Between Life and Death: Examining Sparse Reward Designs in Healthcare RL

Yuxuan Shi[†], **Matthew Lafrance**[†], **Shengpu Tang**
{yuxuan.shi,matthew.lafrance,shengpu.tang}@emory.edu

Department of Computing Science, Emory University

[†] equal contribution

Abstract

In reinforcement learning (RL) for healthcare, reward functions often encode clinical endpoints like survival and death. This results in a sparse reward structure with non-zero rewards only at terminal transitions. However, the exact numerical rewards assigned to survival and death vary in existing literature, raising concerns about whether they will end up optimizing for the same objective. In this work, we theoretically and empirically examine three common sparse reward designs: survival-only, death-only, and mixed. We prove that, under the assumptions of terminal-only rewards, guaranteed absorption, and no discounting, the corresponding value functions of the three designs have an equivalence relationship and lead to the same optimal policy. We verify these theoretical results in randomly generated MDPs and demonstrate how relaxing these assumption affect the equivalence relationship. Finally, we consider a more complex grid-world domain in which the assumptions are violated, where we found the survival-only and mixed designs consistently lead to better policies than the death-only design. Our findings provide important initial insights into the choices of sparse reward designs and how they shape policy learning in healthcare RL applications.

1 Introduction

In reinforcement learning (RL), the reward function defines the objective of the task and serves as the primary signal that guides the behavior and learning of agents (Sutton & Barto, 2018). For many healthcare applications of RL, particularly for critical care (Komorowski et al., 2018), rewards are often defined using clinical endpoints such as survival and death, as these are meaningful outcomes and are easy to extract from data (Gottesman et al., 2019). This leads to a sparse reward structure where a nonzero reward is assigned only upon trajectory termination, and all non-terminal transitions are assigned a zero reward.

Despite the intuitive goal of encouraging survival and avoiding death, the existing literature exhibits notable variability in how these two outcomes are encoded as rewards (Table 2 in Appendix A). Although almost all studies assign a positive reward for survival, they differ in how they handle death, where some use a negative reward, while others use zero. This inconsistency raises an important question: are these studies truly optimizing for the same objective and learning the same policy?

In this work, we analyze three sparse reward designs—survival-only, death-only, and mixed—to study their theoretical and empirical impact on value estimation and policy learning. We prove that, under a set of sufficient conditions including terminal-only sparse rewards, guaranteed absorption, and no discounting, the resulting value functions satisfy a set of linear equivalence relationships and yield identical optimal policies. We also show that each value function has a probabilistic interpretation with respect to the terminal outcomes. We then empirically validate our theoretical

propositions in procedurally generated directed acyclic graph (DAG)-based environments, where the assumptions are satisfied by design. By relaxing these assumptions, such as introducing discounting, intermediate rewards, or state-transition cycles, we observe how the value equivalences break down and how such violations affect policy learning. Lastly, we extend our analysis to a more complex grid world environment motivated by clinical problems, to examine the implications of reward design on policy behavior in a more interpretable setting. Importantly, we found that the policies learned under survival-only and mixed reward are equivalent, and consistently outperform those learned under the death-only reward.

2 Problem Setup

We consider Markov decision processes (MDPs) defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the transition model with $P(s'|a, s)$ specifying the probability of transitioning from state s to s' given action a, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \Delta(\mathbb{R})$ the reward function with $R(s, a, s')$ denoting the expected immediate reward obtained from taking action a in state s and transitioning to s', and discount factor $\gamma \in [0, 1]$. If the reward function only depends on the next state s', we denote it as $R(s')$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ defines the probability distribution of selecting action a in state s. By following policy π in MDP $\mathcal{M}$, we generate a trajectory $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$, for which the return is defined as the cumulative discounted reward, $G = \sum_{t=1}^{\infty} \gamma^t r_t$. The state-value function of policy π represents the expected return starting from state s and following π thereafter: $V^\pi(s) = \mathbb{E}_\pi[G|s_1 = s]$. The action-value function of policy π is defined by further restricting the action taken from the starting state: $Q^\pi(s, a) = \mathbb{E}_\pi[G|s_1 = s, a_1 = a]$.

