# Reward Sparsity in Multi-Objective Reinforcement Learning: A Taxonomy and Key Challenges

**Raghav Thakar**[1] , **Jamie Santos**[1] , **Ayhan Alp Aydeniz**[1] , **Gaurav Dixit**[1] and **Kagan Tumer**[1]
[1]The Collaborative Robotics and Intelligent Systems (CoRIS) Institute
Oregon State University, Corvallis, Oregon, USA
{thakarr, santjami, aydeniza, dixitg, kagan.tumer}@oregonstate.edu

## Abstract

Multi-Objective Reinforcement Learning (MORL) commonly assumes that agents receive regular feedback for every objective during training. In practice, however, reward signals are often imperfect: one objective may provide dense feedback, while another may yield rewards only after rare events, long delays, specific thresholds, or infrequent measurements. This conceptual paper formalizes the problem through two related concepts—multi-objective sparsity and heterogeneous objective sparsity—and explores their effects on Pareto optimization. We analyze how imbalances in reward density can cause dense rewards to dominate scalarized objective functions, thereby masking sparse objectives. We also show how these imbalances can produce false Pareto dominance, in which promising exploratory policies are prematurely discarded because rewards associated with sparse objectives have not yet been observed. In addition, we examine how objective-specific sparsity-mitigation mechanisms may conflict with one another and how sparse feedback complicates the interpretation of incomplete Pareto fronts. To organize these challenges, we introduce a taxonomy that classifies objective sparsity as event-contingent, delayed, threshold-activated, or measurement-limited. We further argue that applying a single, uniform exploration strategy to the entire reward vector can create conflicting optimization signals, underscoring the need for objective-specific methods to reliably approximate the true Pareto front.

## 1 Introduction

Many real-world sequential decision-making tasks require balancing conflicting objectives, such as trading off travel time, passenger comfort, and energy consumption in autonomous driving. While standard reinforcement learning (RL) maps task feedback to a single scalar reward [Sutton and Barto, 2018], Multi-Objective Reinforcement Learning (MORL) employs vector-valued rewards to learn policies that represent meaningful trade-offs across distinct objectives [Roijers *et al.*, 2013; Hayes *et al.*, 2022]. The goal is therefore not only to maximize return, but to learn policies that represent meaningful trade-offs among competing objectives.

Standard MORL frameworks implicitly assume that the expected return for every objective can be reliably estimated throughout training. This assumption collapses when objectives exhibit sparse feedback—a classic RL bottleneck that hinders exploration, reduces sample efficiency, and complicates temporal credit assignment [Hertweck *et al.*, 2020; Riedmiller *et al.*, 2018]. In multi-objective settings, this challenge is compounded by an additional challenge of *heterogeneous objective sparsity*: one objective may provide dense, immediate feedback, while another may trigger only after rare events, threshold satisfaction, or be limited by feedback measurement bottlenecks.

Existing works have addressed partial aspects of this problem. Sparse-reward methods are usually designed for a single reward signal, so they often assume that one intervention can address the main source of sparsity. MORL methods, by contrast, usually focus on how to represent trade-offs, either through scalarization, Pareto dominance, or coverage-set learning [Roijers *et al.*, 2013; Hayes *et al.*, 2022; Van Moffaert and Nowé, 2014]. And although reward shaping has been extended to multi-objective settings with policy-invariance guarantees, the broader problem of learning under heterogeneous objective sparsity remains underexplored in the literature [Mannion *et al.*, 2017; Mannion *et al.*, 2018].

However, neither line of work fully addresses the unique challenges that emerge due to multiple sparse objectives, and heterogeneous objective sparsity. When some objectives are dense and others are sparse, scalarization of rewards can cause dense objectives to dominate learning before sparse objectives are ever observed, even when the sparse objectives have high decision-maker importance [Vamplew *et al.*, 2008]. Similarly, Pareto-based filtering can discard policies whose sparse-objective value has not yet been revealed, producing misleading estimates of the Pareto front [Van Moffaert and Nowé, 2014; Zitzler *et al.*, 2003]. Lastly, different objectives may be sparse for different reasons. Applying a single sparsity-mitigation mechanism to the entire reward vector can therefore be ineffective, and may even improve learning for one objective while harming discovery of another.

