arm.

setting	seed	n	Spearman	hit rate	chance
team-only against decoupled					
Search 1	12	10	−0.11 / +0.30	0.20 / 0.40	0.17
Search 2	12	6	−0.10 / +0.27	0.17 / 0.33	0.17
Gigastep 1	12	6	+0.25 / +0.43	0.31 / 0.42	0.20
Gigastep 2	12	6	+0.48 / +0.56	0.42 / 0.46	0.20
Gigastep 1, share 0	12	6	+0.34 / +0.57	0.38 / 0.51	0.20
Search 1	11	6	−0.25 / +0.30	0.12 / 0.39	0.17
Gigastep 1	11	6	+0.27 / +0.44	0.32 / 0.43	0.20
Gigastep 1, share 0	11	6	+0.36 / +0.57	0.37 / 0.51	0.20
Gigastep 1, share 2/3	11	6	+0.33 / +0.47	0.33 / 0.43	0.20
matched sum at $w = 4$ against decoupled					
Search 1	11	6	+0.32 / +0.30	0.39 / 0.39	0.17
Search 2	11	6	+0.26 / +0.25	0.32 / 0.30	0.17
Gigastep 1	11	6	+0.42 / +0.44	0.41 / 0.42	0.20

setting	seed	Spearman difference	hit-rate difference	regret difference
team-only against decoupled				
Search 1	12	+0.41 [+0.22, +0.57]	+0.20 [+0.11, +0.28]	+5.88 [+3.22, +8.17]
Search 2	12	+0.37 [+0.34, +0.41]	+0.16 [+0.13, +0.18]	+4.34 [+3.37, +5.30]
Gigastep 1	12	+0.18 [+0.15, +0.22]	+0.11 [+0.06, +0.15]	+1.06 [+0.74, +1.35]
Gigastep 2	12	+0.08 [+0.05, +0.10]	+0.03 [+0.01, +0.06]	+0.26 [+0.11, +0.44]
Gigastep 1, share 0	12	+0.23 [+0.19, +0.27]	+0.13 [+0.08, +0.18]	+0.48 [+0.24, +0.73]
Search 1	11	+0.55 [+0.47, +0.61]	+0.26 [+0.23, +0.30]	+7.15 [+6.30, +7.99]
Gigastep 1	11	+0.17 [+0.13, +0.21]	+0.11 [+0.06, +0.16]	+1.17 [+0.90, +1.47]
Gigastep 1, share 0	11	+0.21 [+0.17, +0.25]	+0.14 [+0.09, +0.19]	+0.56 [+0.32, +0.79]
Gigastep 1, share 2/3	11	+0.14 [+0.11, +0.17]	+0.09 [+0.06, +0.13]	+0.75 [+0.55, +0.96]
matched sum at $w = 4$ against decoupled				
Search 1	11	−0.02 [−0.04, −0.00]	−0.01 [−0.03, +0.02]	−0.29 [−0.80, +0.28]
Search 2	11	−0.01 [−0.03, +0.01]	−0.01 [−0.03, +0.01]	−0.14 [−0.60, +0.36]
Gigastep 1	11	+0.02 [+0.00, +0.04]	+0.02 [−0.01, +0.04]	+0.11 [−0.03, +0.24]