Given a policy π, the MDP $\mathcal{M}$ reduces to a Markov reward process (MRP) (i.e., a Markov chain with rewards) $\mathcal{M}^\pi = (\mathcal{S}, P^\pi, R^\pi, \gamma)$, where $P^\pi(s'|s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[P(s'|s, a)]$ is the probability of transitioning to state s' from state s when actions are selected according to policy π. Likewise, $R^\pi(s, s') = \mathbb{E}_{a\sim\pi(\cdot|s)}[R(s, a, s')]$ denotes the expected immediate reward averaged over actions drawn from policy π. The state-value function $V(s)$ of the MRP is defined analogously to the MDP setting, but there is no action-value function. Conceptually, an MRP can be seen as an MDP in which a single action is available at each step, or when actions are marginalized according to policy π. When we need to refer to a generic MRP not induced by a policy, we use the notation $\mathcal{M}^0 = (\mathcal{S}, P, R, \gamma)$. While we are ultimately interested in MDPs, we use MRPs as an intermediate step, where the theoretical results can be directly extended to MDPs.

Sparse Reward Designs. We consider episodic MDPs with an indefinite horizon where we have two absorbing states, $\mathcal{S}_\infty = \{s_{\text{surv}}, s_{\text{death}}\} \subseteq \mathcal{S}$. All actions from an absorbing state $s_\infty \in \mathcal{S}_\infty$ leads to a transition back to the same state s_∞ with a reward of zero, i.e., $R(s_\infty, a, s_\infty) = 0$ for all $a \in \mathcal{A}$. We define a family of three MDPs $\mathcal{M}_+, \mathcal{M}_-, \mathcal{M}_\pm$, where $\mathcal{M}_\bullet = (\mathcal{S}, \mathcal{A}, P, R_\bullet, \gamma)$ for $\bullet \in \{+, -, \pm\}$. These MDPs differ only in their reward functions $R_\bullet$ (Table 1), while sharing the same state space, action space, transition dynamics, and discount factor. For each variant $\mathcal{M}_\bullet$, we define the corresponding value functions for policy π, denoted as $V^\pi_\bullet$ and $Q^\pi_\bullet$. Analogously, we define a family of MRPs $\mathcal{M}^0_+, \mathcal{M}^0_-, \mathcal{M}^0_\pm$ (Table 3). Our notation is summarized in Appendix B.

Name	Symbol	Reward for s_{surv}	Reward for s_{death}
Survival MDP	$\mathcal{M}_+$	$R_+(s_{\text{surv}}) = +1$	$R_+(s_{\text{death}}) = 0$
Death MDP	$\mathcal{M}_-$	$R_-(s_{\text{surv}}) = 0$	$R_-(s_{\text{death}}) = -1$
Mixed MDP	$\mathcal{M}_\pm$	$R_\pm(s_{\text{surv}}) = +1$	$R_\pm(s_{\text{death}}) = -1$

Table 1: Reward definitions for three terminal-state MDP variants used in this work.

3 Theoretical Analyses

In this section, we formalize the relationships between value functions under different reward designs in both MRPs and MDPs. Building on a set of assumptions, we derive several identities that connect the value functions resulting from the three different sparse reward designs, clarify the functional properties of value functions and their downstream influence on learned policies.

3.1 Constraints and Assumptions

In our analysis, we use the following assumptions to further constrain the structure of the MDP/MRP.

Assumption 1 (Terminal-only sparse rewards)**.** Non-terminal transitions all have reward of zero, i.e., $R(s') = 0$ for all $s' \in \mathcal{S} \setminus \mathcal{S}_\infty$.

Assumption 2 (Guaranteed absorption)**.** For the given MRP, or the given MDP under *any* policy π, the agent reaches a terminal state with probability 1 in a finite number of steps: $\Pr[s_T \in \mathcal{S}_\infty] = 1$ for some $T < \infty$.

