DIFFERENT BONUSES, SIMILAR OUTCOMES: UNDERSTANDING AUXILIARY REWARD NON-SEPARATION IN PPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Auxiliary-reward methods are usually distinguished by the quantity that generates their bonus, yet the bonus itself is not the object optimized by a policy-gradient method. We study a general learning question: how can mathematically different auxiliary objectives produce nearly indistinguishable behaviour after integration into the same policy-optimization pipeline? Six PPO-based systems are compared in a controlled dynamic-survival task: PPO alone, intrinsic curiosity (ICM), random-network distillation (RND), hashed pseudocount exploration, advisory planning, and ICM with advisory planning. Across 40 matched seeds, 240 policies are trained in 120,000 episodes and evaluated in 336,000 held-out episodes. Raw intrinsic signals remain clearly distinguishable; their mean per-step magnitudes span more than three orders of magnitude, while stochastic test survival occupies only a 32.63% to 33.25% range. The hybrid is numerically first, but its paired effects are only $+0.30$ to $+0.61$ percentage points; every Holm-adjusted survival test has $p = 1.000$, and every confidence interval excludes the prespecified five-point benefit. The ICM-by-planner interaction is also unsupported. We formalize an effective-update view in which generalized advantage estimation, episodewise standardization, PPO clipping, and mapping through policy score vectors form a many-to-one map from raw auxiliary rewards to actor updates. The experiment does not establish equal gradients or general algorithmic equivalence. It does establish a controlled case in which semantic and numerical differences among auxiliary signals fail to yield practically distinguishable downstream outcomes.

1 INTRODUCTION

$$\text{ICM} \neq \text{RND} \neq \text{Pseudo-count} \neq \text{Planning}.$$

These are not esthetic changes to a single bonus. Different definitions of useful experience are encoded in them. Pseudocount methods reward low estimated visitation (Bellemare et al., 2016; Tang et al., 2017); ICM rewards error in predicting an action-conditioned transition (Pathak et al., 2017); random-network distillation (RND) rewards prediction error against a fixed random representation (Burda et al., 2019); and a planner can reward agreement with a structured preference over actions. It is evident how they differ mathematically at the signal generation stage. A different empirical question is whether that mathematical distinction remains behaviorally visible after the mechanisms are incorporated into the same policy-gradient pipeline.

This observation defines the paper's central question:

> *Does that mathematical distinction necessarily produce different behavior after all of the signals are passed through the same PPO machinery?*

We test this question under a tightly matched experiment and find essentially no practically meaningful separation. Across held-out stochastic evaluation, six PPO-based systems occupy only a 32.63% to 33.25% survival range, even though the recorded intrinsic signals differ by more than three orders of magnitude. The numerically leading ICM-planner hybrid improves on individual comparators by only $+0.30$ to $+0.61$ percentage points; all five confidence intervals include zero, their upper

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.

endpoints are at most 1.80 points, and every Holm-adjusted survival test has $p = 1.000$. The result is therefore non-separation, not a ranking.

Why might objective-level distinctions fail to survive? An auxiliary reward is added to environmental reward, propagated through temporal credit assignment and a learned critic, standardized within each episode, and filtered by PPO's clipped objective (Schulman et al., 2017; 2016). Only the component that remains correlated with policy score vectors can change the actor. Raw reward functions that look unrelated can consequently induce similar effective sample weights, whereas rewards with similar scales can induce different updates. The relevant object is not the standalone bonus but the update that survives the complete learning pipeline.

We investigate this conundrum in a dynamic pure-survival task where a bad decision can terminate the episode as well as any further data collection. The same observations, actor-critic, action mask, training maps, optimization settings, and matched random streams are shared by six systems. The only distinction between them is whether their learning reward includes ICM, RND, hashed pseudocount novelty, one-step planner agreement, or ICM with planner agreement. Auxiliary reward is not included in any reported results, and the planner never replaces the action sampled by PPO. Additionally, a factorial contrast shows no evidence of a positive interaction between the planner and the ICM.

Three contributions are made.

- **A controlled empirical result.** A matched-seed comparison of 456,000 training and evaluation episodes and 240 policies reveals that while their configured systems are unable to generate practically distinguishable held-out outcomes, recorded intrinsic signals can still be numerically distinguishable.
- **An effective-update formulation.** Temporal propagation, normalization, clipping, and score-weighted mapping are the steps that make up the path from auxiliary reward to actor update. Without asserting that the experiment's unrecorded gradients were equal, this formulation pinpoints specific ways in which different rewards can be compressed.
- **A rigorous assessment protocol for evaluating negative comparisons.** The purpose of numerical rank is to separate it from confirmatory evidence by employing matched seeds, disjoint maps, family-wise error control, a predetermined practical threshold, an interaction contrast, artifact-completeness checks, and explicit limits on the claim.

This separation is intended to ensure a narrow conclusion. It does not claim general equivalence between planning, PPO, pseudocounts, RND, and ICM. However, conceptual differences between their auxiliary objectives did not survive strongly enough to separate downstream survival under one environment family, coefficient setting, architecture, and budget.