In this paper, we argue that multi-objective sparsity and heterogeneous objective sparsity are both realistic conditions

that create unique MORL challenges that require explicit identification, analysis, and addressing. Thus, the contributions of this paper are as follows:

1. We identify multi-objective and heterogeneous objective sparsity as distinct sources of difficulty in MORL.
2. We provide a taxonomy of objective sparsity types, namely event-contingent, temporally delayed, threshold-activated, and measurement-limited sparsity.
3. We analyze how sparse objectives can distort Pareto-front discovery through false Pareto dominance, objective masking under scalarization, conflicting sparsity-mitigation mechanisms, and ambiguous incomplete fronts.

## 2 Preliminaries

Sequential decision-making problems are typically modeled as Markov decision processes (MDPs) [Sutton and Barto, 2018; Puterman, 2014]. An MDP is defined by the tuple

$$\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle,$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is the transition function, $r(s, a)$ is a scalar reward, and $\gamma \in [0, 1)$ is a discount factor. A policy $\pi(a \mid s)$ induces an expected discounted return

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, \pi\right]. \tag{1}$$

Multi-objective Markov decision processes (MOMDPs) replace the scalar reward with a vector-valued reward [White, 1982; Roijers *et al.*, 2013]. A MOMDP is defined as

$$\langle \mathcal{S}, \mathcal{A}, P, \mathbf{r}, \gamma \rangle,$$

where $\mathbf{r}(s, a) \in \mathbb{R}^k$ contains one reward component for each of $k$ objectives. Under a policy $\pi$, the value function is therefore vector-valued:

$$\mathbf{V}^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t \mid s_0 = s, \pi\right]. \tag{2}$$

Each component $V_i^{\pi}(s)$ represents the expected return for objective $i$. Since policies may improve one objective while degrading another, there is generally no total ordering over policies in a MOMDP. The goal of MORL is therefore often to identify a policy, a set of policies, or a preference-conditioned policy that captures meaningful trade-offs among objectives [Roijers *et al.*, 2013; Hayes *et al.*, 2022].

One common way to compare policies is through Pareto dominance. Given two policies $\pi$ and $\pi'$, policy $\pi$ Pareto-dominates $\pi'$ if it is at least as good in every objective and strictly better in at least one. Assuming maximization on each objective:

$$\pi \succ \pi' \iff \mathbf{V}^{\pi} \geq \mathbf{V}^{\pi'} \ \wedge \ \mathbf{V}^{\pi} \neq \mathbf{V}^{\pi'}. \tag{3}$$

A policy is Pareto-optimal if no other policy dominates it, and the corresponding set of non-dominated value vectors forms the Pareto front [Van Moffaert and Nowé, 2014; Roijers *et al.*, 2013]. In this paper, we distinguish between the *true* Pareto front, defined by the underlying expected return vectors, and the *observed* or *estimated* Pareto front produced by a learning algorithm during training. This distinction is central when some objectives are sparse, since low observed return may reflect missing information rather than true policy inferiority.

Another common approach is scalarization, which maps a vector-valued return to a scalar utility. For a nonnegative weight vector $\mathbf{w}$, linear scalarization evaluates a policy as

$$u_{\mathbf{w}}(\mathbf{V}^{\pi}) = \mathbf{w} \cdot \mathbf{V}^{\pi} = \sum_{i=1}^{k} w_i V_i^{\pi}. \tag{4}$$

Equivalently, scalarization may be applied to the instantaneous reward:

$$r_t^{\mathbf{w}} = \mathbf{w} \cdot \mathbf{r}_t = \sum_{i=1}^{k} w_i r_{i,t}. \tag{5}$$

This reduction allows standard single-objective RL algorithms to be applied to multi-objective problems [Natarajan and Tadepalli, 2005; Castelletti *et al.*, 2002; Van Moffaert *et al.*, 2013]. However, scalarization only optimizes with respect to a specified preference vector. When scalarization is linear, it may fail to recover Pareto-optimal solutions in non-convex regions of the Pareto front [Roijers *et al.*, 2013; Van Moffaert *et al.*, 2013]. More importantly for this work, scalarized learning assumes that each reward component contributes sufficiently reliable information during training. When objectives differ in sparsity, this assumption can fail.

## 3 A Brief Review and Taxonomy of Reward Sparsity

Sparse rewards represent one of the central challenges in RL, arising whenever the reward signal provides useful feedback only rarely relative to the frequency of agent–environment interaction [Ladosz *et al.*, 2022]. Formally, we say that a reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is *sparse* if the set of state–action pairs for which $R(s, a) \neq 0$ has low measure relative to $\mathcal{S} \times \mathcal{A}$. In the extreme case, an agent may interact with its environment for thousands of steps before receiving any informative feedback, making policy gradient estimation highly noisy and exploration fundamentally difficult [Sutton and Barto, 2018].