Assumption 3 (No discounting)**.** $\gamma = 1$.

Remark. Together, these assumptions result in value functions that have a clean probabilistic interpretation. Assumption 1 is satisfied by all three reward variants $R_+, R_-, R_\pm$ described above. Assumption 2 ensures that from any non-terminal state, the trajectory will reach a terminal state (either survival or death, and receive the corresponding reward) with probability one in a finite number of steps. This excludes settings with cycles that can lead to infinite loops (for certain policies). Thus, with Assumption 3, the undiscounted value of a non-terminal state directly encodes the probability of eventually reaching either terminal states. For example, in the survival-only MRP, the value of each state corresponds to the probability of reaching survival, whereas in the death-only MRP, the value of each state corresponds to the negative probability of reaching death.

3.2 Sum of Value Functions in MRP and MDP

Given the reward definitions in Section 2, we note that $R_+(s) + R_-(s) = R_\pm(s)$ for all states. As a result, we can show the value functions corresponding to these reward functions satisfy an additive identity for both MRPs and MDPs.

Proposition 1 (Value sum in MRPs)**.** Given $\mathcal{M}^0_+$, $\mathcal{M}^0_-$, and $\mathcal{M}^0_\pm$, if Assumption 1 holds, then $V_+(s) + V_-(s) = V_\pm(s)$ for all $s \in \mathcal{S}$.

Proposition 2 (Value sum in MDPs)**.** Given $\mathcal{M}_+$, $\mathcal{M}_-$, and $\mathcal{M}_\pm$ and a policy π, if Assumption 1 holds, then $V^\pi_+(s) + V^\pi_-(s) = V^\pi_\pm(s)$ for all $s \in \mathcal{S}$, and $Q^\pi_+(s,a) + Q^\pi_-(s,a) = Q^\pi_\pm(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$.

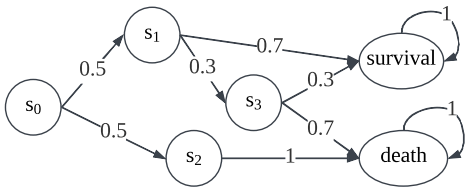
Note that Propositions 1 and 2 do not require Assumptions 2 and 3 to hold. The identities follow directly from the linearity of value functions with respect to reward: under a fixed policy π (in MDPs, and in MRPs which are equivalent to MDPs with a fixed policy), the same trajectory distribution is induced across all three variants, and the total return decomposes additively.

3.3 MRP: Relationship between V_+ and V_-

Proposition 3 (Value difference in MRPs)**.** Given $\mathcal{M}^0_+$ and $\mathcal{M}^0_-$, if Assumptions 1, 2 and 3 hold, then $V_+(s) - V_-(s) = 1$ for all $s \in \mathcal{S} \setminus \mathcal{S}_\infty$.

Since the value functions can be interpreted as probabilities under our assumptions, Proposition 1 implies that the probability of survival ($V_+(s)$) and the probability of death ($-V_-(s)$) sum to 1, which is intuitively true if the system does not have cycles. A formal inductive proof is provided in Appendix C.1. All three assumptions contributed to the inductive argument; thus, the relationship might no longer hold if these assumptions are relaxed (see experiments in Section 4). To illustrate the value function identities in MRP (Propositions 1 and 3), we present the following worked example.

Example 1. As shown in Figure 1-left, the MRP contains four non-terminal states $\{s_0, s_1, s_2, s_3\}$ and two absorbing terminal states $\{s_{\text{surv}}, s_{\text{death}}\}$, with the arrows between states showing the transition probabilities $P(s'|s)$. This MRP satisfies Assumption 2 as the state transition diagram forms a DAG such that all four non-terminal states are guaranteed to reach either survival or death in a finite number of steps. The value functions $V_+, V_-, V_\pm$ corresponding to the three reward function variants $R_+, R_-, R_\pm$ (Assumption 1) with $\gamma = 1$ (Assumption 3) are shown in Figure 1-right. One may verify that both Propositions 1 and 3 hold.