2 Related Work

Intrinsic motivation and novelty. ICM uses error in an action-conditioned feature-dynamics model (Pathak et al., 2017). RND trains a predictor to match a fixed random network and uses residual error as a state-novelty signal (Burda et al., 2019). Count and pseudocount methods use visitation density to approximate familiarity (Bellemare et al., 2016; Ostrovski et al., 2017; Tang et al., 2017). They differ in terms of representation dependence, stationarity, semantics, and vulnerability to stochastic observations. Previous research has also demonstrated that distinguishing between episodic and lifelong novelty can be important and that prediction error can reward uncontrollable novelty (Raileanu & Rocktäschel, 2020; Sekar et al., 2020). The current study investigates whether configured instances are still behaviorally separable following a common PPO transformation rather than attempting to find a universal ordering.

Planning and shaping. Options, subgoals, demonstrations, imagined rollouts, and shaped values have all been used to integrate planning with model-free learning. (Bacon et al., 2017; Levy et al., 2018; Ha & Schmidhuber, 2018; Schrittwieser et al., 2020). The experimental object is altered by the planner's power: reward advice assesses whether the policy learns from the planner's preference, while replacing a policy action assesses a composite controller. Here, only the former is employed.

The examined bonuses are not presumed to maintain an ideal policy, in contrast to potential-based shaping; instead, their impact is quantified empirically.

Safe exploration and controlled evaluation. Constrained optimization, shielding, and benchmarked hazard costs have all been used to illustrate safety (Altman, 1999; Achiam et al., 2017; Alshiekh et al., 2018; Ray et al., 2019). Our common action mask is more constrained: it does not certify the subsequent state following stochastic hazard motion, but it does exclude currently observed lethal moves. This residual uncertainty eliminates blatantly incorrect actions for each method while making survival sensitive to learning. In contrast to leaderboard-style evaluation, a small numerical rank is distinguished from evidence that an auxiliary mechanism modifies behavior using a predetermined practical effect and paired uncertainty.

Auxiliary objectives and optimization-mediated effects. Previous studies have demonstrated that deep-RL behavior is significantly influenced by implementation and optimization choices, rather than solely relying on nominal algorithm identity. (Engstrom et al., 2020; Andrychowicz et al., 2021). For intrinsic exploration, Creus Castanyer et al. (2024) also demonstrate that optimization properties, including reward non-stationarity, can have a material impact on the learning of intrinsic objectives. Through shared representations and gradient directions, auxiliary-task and multi-task work also shows how other objectives interact with the main objective (Jaderberg et al., 2017; Du et al., 2018; Yu et al., 2020; Cobbe et al., 2021). These studies demonstrate that downstream learning effects cannot be determined solely by objective semantics. Our inquiry is more focused and comes after this body of work: can heterogeneous auxiliary rewards' objective-level differences still be discernible when they go through the same return estimator, critic, normalization, clipping rule, and policy-score mapping?

3 From Auxiliary Rewards to Effective PPO Updates

3.1 The shared transformation

Let method m use the learning reward

$$\tilde{r}_t^{(m)} = r_t^{\text{ext}} + \beta_t b_t^{(m)} + p_t^{(m)}, \tag{1}$$

where $b_t^{(m)}$ is an intrinsic bonus, β_t its coefficient, and $p_t^{(m)}$ an optional planner-agreement reward. At the conceptual level, every method passes through the same sequence

$$\tilde{r}^{(m)} \longrightarrow \hat{A}_t^{(m)} \longrightarrow \bar{A}_t^{(m)} \longrightarrow L_m^{\text{clip}}. \tag{2}$$

PPO does not optimize "curiosity," "novelty," or "planner agreement" as semantic concepts. It optimizes the temporal and action-discriminative advantage information that remains after return construction, generalized advantage estimation (GAE), normalization, and clipping. This is the conceptual lens used throughout the experiment.

For compactness, define $u_t^{(m)} = \beta_t b_t^{(m)} + p_t^{(m)}$. For a trajectory of length T, the method-specific generalized advantage estimate is

$$\hat{A}_t^{(m)} = \sum_{\ell=0}^{T-t-1} (\gamma\lambda)^\ell \delta_{t+\ell}^{(m)}, \tag{3}$$

$$\delta_t^{(m)} = r_t^{\text{ext}} + u_t^{(m)} + \gamma V_{\phi_m}(s_{t+1}) - V_{\phi_m}(s_t). \tag{4}$$

The critic is method-specific after learning; therefore, even the decomposition into extrinsic and auxiliary advantage is coupled through V_{ϕ_m}.

Before the actor update, the implementation standardizes advantages within the episode. For vector $\boldsymbol{a} \in \mathbb{R}^T$, define

$$\mathcal{N}(\boldsymbol{a}) = \frac{\boldsymbol{a} - \bar{a}\mathbf{1}}{s(\boldsymbol{a}) + 10^{-8}}, \qquad \bar{\boldsymbol{A}}^{(m)} = \mathcal{N}(\hat{\boldsymbol{A}}^{(m)}), \tag{5}$$

where $s(\boldsymbol{a})$ is its empirical standard deviation. PPO then maximizes

$$L_m^{\text{clip}}(\theta) = \frac{1}{T}\sum_{t=1}^{T} \min\Big(\rho_t(\theta)\bar{A}_t^{(m)}, \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\bar{A}_t^{(m)}\Big), \tag{6}$$

with $\rho_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$.