Sparse reward problems arise naturally across a broad range of real-world applications. A robot arm learning to grasp objects receives no reward until a successful grasp is achieved, a financial trading agent may only observe the profitability of its decisions after a position is closed days later, and a planetary rover operating under communication constraints receives feedback only during infrequent uplink windows, regardless of task performance. Despite sharing the surface characteristic of reward scarcity, these examples differ fundamentally in the *cause* of sparsity, and consequently in the algorithmic strategies best suited to address them. This motivates the taxonomy we develop in the remainder of this section.

**Event-Contingent Sparsity**

Event-contingent sparsity arises when reward is withheld until a specific discrete event occurs, most commonly, the agent reaching a designated goal state $s_g \in \mathcal{S}$. Formally, the reward function takes the form

$$R(s, a, s') = \begin{cases} 1 & \text{if } s' = s_g, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

so that feedback is entirely absent along any trajectory that fails to reach the goal.

The primary difficulty is one of *exploration*: without intermediate shaping signals, the agent must stumble upon the goal state by chance before any learning can occur. In high dimensional or large state spaces, the probability of doing so under a random policy is negligible, rendering naive exploration strategies such as $\varepsilon$-greedy ineffective. Credit assignment across the trajectory is also problematic, since all transitions preceding the goal receive identical zero reward.

Goal-conditioned formulations and hindsight relabeling have proven particularly effective in this setting. Hindsight Experience Replay (HER) [Andrychowicz *et al.*, 2017] addresses the exploration problem by retroactively treating any reached state as a valid goal, thereby converting failed trajectories into informative training signal. Curriculum-based methods [Florensa *et al.*, 2017] alleviate the exploration burden by initializing episodes progressively closer to the goal, while intrinsic motivation approaches [Burda *et al.*, 2018; Pathak *et al.*, 2017] augment the sparse extrinsic reward with a dense curiosity-driven signal to encourage broader state coverage.

**Temporal Delay Sparsity**

Temporal delay sparsity occurs when a reward signal exists but is not delivered until a significant number of timesteps after the actions that caused it. Unlike event-contingent sparsity, the reward is not structurally absent; rather, it is episodically aggregated and revealed only at a terminal or checkpoint step. Formally, an agent operating over a horizon $T$ may receive no intermediate reward ($r_t = 0$ for $t < T$), with all feedback collapsed into a single terminal signal $r_T$.

The core difficulty is *long-horizon credit assignment*, or identifying which actions among potentially thousands were responsible for the eventual outcome. Standard temporal difference methods suffer from vanishing or highly variable gradient estimates when the return must be propagated across many steps, and experience replay buffers are less efficient when informative transitions are rare. Though temporal delay sparsity and event-contingent sparsity share common symptoms, temporally delayed rewards are causally determined early but revealed late, whereas, in event-contingent sparsity, the reward does not exist until the event happens.

RUDDER [Arjona-Medina *et al.*, 2019] addresses temporal delay sparsity directly by learning a return decomposition that redistributes the delayed reward back to the individual transitions most causally responsible for it, effectively transforming a delayed-reward MDP into one with dense, well-attributed feedback. Reward redistribution has since been explored through transformer-based sequence models [Ren *et al.*, 2021] and potential-based reward shaping [Ng *et al.*, 1999], the latter providing theoretical guarantees of policy invariance under reshaping. Eligibility traces [Sutton and Barto, 2018] offer a classical mechanism for softening the credit assignment problem by maintaining a decaying memory of recent state–action pairs.

**Threshold-Activated Sparsity**

Threshold-activated sparsity arises when reward is suppressed until the agent's cumulative or instantaneous performance crosses a predetermined threshold $\tau$. Formally,

$$R(s, a) = \begin{cases} f(s, a) & \text{if } \Phi(s, a) \geq \tau, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\Phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a performance measure and $f$ is the underlying reward signal. Crucially, threshold-activated sparsity introduces a discontinuity in the reward landscape, making gradient-based optimization particularly susceptible to becoming trapped in flat, zero-reward regions. In multi-objective settings, the problem is compounded as different objectives may activate at different thresholds, creating a complex, non-stationary feedback structure that is difficult to navigate without prior knowledge of the threshold geometry [Kaushik *et al.*, 2018].