State	$V_+(s)$	$V_-(s)$	Diff	Sum / $V_\pm(s)$
s_0	0.395	−0.605	1.00	−0.21
s_1	0.79	−0.21	1.00	0.58
s_2	0.00	−1.00	1.00	−1.00
s_3	0.30	−0.70	1.00	−0.40

Figure 1: Left - MRP state transition graph. Right - state values under $\mathcal{M}_+^0$, $\mathcal{M}_-^0$, $\mathcal{M}_\pm^0$.

3.4 MDP: Relationships Between V_+^π, V_-^π and Q_+^π, Q_-^π

Here, we extend the MRP value function identities to MDPs with a *fixed* policy π.

Proposition 4 (State-value difference in MDPs with shared policy)**.** Given $\mathcal{M}_+$ and $\mathcal{M}_-$ and a policy π, if Assumptions 1, 2 and 3 hold, then $V_+^\pi(s) - V_-^\pi(s) = 1$ for all $s \in \mathcal{S} \setminus \mathcal{S}_\infty$.

Proposition 5 (Action-value difference in MDPs with shared policy)**.** Given $\mathcal{M}_+$ and $\mathcal{M}_-$ and a policy π, if Assumptions 1, 2 and 3 hold, then $Q_+^\pi(s,a) - Q_-^\pi(s,a) = 1$ for all $s \in \mathcal{S} \setminus \mathcal{S}_\infty,\ a \in \mathcal{A}$.

We can show Propositions 4 and 5 using the same induction argument of Proposition 3 (full proof in Appendix C.2). Alternatively, we can view the MDP $\mathcal{M}$ with a fixed policy π to be an MRP $\mathcal{M}^\pi$, thereby directly extending the result of Proposition 3 to Proposition 4. For Proposition 5, similar reasoning applies if we view the system as an MRP on a different "state space" of $\mathcal{S} \times \mathcal{A}$. See Figure 2 below for a worked example.

3.5 MDP: Policy Learning under $\mathcal{M}_+, \mathcal{M}_-, \mathcal{M}_\pm$

We now consider the implications of our value function identities on policy learning.

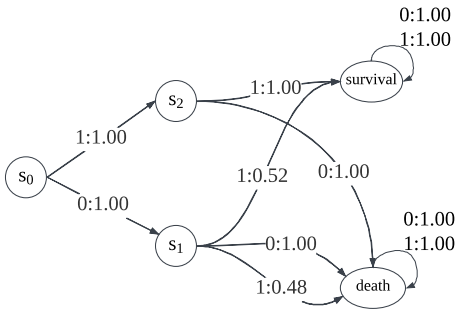
Proposition 6 (Optimality equivalence under $\mathcal{M}_+, \mathcal{M}_-$ and $\mathcal{M}_\pm$)**.** If Assumption 1, 2 and 3 hold, the optimal policies under $\mathcal{M}_+$, $\mathcal{M}_-$, and $\mathcal{M}_\pm$ are identical: $\pi_\pm^* = \pi_+^* = \pi_-^*$.

The equivalence between π_+^* and π_-^* follows directly from Propositions 4 and 5 where for any policy π, we have $V_+^\pi(s) = V_-^\pi(s) + 1$ for all $s \in \mathcal{S} \setminus \mathcal{S}_\infty$, meaning V_+^π is a positive affine transformation of V_-^π. Given a pair of policies π_1, π_2 that satisfies $V_-^{\pi_1}(s) \le V_-^{\pi_2}(s)$ for all s, we have the relationship $V_+^{\pi_1}(s) \le V_+^{\pi_2}(s)$ for all s. This implies that the relative ranking of any policy pair remains the same under $\mathcal{M}_+$ and $\mathcal{M}_-$, and thus the optimal policy is the same.