Away from clipping boundaries, let $c_t^{(m)}(\theta) = \rho_t(\theta)$ when the sample lies on the active branch of the objective and zero otherwise. Specifically, the active set contains $\bar{A}_t^{(m)} \geq 0$ with $\rho_t \leq 1+\epsilon$ and $\bar{A}_t^{(m)} < 0$ with $\rho_t \geq 1-\epsilon$. The actor gradient can be written

$$\boldsymbol{g}_m(\theta) = \frac{1}{T}\sum_{t=1}^{T} w_t^{(m)} \nabla_\theta \log \pi_\theta(a_t \mid s_t), \qquad w_t^{(m)} = c_t^{(m)}(\theta)\bar{A}_t^{(m)}. \tag{7}$$

If $\boldsymbol{S} \in \mathbb{R}^{T\times d}$ stacks the policy score vectors by row, then $\boldsymbol{g}_m = T^{-1}\boldsymbol{S}^\top \boldsymbol{w}^{(m)}$. We call $\boldsymbol{w}^{(m)}$ the *effective advantage weights*. The operative map is therefore

$$\{u_t^{(m)}\} \longrightarrow \hat{\boldsymbol{A}}^{(m)} \longrightarrow \mathcal{N}(\hat{\boldsymbol{A}}^{(m)}) \longrightarrow \boldsymbol{w}^{(m)} \longrightarrow \boldsymbol{S}^\top \boldsymbol{w}^{(m)}. \tag{8}$$

Semantic distinction at the left edge is neither sufficient nor necessary for separation at the right edge.

3.2 Four routes to non-separation

The formulation makes four compression mechanisms explicit.

Positive-affine invariance. Ignoring the numerical stabilizer, $\mathcal{N}(a\boldsymbol{a}+b\boldsymbol{1}) = \mathcal{N}(\boldsymbol{a})$ for every $a > 0$. Consequently, an auxiliary term that mainly shifts or positively rescales the complete within-episode advantage cannot be distinguished by the actor before clipping. Raw magnitude alone is therefore a poor proxy for actor influence.

Score-space null components. On a fixed batch and update point, two methods have the same actor gradient whenever $\boldsymbol{S}^\top(\boldsymbol{w}^{(m)} - \boldsymbol{w}^{(n)}) = 0$. A familiar population-level special case is a state-dependent, action-independent component $q(s_t)$, because

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[q(s)\nabla_\theta \log \pi_\theta(a \mid s)\right] = 0. \tag{9}$$

Thus, reward variation can be large yet lie largely outside the policy's score-weighted action-discriminative subspace.

Clipping. Differences concentrated on samples whose likelihood ratios are already clipped contribute zero local gradient through those samples. Methods may therefore disagree in normalized advantages without disagreeing proportionally in the executed PPO step.

Trajectory and critic coupling. After updates diverge, methods collect different trajectories and fit different critics; the score matrix in equation 7 is no longer shared. This coupling can amplify or erase earlier differences. On a common batch, a useful bound is

$$\|\boldsymbol{g}_m - \boldsymbol{g}_n\|_2 \leq \frac{\|\boldsymbol{S}\|_2}{T}\|\boldsymbol{w}^{(m)} - \boldsymbol{w}^{(n)}\|_2, \tag{10}$$

but the converse need not hold because $\boldsymbol{S}^\top$ has a null space.

What the pipeline can compress is described in these statements. Since score matrices, gradient vectors, clipping indicators, and transition-level advantage components were not recorded, they are unable to determine which mechanism produced the empirical result. Direct update equality is still a hypothesis for future instrumentation, but the experiment can demonstrate outcome non-separation.

4 Controlled Experimental Design

4.1 GridWorld as a comparison instrument

The environment is not the suggested contribution of the paper; rather, it is a tool for a controlled comparison. The agent observes five spatial channels that encode its position, terrain-dependent cost, dynamic risk, static mines, and a moving enemy in this dynamic 8×8 survival GridWorld. There are four cardinal actions to choose from. Evaluation success is defined as staying alive for 100 steps;

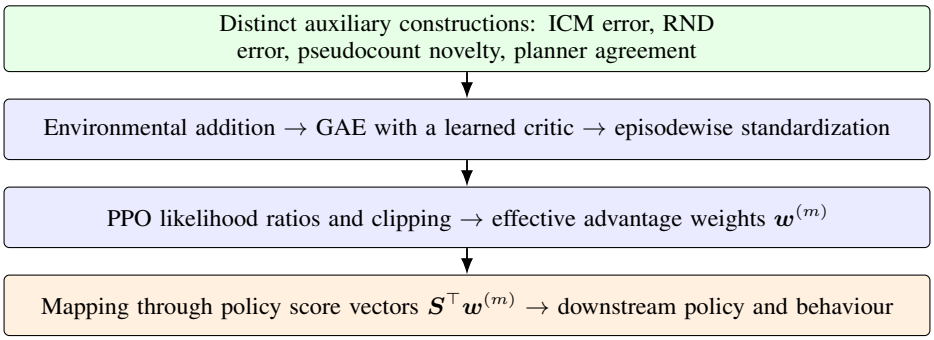


Figure 1: The effective-update view.