A representative real-world instance is industrial process control. In a chemical reactor, rewards associated with throughput and energy efficiency may be withheld entirely until product purity crosses a regulatory minimum threshold $\tau$, so that the agent may improve continuously along secondary objectives yet receive no feedback whatsoever while the primary quality constraint remains unsatisfied. This setting is inherently multi-objective (the competing criteria being purity, throughput, and energy cost) making it a natural instance of threshold-activated sparsity within the MOMDP framework introduced in Section 2.

Model-based methods have shown promise here, as learning a forward model of the environment may allow the agent to reason about threshold-crossing transitions without requiring direct observation of reward [Kaushik *et al.*, 2018]. Constrained RL formulations [Achiam *et al.*, 2017] can encode threshold requirements as hard constraints and thus separate performance optimization from feasibility.

**Measurement-Limited Sparsity**

Measurement-limited sparsity is distinct from the preceding categories. Here, sparsity is not a feature of the reward definition but a constraint imposed by the physical or economic cost of observing or communicating the system state. The reward function itself may be well-defined and dense in principle, but the agent may only be able to query it infrequently. Formally, let $\omega_t \in \{0, 1\}$ be an observation indicator; the agent receives $r_t = R(s_t, a_t) \cdot \omega_t$, where $\omega_t = 1$ only at timesteps for which measurement is feasible, and $\omega_t = 0$ otherwise.

Active reward learning provides one mechanism for reducing the effect of measurement-limited sparsity when reward observations can be actively selected. Daniel et al. [Daniel *et al.*, 2014] consider settings in which direct reward specification or evaluation is costly and instead learn a probabilistic reward model from expert ratings of only a subset of the

agent's rollouts. Instead of addressing the underlying measurement constraint, this work converts sparse reward observations into an actively maintained surrogate signal that can support policy improvement between costly measurements.

# 4 How the Presence of Sparse Objectives Affects the Availability of Trade-Offs

Reward sparsity affects MORL in two distinct ways. First, it can make an existing trade-off difficult to discover because the relevant reward signal is rarely observed. Second, and more subtly, it can determine whether intermediate trade-offs are available at all. In this section, we outline how the form of reward sparsity affects the which trade-offs are available to be uncovered.

## 4.1 Sequence-Contingent Sparsity Can Eliminate Intermediate Trade-Offs

In event-contingent sparse objectives, reward is obtained only when a particular event or outcome occurs. For example, an agent may receive reward for retrieving an object in a maze only after executing a specific sequence of navigation actions. Before the object is retrieved, the objective provides no graded signal indicating partial progress.

This kind of sparsity can limit the availability of trade-offs. If an objective is only realized after completing a specific behavioral sequence, then policies that deviate from that sequence may receive no reward for that objective at all. As a result, the objective may behave like a binary constraint: either the agent completes the sequence and obtains the reward, or it does not. In such cases, there may be few meaningful intermediate policies that partially optimize the sparse objective while trading it against another objective.

For example, suppose one objective rewards object retrieval and another rewards minimizing energy use. If the object-retrieval reward is only obtained after the full retrieval sequence is completed, then policies that save energy by stopping halfway do not represent smooth trade-offs between retrieval and energy. They simply fail the retrieval objective. The resulting Pareto front may therefore contain discontinuities or isolated regions rather than a smooth continuum of trade-offs.

## 4.2 Delayed Feedback Can Hide Existing Trade-Offs

Delayed sparsity has a different effect. In this case, the reward signal is sparse because feedback arrives late, not because intermediate behavior is irrelevant. The agent may make many decisions that influence the final outcome, but the reward is only revealed at the end of an episode or after a long delay.

Unlike sequence-contingent sparsity, delayed feedback does not necessarily eliminate trade-offs. The underlying problem may contain many policies that balance objectives in different ways. However, these trade-offs may be difficult to uncover because the learner cannot easily assign credit to the earlier actions that produced the delayed reward.

For example, if one objective measures the total distance walked by a robot biped, and another measures cumulative resource consumption, then different policies may achieve different trade-offs. A high-distance covering policy may use more energy, while a stationary policy may conserve more energy. These trade-offs exist in the underlying MOMDP, even if the reward for distance covered is only observed at the end of the episode. In this case, sparsity primarily obscures the Pareto front rather than changing its structure.