For $\pi_\pm^*$, Proposition 2 implies that for any policy π, the action-values satisfy $V_\pm^\pi(s) = V_+^\pi(s) + V_-^\pi(s)$. Given the identity $V_+^\pi(s) - V_-^\pi(s) = 1$ from Proposition 4, it follows that if Assumptions 1, 2 and 3 hold, $V_\pm^\pi(s) = 2V_+^\pi(s) - 1 = 2V_-^\pi(s) + 1$ for all $s \in \mathcal{S} \setminus \mathcal{S}_\infty$. Since $V_\pm^\pi$ is also a positive affine transformation of V_-^π, the same reasoning as above leads to the conclusion that the optimal policy is also the same.

In Example 2, one may verify the learned optimal policies across all three variants are identical (always selecting a_1), confirming the optimality equivalence when all three assumptions hold. However, when any of the three assumptions are relaxed, we can no longer guarantee the equivalence of optimal policies. In our empirical experiments, we investigate such settings to examine how these theoretical guarantees break down in practice and to assess the extent to which these assumptions constraints policy learning behavior.

Example 2. As shown in Figure 2 (a), the MDP contains three non-terminal states s_0, s_1, s_2 and two absorbing terminal states $s_{\text{surv}}, s_{\text{death}}$. The arrows indicate transition probabilities $P(s'|s, a)$ under each action, with each edge labeled in the format (action: probability), For instance, the outgoing edge labeled $1 : 0.52$ from s_1 represents taking action a_1 and transitioning to the corresponding next state with probability 0.52. We evaluate the MDP under two policies: a uniform random policy and an optimal policy, computing the action-value functions Q^π_+ and Q^π_- using the reward variants R_+, R_- and $R_\pm$ (Assumption 1) with $\gamma = 1$ (Assumption 3), and due to the DAG structure, Assumption 2 is also satisfied. As shown in Figure 2 (b–c), Propositions 2 and 5 holds exactly for all state-action pairs under both evaluation settings.



(b)

State-Action	$Q^\pi_+(s,a)$	$Q^\pi_-(s,a)$	Diff	Sum / $Q^\pi_\pm(s,a)$
(s_0, a_0)	0.26	-0.74	1.00	-0.48
(s_0, a_1)	0.50	-0.50	1.00	0.00
(s_1, a_0)	0.00	-1.00	1.00	-1.00
(s_1, a_1)	0.52	-0.48	1.00	0.04
(s_2, a_0)	0.00	-1.00	1.00	-1.00
(s_2, a_1)	1.00	0.00	1.00	1.00

(c)

State-Action	$Q^*_+(s,a)$	$Q^*_-(s,a)$	Diff	Sum / $Q^*_\pm(s,a)$
(s_0, a_0)	0.52	-0.48	1.00	0.04
(s_0, a_1)	1.00	0.00	1.00	1.00
(s_1, a_0)	0.00	-1.00	1.00	-1.00
(s_1, a_1)	0.52	-0.48	1.00	0.04
(s_2, a_0)	0.00	-1.00	1.00	-1.00
(s_2, a_1)	1.00	0.00	1.00	1.00

Figure 2: MDP toy example with labeled action-transition probabilities (a), Q-values under uniform random policy (b), and Q-values under optimal policy (c).

4 Empirical Results

To understand the extent to which our theoretical insights apply in various scenarios, we use a series of synthetic environments. Specifically, we first use procedurally generated DAG MDPs to verify our theorems when all the assumptions are satisfied, and then explore how breaking each assumption affects our conclusions. Finally, we consider a previously studied grid world environment, which represents a more complex setting and a style of domains typically used in benchmarks.

4.1 DAG MDPs

Setup. We consider stochastic MDPs for which the state transition diagram forms a DAG, like the one in Example 2. The DAGs are procedurally generated based on a maximum number of states and actions, as well as termination probabilities of each state transitioning to the two absorbing states. Seeds are used for reproducibility. By default, all three assumptions are satisfied.