Table 1: Controlled system matrix.

System	Intrinsic signal	Planner reward	Executed action
PPO	-	-	PPO sample
PPO+ICM	ICM	-	PPO sample
PPO+RND	RND	-	PPO sample
PPO+PC	PC	-	PPO sample
PPO+Planner	-	agreement	PPO sample
PPO+ICM+Planner	ICM	agreement	PPO sample

there is no destination reward. An apparently safe action can turn deadly before the next decision is made because the enemy moves randomly after the agent takes action. Complete definitions of the environment and rewards are given in Appendix A.

All systems use the same PPO actor-critic, action mask, observations, maps, training schedule, and optimization hyperparameters. The mask removes moves that are currently outside the grid or visibly lethal, but cannot predict the enemy's next displacement. The six conditions are shown in Table 1. Intrinsic coefficients decay from 0.10 to 0.05. The one-step planner scores currently available neighbours using cost, risk, and within-episode revisitation. It supplies a decaying agreement reward only when the sampled PPO action matches its recommendation in a low-risk successor; it never selects or replaces the executed action.

No cross-method bonus normalization, coefficient sweep, or method-specific retuning is performed. This is deliberate. Scale-matched or individually retuned variants address a different question: whether the methods can be made behaviourally separable through calibration. The present study instead asks whether conceptual differences among auxiliary objectives are sufficient to produce meaningful separation under a fixed, shared PPO pipeline. The configured systems are therefore compared under common optimization choices; conclusions about what method-specific tuning could achieve are outside the study's scope.

4.2 Pairing, held-out splits, and scale

The comparison is deliberately paired so that variation unrelated to the auxiliary mechanism is minimized. Thirty training, ten validation, and thirty test maps are generated independently. Within each of 40 matched seeds, all six variants start from identical PPO weights, use the same 30 training maps and start states, and receive matched map rotations, environment streams, action-sampling streams, and evaluation streams. Across variants, observations, architecture, action mask, training schedule, and optimizer settings are held fixed.

The design therefore contains

$$40 \text{ matched seeds} \times 6 \text{ systems} = 240 \text{ trained policies}. \tag{11}$$

Each variant is trained for 500 episodes, yielding **120,000 training episodes**. Every policy is then evaluated with 20 stochastic rollouts per validation map, 20 stochastic rollouts per test map, and 20 secondary greedy rollouts per test map, yielding **336,000 held-out evaluation episodes**. Each

seed-level primary test mean aggregates 600 held-out stochastic test episodes. The scale and pairing distinguish the result from an underpowered five-seed null comparison.

The primary outcome is the fraction of test episodes that reach 100 steps alive. Extrinsic test return is secondary; auxiliary reward is never included in a reported return. A pre-training map gate excludes trivially survivable and observably unsolvable split realizations. A separate two-seed calibration requires every variant to exceed 0.05 stochastic test survival before the confirmatory run is authorized. Calibration ranking is not used for model selection. Full details appear in Appendix D.

4.3 Confirmatory analysis

For outcome Y and comparator c, the paired seed-level effect is

$$d_{i,c} = Y_{i,\mathrm{hyb}} - Y_{i,c}, \qquad i = 1, \ldots, 40. \tag{12}$$

Two-sided paired t intervals and tests are reported. Family-wise error across the five hybrid-comparator tests is controlled separately for survival and reward by Holm's procedure. Hybrid survival superiority was prespecified to require all three conditions against every comparator: adjusted $p < 0.05$, a confidence-interval lower bound above zero, and an absolute gain of at least five percentage points. A secondary factorial contrast tests non-additive ICM-planner interaction:

$$\Delta_i^{\mathrm{int}} = Y_i^{\mathrm{ICM+Planner}} - Y_i^{\mathrm{ICM}} - Y_i^{\mathrm{Planner}} + Y_i^{\mathrm{PPO}}. \tag{13}$$

5 Results

5.1 All six methods occupy a 0.62-point range

All six methods occupy a 0.62-percentage-point range in mean held-out survival. Across 40 matched seeds, the lowest mean is 32.63% and the highest is 33.25% (Table 2). This narrow range is the primary descriptive result.

Table 2: Held-out stochastic test performance over 40 matched seeds. Survival intervals are 95% intervals over seed-level means; numerical rank is secondary.