## 4.3 Threshold-Activated Sparsity Can Create Hidden Fronts

A third case occurs when reward is activated only after performance crosses a threshold. Unlike purely event-contingent sparsity, threshold-activated sparsity may allow gradual improvement below the threshold, but this improvement remains invisible to the reward function until the threshold is reached.

This creates a hidden-front problem. Policies below the threshold may differ meaningfully in their proximity to satisfying the sparse objective, but they all receive the same observed reward. Consequently, the learner may be unable to distinguish policies that are close to revealing a new trade-off region from policies that are genuinely poor with respect to that objective.

For example, suppose an objective rewards a robot only if it transports at least $90\%$ of a payload, while another objective penalizes energy consumption. Policies transporting $40\%$, $70\%$, or $89\%$ of the payload may all receive zero reward for the payload objective, even though they differ substantially in how close they are to activating it. Once the threshold is crossed, new trade-offs between payload delivery and energy use may suddenly become visible. Thus, threshold-activated sparsity can make the empirical Pareto front appear incomplete until a critical performance boundary is crossed.

## 4.4 Measurement-Limited Rewards may Produce Deceptive Policy Evaluations

In multi-objective settings, measurement-limited sparsity is particularly damaging because observation gaps may affect all objectives simultaneously, yet not necessarily uniformly. Different objectives may have different sensitivities to feedback absence (a safety-critical objective may degrade rapidly without measurement while a secondary performance objective remains stable), making it difficult to maintain a coherent estimate of the Pareto front during "silent" intervals. This uneven information loss can cause the learned policy to drift toward solutions that appear Pareto-optimal under the limited observations available, but are in fact dominated when the full reward structure is revealed.

An additional issue may arise when a policy is re-evaluated with sporadically absent rewards, as it may cause the same policy to yield a slightly different cumulative return vector. This may cause the perception of an available trade-off to shift, or may cause the a policy's 'true' multi-objective performance to be miscalculated.

# 5 The Unique Challenges of Multiple, Possibly Sparse, Objectives

Reward sparsity and multi-objective learning have each been studied extensively, but their intersection raises challenges

that are not reducible to either setting alone. In single-objective RL, sparse rewards are primarily understood as creating difficulties for exploration, sample efficiency, and temporal credit assignment. Methods such as Hindsight Experience Replay [Andrychowicz *et al.*, 2017], curiosity-driven exploration [Pathak *et al.*, 2017], count-based exploration [Bellemare *et al.*, 2016], and return decomposition methods such as RUDDER [Arjona-Medina *et al.*, 2019] all address, in different ways, the problem that an agent may receive little or no informative feedback for long periods of training.

In MORL, by contrast, the central difficulty is usually framed as one of trade-off representation: how to learn a policy, a set of policies, or a preference-conditioned policy that captures the relevant Pareto-optimal solutions [Roijers *et al.*, 2013; Vamplew *et al.*, 2011; Hayes *et al.*, 2022]. Existing MORL methods often assume that objective returns can be meaningfully estimated and compared during learning, whether through scalarization, Pareto dominance, or hypervolume maximization [Auger *et al.*, 2009]. However, this assumption becomes problematic when the individual reward components differ in their sparsity. In such cases, low observed return for an objective may not simply indicate that the current policy performs poorly with respect to that objective, but also that the objective has possibly not yet been sufficiently explored, observed, or credited.

This distinction gives rise to the problem of *multiple, possibly sparse objectives*: multi-objective learning problems in which each reward component may exhibit a different degree or type of sparsity. In this section, we specify some distinct learning challenges this setting may bring.

### 5.1 False Pareto Dominance

The first major challenge brought by reward sparsity is that it can distort Pareto comparisons. The Pareto dominance relation between two policies is determined simply by comparing their objective values. However, when rewards are sparse, observed returns may not be accurate estimates of the actual quality of policies.

Consider a problem with two competing objectives: a dense objective 1 and a sparse objective 2. Consider two policies: $\pi_a$, that predominantly seeks to optimize objective 1, and $\pi_b$, that predominantly seeks to optimize objective 2. When compared purely on their objective returns, $\pi_a$ likely dominates $\pi_b$ owing to objective 1 being easier to optimize quickly, and $\pi_b$ focusing on, but struggling to optimize objective 2 (and thus optimizing neither).