Evaluation. Given an MDP's state transition DAG, we compute value functions (V and Q) under each reward setting. We consider two different experimental conditions; policy evaluation and policy learning. For policy evaluation, we assume a policy is provided (in our case, the uniformly random policy), and we evaluate the given policy to get Q^π_+, Q^π_-, and $Q^\pi_\pm$. For policy learning, we run the policy iteration algorithm separately on $\mathcal{M}_+$, $\mathcal{M}_-$, and $\mathcal{M}_\pm$, and obtain the resulting Q^*_+, Q^*_-, and $Q^*_\pm$. We visualize the relationship between Q_+ and Q_- using scatter plots where each point is a state-action pair, and if our conclusions hold, all the points would be on a straight line of $Q_+ - Q_- = 1$. We also compare $Q_+ + Q_-$ with the corresponding $Q_\pm$. Finally, we compare the learned optimal policies π^*_+, π^*_- and $\pi^*_\pm$ to see if they are identical.

Results. We start with the ideal setting where all assumptions are satisfied, and then break each assumption in turn to see how they affect our theoretical conclusions.

Ideal Setting. In the ideal case, we evaluate a 200-state DAG MDP and find that Proposition 2 and Proposition 5 hold across all state-action pairs, as shown in Figure 3. In the Q_+ vs. Q_- plots, the dotted line represents the identity $Q_+ - Q_- = 1$; in the $Q_+ - Q_-$ vs. $Q_\pm$ plots, the dotted line represents the identity $Q_+ + Q_- = Q_\pm$. All points lie precisely on the dotted lines. Note also that π^*_+ is the same as π^*_-. This aligns with our expectations, as the relative ordering for optimal states in both policies should be the same. $\pi^*_\pm$ is also exactly the same between the other two policies.

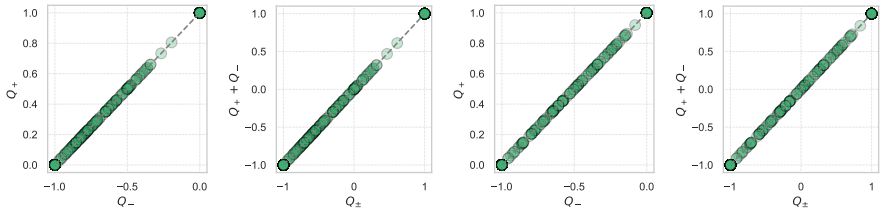


Figure 3: Q function comparisons on the 200-state MDP. Left two: Q^π for a uniform random policy. Right two: Q^* for the optimal policy from policy iteration.

Breaking Assumption 1. To break the assumption of sparse rewards, we added small positive intermediate rewards for non-terminal transitions. We use the same large MDP transition graph as Figure 3, while adding an intermediate reward of 0.1 to each non-terminal transition. Figure 4 shows the resulting Q function comparisons. Surprisingly, Proposition 5 still holds, as all points still lie on the diagonal line of $Q_+ - Q_- = 1$. This is because the same symmetric intermediate reward is added to both $\mathcal{M}_+$ and $\mathcal{M}_-$, causing the additional reward to cancel out when computing the difference. However, some points exceed the range of values $[0, 1]$ and $[-1, 0]$ for Q^π_+ and Q^π_- respectively, indicating that the Q-values can no longer be interpreted as as a probability. Notably, π^*_+ is still exactly the same as π^*_- for all states. This is expected, as Proposition 5 still holds, which guarantees identical policy improvement steps and thus identical optimal policies by Proposition 6. Interestingly, even though the identity $Q_+ + Q_- = Q_\pm$ no longer holds and the points deviate from the dotted line, we observe that $\pi_\pm$ is also identical to π_+ and π_-.