System	Survival	95% CI	Extrinsic return
PPO	32.95%	[32.22, 33.67]	−21.10
PPO+ICM	32.70%	[32.04, 33.37]	−21.39
PPO+RND	32.75%	[31.85, 33.66]	−21.11
PPO+PC	32.63%	[31.80, 33.47]	−21.09
PPO+Planner	32.78%	[31.99, 33.56]	−21.42
PPO+ICM+Planner	**33.25%**	[32.46, 34.03]	−21.78

We only observe the rank after determining the range: PPO+ICM+Planner has the highest numerical value at 33.25 percent. In terms of mean extrinsic return, it comes in last. These descriptive rankings do not have to match because return accrues risk and expense at every stage, while survival is a fixed-horizon event.

5.2 Confirmatory effects exclude the prespecified benefit

The confirmatory result is not $33.25 > 32.95$. The hybrid-minus-comparator effect is the pertinent paired quantity. Across the five comparisons,

$$\Delta_{\max} = 0.61 \text{ percentage points}, \qquad \max_c \mathrm{CI}_{95,c}^{\mathrm{upper}} = 1.80 < 5 \text{ points}. \tag{14}$$

Thus, every confidence interval includes zero, every Holm-adjusted p value is 1.000, and even the largest upper confidence endpoint lies below the prespecified five-point practically meaningful advantage. The global superiority criterion is not met against any comparator, let alone all five.

Table 3: Paired hybrid-minus-comparator survival effects in percentage points.

Comparator	Effect	95% CI	Holm p
PPO	$+0.30$	$[-0.54, 1.14]$	1.000
PPO+ICM	$+0.54$	$[-0.34, 1.42]$	1.000
PPO+RND	$+0.49$	$[-0.44, 1.42]$	1.000
PPO+PC	$+0.61$	$[-0.58, 1.80]$	1.000
PPO+Planner	$+0.47$	$[-0.28, 1.23]$	1.000

The survival interaction is $+0.71$ points (95% CI $[-0.46, 1.89]$, $p = .228$); the return interaction is -0.063 (95% CI $[-1.776, 1.649]$, $p = .941$). No positive ICM-planner interaction is supported. Validation and greedy results, paired return effects, and artifact counts are reported in Appendix E.

6 Different Signals, Similar Outcomes

The auxiliary signals themselves were not similar. Across the 20,000 training episodes in each single-mechanism intrinsic condition, their mean per-step values span more than three orders of magnitude, and their episode totals have sharply different relationships with episode length (Table 4). Planner shaping was not recorded as a separate transition-level column and is therefore excluded from this diagnostic.

Table 4: Raw auxiliary-signal diagnostics for the single-mechanism intrinsic conditions. Correlation is between total intrinsic reward and episode length.

Method	Mean bonus/step	Corr. with episode length
ICM	6.53×10^{-4}	.157
RND	2.20×10^{-3}	.914
Pseudocount	9.45×10^{-1}	.999

The central empirical contrast is therefore

$$\boxed{\text{different raw signals} \not\Rightarrow \text{different downstream survival}}. \tag{15}$$

Table 4 rules out the trivial explanation that the mechanisms generated the same raw quantity.

Magnitude. ICM and RND are tiny relative to the environmental reward on their recorded scale, and their contribution is smaller still after multiplication by the decaying intrinsic coefficient. Raw scale alone, however, cannot explain the entire result: the pseudocount bonus is roughly three orders of magnitude larger than ICM.

Temporal structure. The pseudocount bonus is close to one per step, and its episode total correlates at .999 with episode length. Its large raw total therefore contains a strongly duration-linked component rather than necessarily supplying equally large action-discriminative information.

Information overlap. The planner scores cost and risk channels that are already observed by PPO and penalized by the environmental reward. Its agreement reward may therefore reinforce advantage information already available to the actor rather than introduce a new preference over actions.

Standardization. Positive affine transformations of a complete within-episode advantage vector yield the same standardized vector, aside from the numerical stabilizer, as demonstrated by the normalizer in equation 5. Therefore, even when the original rewards are still numerically distinct, offset and scale differences may vanish prior to clipping. When taken as a whole, these findings link the measured diagnostics to the effective-update paths created in Section 3.

7 DISCUSSION

7.1 WHAT THE EXPERIMENT ESTABLISHES

None of the tested auxiliary mechanisms generated a practically significant survival advantage under this particular environment, architecture, 500-episode training budget, and coefficient configuration. Every paired hybrid-comparator interval excludes the predetermined five-point benefit; the largest upper endpoint is 1.80 points. The six systems span a range of 0.62 points. As a result, the experiment creates a controlled example of *auxiliary-objective non-separation*: under a shared PPO pipeline, conceptually and numerically different signal generators do not produce meaningfully separated held-out survival.

The outcome does not imply universal algorithmic equivalence or ineffectiveness in other contexts; rather, it is an existence statement about held-out survival under the evaluated configuration. Therefore, without the paired uncertainty, practical threshold, and contradictory return ranking, it would be deceptive to promote the hybrid's top numerical rank.

7.2 WHY MIGHT DIFFERENT BONUSES FAIL TO SEPARATE?

The data are consistent with five mechanisms: action-independent prediction noise from stochastic enemy motion; temporal redundancy in the duration-linked pseudocount signal; magnitude suppression for small ICM and RND contributions; alignment between planner advice and current environmental advantages; and compression via episodewise standardization and PPO clipping.