The above scenario gives rise to what we may call, *false Pareto dominance*: an apparent dominance relation induced by missing information rather than true inferiority. Even if $\pi_a$ reliably achieves only moderate dense rewards, while $\pi_b$ explores regions that could eventually produce high sparse rewards but has not yet received them, standard Pareto filtering may discard $\pi_b$ before its sparse-objective potential becomes observable. Discarding or ignoring exploratory policies would hamper any progress made toward the sparse objectives requiring that exploration, severely limiting an algorithm's ability to produce policies that optimize them.

### 5.2 Objective Masking Under Scalarization

A second challenge is that sparse objectives can be masked by denser objectives during optimization. Scalarization is a central tool in MORL. However, when objectives are scalarized into a single reward or utility signal, the resulting scalar feedback combines not only the relative importance of each objective, but also the magnitude, frequency, and timing of the corresponding reward components.

Consider a scalarized reward for a $k$-objective problem:

$$r_t^{\mathbf{w}} = \sum_{i=1}^{k} w_i r_{i,t}. \tag{8}$$

If one objective provides a small reward at every timestep while another provides a large reward only after a rare event or is held back by measurement limits, then the dense objective may dominate the scalarized return for most of training. This dominance is not a consequence of the decision maker's preferences but of the availability of the reward signal. The sparse objective contributes minimally to the scalarized reward until the agent discovers the states, trajectories, or threshold conditions that activate it, or may keep doing so if measurement is limited. Thus, even a high weight $w_i$ may have no practical effect before discovery. Conversely, assigning a very large weight to compensate for sparsity can produce unstable updates once the sparse reward is finally observed.

This produces an *objective masking* effect: a dense objective can make progress appear meaningful while a sparse objective remains effectively absent from the learning signal. In such cases, scalarized learning may converge toward policies that optimize the most frequently observed objectives rather than policies that reflect the intended trade-off among objectives. This limitation differs from the frequently cited concern that linear scalarization cannot recover policies on non-convex regions of the Pareto front [Van Moffaert *et al.*, 2013]. Even if a scalarization function is capable of representing the desired solution, the learning process may fail to discover that solution because the sparse objective is unable to contribute useful temporal-difference targets, policy-gradient estimates, or action preferences during early exploration.

### 5.3 Conflicting Mechanisms to Address Reward Sparsity

In single-objective sparse-reward RL, the mechanism used to address sparsity is typically matched to the structure of the reward. Sparse binary goal-reaching rewards can be addressed by goal relabeling methods such as Hindsight Experience Replay (HER), which extract learning signal from failed trajectories by reinterpreting them as successful with respect to alternative achieved goals [Andrychowicz *et al.*, 2017]. Objectives whose rewards are sparse because relevant states are rarely discovered may instead benefit from intrinsic motivation or count-based exploration, which explicitly bias the agent toward novel or uncertain regions of the state space [Pathak *et al.*, 2017; Bellemare *et al.*, 2016]. Objectives with long-delayed feedback may require credit-assignment mechanisms such as return redistribution, as in RUDDER, rather

than exploration alone [Arjona-Medina *et al.*, 2019]. Thresholded objectives may require curricula that gradually expose the agent to solvable instances near the reward boundary [Florensa *et al.*, 2017]. Safety-related reward components or constraints may require conservative exploration that avoids high-risk states even when those states are novel [García and Fernández, 2015].

In multi-objective settings, these mechanisms can have objective-dependent effects. For instance, a novelty bonus that helps uncover a sparse task-completion reward may increase visits to unfamiliar states, but those same visits may make a safety-related objective or constraint harder to satisfy. A curriculum that begins near one objective's success region may provide useful learning signal for that objective while undersampling states needed to discover another. Similarly, return redistribution can help with delayed rewards once informative trajectories have been observed, but it does not by itself solve the problem of discovering rare reward-triggering states. Thus, different sparse objectives may expose different failure modes during learning.

This creates a challenge for any single, global sparsity-addressing strategy. If the agent receives a vector reward

$$\mathbf{r}_t = [r_{1,t}, \ldots, r_{k,t}], \tag{9}$$

then each component $r_i$ may be sparse for a different structural reason. Applying one auxiliary mechanism uniformly across the entire problem implicitly assumes that the same intervention is appropriate for all reward components. This assumption may fail when objectives differ in whether their sparsity arises from rare-state discovery, delayed feedback, threshold effects, or measurement limits.