Breaking Assumption 2. As shown in Figure 5-left, we create a third absorbing state s_3 to the MDP from Figure 2 which acts as a pseudo-terminal state, such that an agent at s_3 will never reach a true terminal state that receives reward. The reward function remains the same, where transitions to survival or death have the same reward as before, and transitioning to s_3 has reward 0. Figure 5-right shows the resulting Q function comparisons, where certain state-action pairs no longer fall on the dotted line, such that the conclusion in Proposition 5 no longer holds. Note that π^*_+ and π^*_- differ for one state for this MDP, where state s_0 has a different optimal action between policies, with π^*_+ choosing action 0 and π^*_- choosing action a_1. $\pi^*_\pm$, however is exactly equal to π^*_+. Given this, we find that of the three policies, π^*_+ and $\pi^*_\pm$ both have a probability of 0.52 of transitioning to survival, whereas π^*_- has a probability of 0.5.

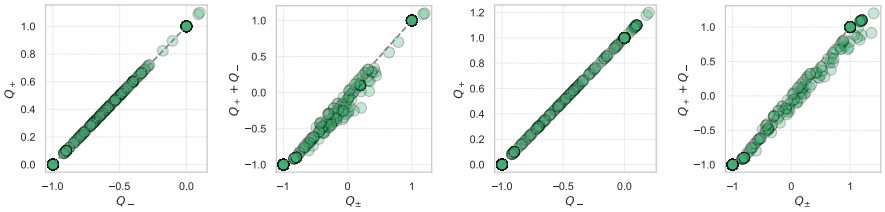


Figure 4: Q function comparison with small non-terminal rewards on the 200-state DAG-MDP. Left two: Q^π for a uniform random policy. Right two: Q^* for the optimal policy from policy iteration.

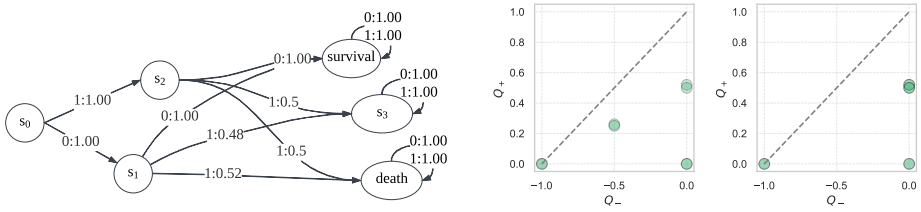


Figure 5: Left: Toy MDP without guaranteed absorption. Right: Q function comparison of this MDP, under a uniform random policy (left) and the optimal policy from policy iteration (right).

Breaking Assumption 3. Here, we change the discount factor γ from 1.0 to $0.9, 0.8, 0.7$ using the same large MDP from Figure 3. Figure 6 depict the resulting Q graphs, in which many points fall below the dotted line, meaning their difference is less than 1, directly due to the discounting of future rewards. Note that many points still fall on the dotted line. These points correspond to the state-action pairs that deterministically transition to a terminal state, as those transitions are not being discounted.

Overall, a smaller γ leads to greater deviation from the identity in Proposition 5, resulting in a monotonic increase in the number of mismatches between π^*_+ and π^*_- as γ decreases. For instance, the two policies differ in only one state when $\gamma = 0.9$ and 0.8, but in three states when $\gamma = 0.7$. In contrast, when comparing either π^*_+ or π^*_- with the mixed policy $\pi^*_\pm$, the behavior becomes unclear: the number of mismatches does not consistently increase or decrease as γ decreases.

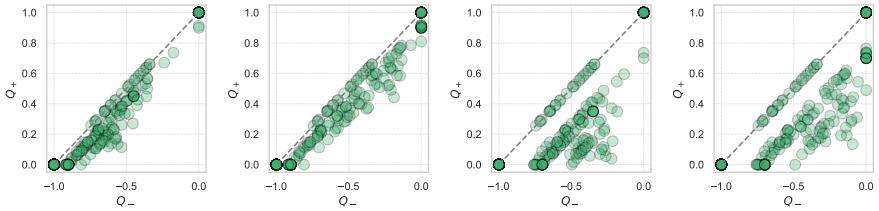


Figure 6: Q function comparison with discounting under a uniform random policy (left) and the optimal policy from policy iteration (right) for different discount (0.9 and 0.7).