Similar final metrics can arise from similar gradients, different gradients that converge to behaviourally similar policies, insufficient training for differences to accumulate, or a common task bottleneck. Section 3 turns these alternatives into quantities that can be distinguished through direct instrumentation.

At least three levels should be measured in comparisons: (i) the raw signal, (ii) the effective update, and (iii) downstream behaviour. Reporting only (i) encourages semantic overinterpretation, and reporting only (iii) is unable to pinpoint the location of the disappearance of separation. A more robust protocol would measure clip-set overlap, compute gradient norms and pairwise cosine similarity on shared batches, log extrinsic and auxiliary TD residuals independently, compare normalized advantage rank correlations, and track policy KL and action-disagreement trajectories. Instead of being combined with the fixed-pipeline comparison, scale-matched or retuned variants ought to be reported as calibration studies.

7.3 FEASIBILITY BEFORE SEPARATION

Feasibility is a secondary hypothesis for future study: full current spatial observation, a common mask for visibly lethal moves, stochastic post-action enemy motion, and a 100-step survival objective may leave limited action-discriminative room for auxiliary differences. A factorial study varying observability, masking, horizon, hazard stochasticity, and trajectory length would be needed to test whether auxiliary methods separate only when the base task is sufficiently learnable for their information to affect action choice, yet not so constrained or saturated that all methods occupy the same behavioural regime.

7.4 LIMITATIONS

Only one 8×8 environment family, four discrete actions, full current spatial observation, one actor-critic architecture, and one training budget are studied. Test maps are fresh but sampled from the same generator. The action mask may reduce room for auxiliary methods to help, while post-action enemy motion leaves residual risk that a one-step planner cannot eliminate. The planner is deliberately myopic, and hyperparameter robustness is unknown.

Advantage decompositions, clipping indicators, policy-score mappings, gradient vectors, and action-distribution distances are not recorded; the candidate mechanisms in Section 7 therefore cannot be causally distinguished. The findings may change under partial observation, continuous control, multi-step planning, longer training, different intrinsic coefficients, normalized bonuses, or a different hazard process. No inference is made about continuous-control benchmarks or physical systems.

8 CONCLUSION

Mathematical differences among auxiliary objectives do not automatically survive a shared policy-optimization pipeline. The recorded ICM, RND, and pseudocount signals in this controlled PPO study are still clearly different, but six systems—including planner variants—finish within 0.62 percentage points of held-out survival. There is no statistically supported comparison, no practically significant prespecified advantage, and no supported ICM-planner interaction for the numerically leading hybrid.

The more general lesson is to assess auxiliary mechanisms using the action-discriminative and effective temporal information that reaches the policy update. The effective-update formulation explains how GAE, critic coupling, standardization, clipping, and score-weighted mapping can form a many-to-one map from rewards to updates. To identify the location of compression during a specific run, direct instrumentation is required. Until then, comparable results should be reported as outcome non-separation rather than as evidence of identical gradients, policies, or algorithms.

AI USE STATEMENT

Generative AI tools were used to help with literature retrieval, discovery, and organization as well as language editing and rewriting for readability and clarity. The authors developed the research questions, hypotheses, experimental design, mathematical arguments, analysis, results interpretation, and scientific conclusions. The reported experimental data was not created using AI tools. The authors, who are accountable for the final manuscript and its contents, examined and validated all AI-assisted text and retrieved references.

ETHICS STATEMENT

There are no private data, learned models trained on personal data, or human participants in this work; instead, a synthetic discrete benchmark is used. The environment does not prove that a system is safe to use in transportation, robotics, the military, or other situations where safety is crucial. Because of its local action mask, collisions following stochastic hazard motion cannot be avoided. Uncertainty-aware sensing, formally defined constraints, human oversight, and independent safety validation would all be necessary for any physical transfer. There is no assertion of autonomous operational use or field readiness.

REPRODUCIBILITY STATEMENT

The environment, rewards, objectives, PPO implementation, model matrix, hyperparameters, map construction, matched-seed protocol, budget, and decision rule are described in Section 4 and Appendices A to G. Source code, tests, configurations, confirmatory maps, episode-level records, seed summaries, statistical outputs, and completeness checks required to replicate the reported confirmatory analyses are all included in the anonymized supplementary archive. Its protocol and reported pass/fail outcome are documented separately; the original machine-readable two-seed calibration artifact was not kept.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 22–31. PMLR, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://doi.org/10.1609/aaai.v32i1.11611.

Eitan Altman. Constrained markov decision processes. *CRC Press*, 1999.

Marcin Andrychowicz, Anton Raichuk, Piotr Stanczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier

Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nIAxjsniDzg.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Marc G Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. https://arxiv.org/abs/1810.12894.

Karl W. Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2020–2027. PMLR, 2021. URL https://proceedings.mlr.press/v139/cobbe21a.html.

Roger Creus Castanyer, Joshua Romoff, and Glen Berseth. Improving intrinsic exploration by creating stationary objectives. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YbZxT0SON4.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018. URL https://arxiv.org/abs/1812.02224.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1etN1rtPB.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJ6yPD5xg.