More importantly, blending sparsity-addressal mechanisms does not necessarily produce meaningful trade-off discovery. Such mechanisms are not neutral estimators of missing rewards. Instead, they may change the distribution of experience, the temporal placement of feedback, the auxiliary objectives being optimized, or the effective task presented to the learner itself. When these interventions are applied globally, the resulting policy may reflect compromises among the interventions rather than among the original objectives. Moreover, mechanisms designed for one sparsity structure may rely on assumptions that do not hold for another. For example, an intervention that improves rare-event discovery may be irrelevant to delayed feedback.

Therefore, sparsity-addressal mechanisms should be treated as objective-specific interventions rather than as generic additions to the multi-objective reward function. Each objective may require its own diagnosis, and only after this diagnosis can the learner decide which mechanisms are appropriate. Mixed Advantage Pareto Extraction (MAPEX) [Thakar *et al.*, 2026] provides a promising avenue to achieve this by isolating the learning process on each objective, and combining each objective's 'specialist' solutions in a post-hoc manner to uncover trade-offs. In general, the selection and coordination of sparsity-addressal mechanisms in sparse MORL settings is a relevant, yet wide open research area.

A useful design principle, therefore, is to separate the rewards used to support exploration from the rewards used for Pareto evaluation. Shaped rewards such as novelty bonuses, curricula, or redistributed rewards may help the learner discover informative regions of the state space, but Pareto dominance should ultimately be computed using the original vector return. This separation prevents auxiliary learning signals from redefining the trade-offs. In other words, shaping may guide the search process, but the nondominated set should be assessed with respect to the original objectives, not the shaped or auxiliary rewards that helped make those objectives observable.

### 5.4 Interpreting Incomplete Pareto Fronts

A further challenge is that reward sparsity complicates the interpretation of incomplete Pareto fronts. In standard MORL, sparse coverage of the objective space is often treated as an approximation failure, such that if a region of the Pareto front is missing, the algorithm may allocate additional search or training effort toward that region. This assumption underlies many Pareto-front discovery methods, where poorly covered regions are interpreted as areas requiring further optimization or policy refinement [Liu *et al.*, 2025; Liu *et al.*, 2024; Alegre *et al.*, 2023].

Under heterogeneous reward sparsity, however, an incomplete observed front is ambiguous. A missing region may indicate that the corresponding trade-off truly does not exist; that the region exists but has not been explored; or that the objective is threshold-activated and therefore appears uninformative until a minimum level of performance is reached. This ambiguity is especially problematic for adaptive MORL algorithms that use the current front to guide future search. If a sparse region is caused by insufficient exploration, additional training may uncover valuable policies, while if it is caused by a genuine absence of feasible trade-offs, the same effort may be wasted. Therefore, in sparse multi-objective settings, incomplete Pareto fronts should not be interpreted purely as coverage failures, and distinguishing between nonexistent trade-offs and undiscovered trade-offs is a non-trivial challenge when dealing with sparse MORL.

## 6 Conclusion

This paper examined how heterogeneous objective sparsity affects MORL. We argued that MORL problems with sparse rewards are more than an extension of single-objective sparse reward problems, and bring unique challenges in the discovery and interpretation of trade-offs among objectives. To characterize this issue, we introduced a taxonomy of reward sparsity, namely event-contingent, temporally delayed, threshold-activated, and measurement-limited sparsity. We then analyzed how these forms of sparsity can affect Pareto-front discovery through false Pareto dominance, objective masking under scalarization, conflicting sparsity-mitigation mechanisms, and ambiguity in incomplete observed fronts. Overall, our discussion brought forward the need for objective-specific approaches that diagnose and address sparsity at the level of individual reward components rather than applying a single global strategy to the entire reward vector.

## References


[Achiam *et al.*, 2017] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. Pmlr, 2017.

[Alegre *et al.*, 2023] Lucas N. Alegre, Ana L. C. Bazzan, Diederik M. Roijers, Ann Nowé, and Bruno C. da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 2003–2012, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[Andrychowicz *et al.*, 2017] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Arjona-Medina *et al.*, 2019] Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Auger *et al.*, 2009] Anne Auger, Johannes Bader, Dimo Brockhoff, and Eckart Zitzler. Theory of the hypervolume indicator: optimal -distributions and the choice of the reference point. In *Proceedings of the Tenth ACM SIGEVO Workshop on Foundations of Genetic Algorithms*, FOGA '09, page 87–102, New York, NY, USA, 2009. Association for Computing Machinery.

[Bellemare *et al.*, 2016] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1479–1487, Red Hook, NY, USA, 2016. Curran Associates Inc.