Andrew Levy, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. In *International Conference on Learning Representations (ICLR)*, 2018.

Georg Ostrovski, Marc G Bellemare, Aaron Van Den Oord, and Rémi Munos. Count-based exploration with neural density models. In *International Conference on Machine Learning (ICML)*, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. https://proceedings.mlr.press/v70/pathak17a.html.

Roberta Raileanu and Tim Rocktäschel. Decoupling exploration and exploitation in deep rl through reward uncertainty estimation. In *International Conference on Machine Learning (ICML)*, 2020.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. In *Proceedings of the NeurIPS 2019 Safety in Reinforcement Learning Workshop*, 2019. URL https://openai.com/research/safety-gym.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. doi: 10.1038/s41586-020-03051-4.

John Schulman, Philipp Moritz, Sergey Levine, Michael I Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017. https://arxiv.org/abs/1707.06347.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, Pieter Abbeel, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning (ICML)*, 2020.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

A Environment and Reward Definition

State and dynamics. The environment is an 8×8 grid with actions up, down, left, and right. The observation $o_t \in \mathbb{R}^{5\times8\times8}$ contains an agent-position channel, cost map, risk map, terrain/mine encoding, and enemy-position channel. Mud, water, and mines are encoded as 0.5, 0.8, and 1.0. Mines are static. One enemy starts at the center, and its horizontal and vertical displacements are sampled independently from $\{-1, 0, 1\}$ after each agent action.

Every map starts with base cost 0.3 and zero risk. Interior mines occur with probability 0.05 and the boundary contains mines. Mud and water are each sampled with probability 0.15. Starts are placed at least two cells from the boundary and at Manhattan distance greater than one from mines and the enemy. Cost decays by 0.98 per step. Mine cells and their four-neighbours receive cost increments of 0.5 and 0.25, capped at one. Risk decays by 0.95 and receives the same increments around the updated enemy position.

Mask. Moves outside the grid or into a currently observed mine or enemy cell are masked during both data collection and PPO updating. If every action is masked, all actions are restored to keep the categorical distribution defined. Because enemy motion occurs after action selection, this is a shared local precaution rather than a safety certificate.

Environmental and learning rewards. After the agent and enemy move and dynamic maps are updated, environmental reward is

$$\begin{aligned} r_t^{\text{ext}} = 0.5 - C_t(x_t) - R_t(x_t) - 0.2d_e\mathbb{I}_{\text{mud}} - 0.1d_e\mathbb{I}_{\text{water}} \\ - 2\mathbb{I}_{\text{mine}} - 3\mathbb{I}_{\text{enemy}} + 5\mathbb{I}_{\text{horizon}}, \end{aligned} \tag{16}$$

where $C_t, R_t \in [0, 1]$, $d_e = \max(0.1, 1 - e/1000)$ for training episode e, and the final term is given only upon reaching the horizon alive. Collision terminates the episode. Method m is optimized with

$$\tilde{r}_t^{(m)} = r_t^{\text{ext}} + \beta_t b_t^{(m)} + p_t^{(m)}. \tag{17}$$

Only r_t^{ext} enters comparable return columns.

B Auxiliary Objectives and Advisory Planner

ICM. A 128-dimensional encoder ϕ, inverse-action model, and forward-feature model are used. The per-transition bonus is

$$b_t^{\text{ICM}} = \frac{1}{128} \left\| \phi(s_{t+1}) - f(\phi(s_t), a_t) \right\|_2^2 . \tag{18}$$

The next-state target is detached, and forward and inverse losses are optimized with Adam at 10^{-3}.

RND. Fixed and trainable 128-unit networks are used:

$$b_t^{\mathrm{RND}} = \frac{1}{128} \left\| \hat{f}(s_t) - f_{\mathrm{fixed}}(s_t) \right\|_2^2. \tag{19}$$

Only the predictor is trained.

Pseudocount. The complete observation is mapped into one of 2^{16} stable BLAKE2 buckets. After the run-level count is incremented,

$$b_t^{\mathrm{PC}} = \frac{1}{\sqrt{N(h(s_t))}}. \tag{20}$$

Planner. Observed lethal neighbours are excluded. Each available successor s' is scored by

$$J(s') = w_c C(s') + w_r R(s') + 2\mathbb{I}[s' \in \mathcal{V}], \tag{21}$$

where $\mathcal{V}$ is the within-episode visited set. The cost/risk weights are $(1, 5)$ when current risk exceeds 0.7, $(2, 3)$ above 0.5, and $(3, 1)$ otherwise. The lowest-scoring action is recommendation g_t. The PPO sample a_t is always executed, and agreement contributes

$$p_t^{(m)} = \alpha_e \mathbb{I}[a_t = g_t \ \wedge \ R_t(x_{t+1}) < 0.5], \tag{22}$$

where α_e decays from 0.5 to 0.1 across 500 episodes. Planner and environment random streams are isolated.

C Training Configuration

Table 5: Shared training configuration.