[Burda *et al.*, 2018] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[Castelletti *et al.*, 2002] Andrea Castelletti, Giorgio Corani, A Rizzolli, Rodolfo Soncinie-Sessa, and Enrico Weber. Reinforcement learning in the operational management of a water system. In *IFAC workshop on modeling and control in environmental issues*, pages 325–330. Keio University Yokohama, Japan, 2002.

[Daniel *et al.*, 2014] Christian Daniel, Malte Viering, Jan Metz, Oliver Kroemer, and Jan Peters. Active reward learning. In *Robotics: Science and systems*, volume 98, 2014.

[Florensa *et al.*, 2017] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*, pages 482–495. PMLR, 2017.

[García and Fernández, 2015] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.*, 16(1):1437–1480, January 2015.

[Hayes *et al.*, 2022] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1), April 2022.

[Hertweck *et al.*, 2020] Tim Hertweck, Martin Riedmiller, Michael Bloesch, Jost Tobias Springenberg, Noah Siegel, Markus Wulfmeier, Roland Hafner, and Nicolas Heess. Simple sensor intentions for exploration, 2020.

[Kaushik *et al.*, 2018] Rituraj Kaushik, Konstantinos Chatzilygeroudis, and Jean-Baptiste Mouret. Multi-objective model-based policy search for data-efficient learning with sparse rewards. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 839–855. PMLR, 29–31 Oct 2018.

[Ladosz *et al.*, 2022] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

[Liu *et al.*, 2024] Ruohong Liu, Yuxin Pan, Linjie Xu, Lei Song, Pengcheng You, Yize Chen, and Jiang Bian. C-morl: Multi-objective reinforcement learning through efficient discovery of pareto front, 10 2024.

[Liu *et al.*, 2025] Erlong Liu, Yu-Chang Wu, Xiaobin Huang, Chengrui Gao, Ren-Jian Wang, Ke Xue, and Chao Qian. Pareto set learning for multi-objective reinforcement learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025.

[Mannion *et al.*, 2017] Patrick Mannion, Sam Devlin, Karl Mason, Jim Duggan, and Enda Howley. Policy invariance under reward transformations for multi-objective reinforcement learning. *Neurocomputing*, 263:60–73, 2017. Multiobjective Reinforcement Learning: Theory and Applications.

[Mannion *et al.*, 2018] Patrick Mannion, Sam Devlin, and Enda Howley. Reward shaping for knowledge-based multi-objective multi-agent reinforcement learning. *The Knowledge Engineering Review*, 33, 12 2018.

[Natarajan and Tadepalli, 2005] Sriraam Natarajan and Prasad Tadepalli. Dynamic preferences in multi-criteria reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 601–608, 2005.

[Ng *et al.*, 1999] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

[Pathak *et al.*, 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 - Volume 70*, ICML'17, page 2778–2787. JMLR.org, 2017.

[Puterman, 2014] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Ren *et al.*, 2021] Zhizhou Ren, Ruihan Guo, Yuan Zhou, and Jian Peng. Learning long-term reward redistribution via randomized return decomposition. *arXiv preprint arXiv:2111.13485*, 2021.

[Riedmiller *et al.*, 2018] Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Van de Wiele, Volodymyr Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing - solving sparse reward tasks from scratch, 2018.

[Roijers *et al.*, 2013] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Int. Res.*, 48(1):67–113, October 2013.

[Sutton and Barto, 2018] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[Thakar *et al.*, 2026] Raghav Thakar, Gaurav Dixit, and Kagan Tumer. Post hoc extraction of pareto fronts for continuous control, 2026.

[Vamplew *et al.*, 2008] Peter Vamplew, John Yearwood, Richard Dazeley, and Adam Berry. On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In *Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, AI '08, page 372–378, Berlin, Heidelberg, 2008. Springer-Verlag.

[Vamplew *et al.*, 2011] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, 84:51–80, 07 2011.

[Van Moffaert and Nowé, 2014] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *J. Mach. Learn. Res.*, 15(1):3483–3512, January 2014.

[Van Moffaert *et al.*, 2013] Kristof Van Moffaert, Madalina M. Drugan, and Ann Nowé. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 191–199, April 2013.

[White, 1982] DJ White. Multi-objective infinite-horizon discounted markov decision processes. *Journal of Mathematical Analysis and Applications*, 89(2):639–647, 1982.

[Zitzler *et al.*, 2003] Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M Fonseca, and Viviane Grunert Da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on evolutionary computation*, 7(2):117–132, 2003.