Quantity	Setting
Policy/value network	MLP, hidden widths 256 and 128
Policy input/output	320 values / 4 logits
Optimizer learning rate	3×10^{-4}
Discount γ	0.99
GAE λ	0.95
PPO clip ϵ	0.2
Entropy coefficient	0.01
Value coefficient	0.5
Update cadence	one update after each episode
Training episodes	500 per seed-variant
Training horizon	linear curriculum, 20 to 100
Evaluation horizon	fixed at 100
Intrinsic coefficient	0.10 to 0.05 over first two-thirds

For every matched seed, all six variants are initialized from the same PPO parameters. The 30 training maps are rotated by index $(e + \mathrm{seed}) \bmod 30$. Starts, map order, environment randomness, and evaluation randomness are matched across variants within seed. Action-sampling, environment, and planner streams are isolated so that adding a module does not silently shift unrelated random draws.

D Protocol, Gates, and Statistical Details

Map gate. Before training, each split is evaluated with constant, random, masked-random, masked-preference, and observation-limited safe policies. A split is rejected if an unmasked constant policy exceeds 0.10 survival or the safe heuristic falls below 0.10. The gate is intended to exclude trivially survivable and observably unsolvable realizations.

Calibration. A separate two-seed run uses 150 episodes, 10/5/5 maps, and five stochastic test rollouts per map. Calibration passes only if every variant reaches 0.05 test survival. This viability floor was met. The original machine-readable calibration result bundle was not retained; the supplementary archive therefore contains the declared calibration configuration and reported pass/fail summary, but not the original per-run calibration output. No confirmatory model was selected or tuned according to calibration rank.

Evaluation inventory. For each of 40 seeds and six variants, 20 stochastic rollouts are run on each of ten validation maps and thirty test maps. Twenty greedy rollouts are additionally run on each test map. Expected completeness is therefore 120,000 training episodes, 336,000 evaluation episodes, 240 checkpoints, and 70 maps.

Inference. All primary comparisons are paired by seed. Two-sided one-sample t tests are applied to paired differences, and 95% t intervals are reported. Holm correction is applied to the five survival tests and, separately, to the five return tests. The practical survival threshold is five percentage points; the secondary return threshold is 0.5. The five-percentage-point threshold was fixed before the confirmatory analysis as a conservative decision threshold rather than estimated from the confirmatory data. The interaction contrast is reported with an unadjusted two-sided test because it is a separately declared factorial question.

E Additional Results

Table 6: Validation and secondary greedy survival across 40 seeds.

System	Validation survival	Greedy test survival
PPO	32.83%	37.16%
PPO+ICM	32.78%	36.53%
PPO+RND	32.68%	37.12%
PPO+PC	32.29%	38.02%
PPO+Planner	32.24%	36.56%
PPO+ICM+Planner	33.36%	38.38%

Table 7: Paired hybrid-minus-comparator extrinsic-return effects.

Comparator	Effect	Holm p
PPO	-0.68	.716
PPO+ICM	-0.39	1.000
PPO+RND	-0.67	.716
PPO+PC	-0.69	.716
PPO+Planner	-0.36	1.000

The hybrid's first rank under greedy survival is secondary and was not used to modify the primary stochastic claim. Its stochastic survival lead does not coincide with a return lead: every paired return difference is negative, although none is significant after correction.

F Implementation Validity and Artifact Completeness

After an implementation audit found state-action misalignment, incorrect value targets, unstable pseudocount hashing, contamination of reported reward by shaping terms, and uncontrolled random streams, earlier development runs were eliminated. Tests for ICM updates, return construction, stable hashing, map rotation, mask consistency, advisory-only planner behavior, reward separation, random-stream isolation, fixed-horizon evaluation, shard completeness, and paired Holm-corrected analysis are all included in the updated pipeline.

All declared checks were passed by the confirmatory inventory, which included 40 seeds, six variants, 500 training episodes per seed-variant, 240 checkpoints, 40 logs, and 70 maps with all declared validation, stochastic-test, and greedy-test rollouts. Results from pipelines that were either unfinished or invalid in the past are excluded.

G Reproducibility Checklist

Configuration files, split maps, seed summaries, training and evaluation episode records, statistical outputs, and tests utilizing non-identifying paths and metadata should all be preserved in the sup-

Table 8: Location of information needed to reproduce the study.

Item	Location
Environment state, dynamics, mask	Appendix A
Environmental and learning rewards	Appendix A
ICM, RND, PC, and planner definitions	Appendix B
PPO architecture and hyperparameters	Appendix C
Map splits, seeds, horizons, rollouts	Section 4; Appendix D
Primary outcome and decision rule	Section 4; Appendix D
Complete primary and secondary results	Section 5; Appendix E
Validity tests and inventory checks	Appendix F
Confirmatory code and machine-readable artifacts	Anonymized supplementary archive
Calibration protocol	Appendix D; `configs/calibration_v2.yaml`

plemental archive. During double-blind review, public repository URLs, account names, commit authors, absolute local paths, and institution-specific identifiers must not be present.