4.2 LifeGate Grid World

Next, we examine a more complicated setup based on the LifeGate example used in Fatemi et al. (2021) with a 10×10 gridworld. As shown in Figure 7-left, the agent can move in four cardinal directions, with rewards only assigned at two types of terminal states: survival (blue) and death (red). Pink squares indicate dead-end states from which survival is no longer possible. Grey squares denote out-of-bounds areas. We set the discount factor $\gamma = 0.9$, as using $\gamma = 1$ leads to non-convergence.

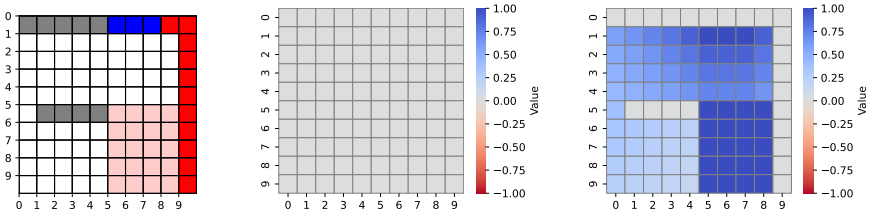


Figure 7: Gridworld (left), heatmap of $V_+(s) + V_-(s) - V_\pm(s)$ (middle), and $V_+(s) - V_-(s)$ (right).

Note that under this setting, only Assumption 1 is satisfied. As a result, we do not expect Propositions 3 to 6 to hold, whereas Propositions 1 and 2 to hold. As shown in Figure 7, we observe that $V_+ - V_- \neq 1$, as indicated by the light blue values in the right heat map for non terminal entries.

However, we do find that $V_+(s) + V_-(s) = V_\pm(s)$, since all values in the middle heatmap are zero. This aligns with our theoretical expectations.

Next, we analyze differences in π^* under $\mathcal{M}_+$, $\mathcal{M}_-$, and $\mathcal{M}_\pm$. As shown in Figure 8, π^*_+ is strictly better than π^*_- in the sense that, for each state, the optimal action(s) under π^*_+ form a subset of those under π^*_-, indicating a more decisive policy. This observation provides intuition for the empirical equivalence between π^*_+ and $\pi^*_\pm$. Conceptually, adding negative reward structure to $\mathcal{M}_+$ does not introduce additional actionable information: incentivizing survival inherently entails avoiding death, given that survival and death are mutually exclusive outcomes. In contrast, $\mathcal{M}_-$ lacks a survival incentive, so the agent is encouraged only to delay death rather than to achieve recovery. While it may still prefer safer paths, the absence of recovery reward can lead to indecision among multiple "non-deadly" routes. Thus, $\mathcal{M}_+$ and $\mathcal{M}_\pm$ encode equivalent decision making objectives, resulting in the identical optimal policy, while $\mathcal{M}_-$ leads to a less directed policy.

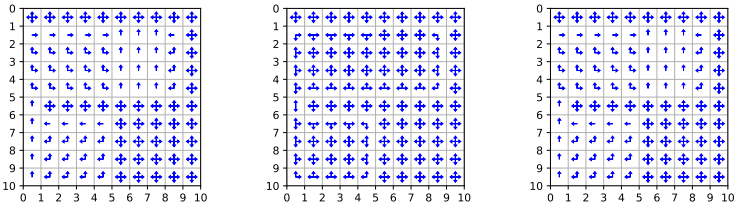


Figure 8: Diagrams of optimal actions of π^*_+ (left), π^*_- (middle), and $\pi^*_\pm$ (right).

5 Discussion and Conclusion

During the empirical experiments, we discover that under Assumption 1, 2 and 3, $\mathcal{M}_+$ and $\mathcal{M}_-$ yield equivalent policies. However, once the assumptions are relaxed, this equivalence break down. In our gridworld experiments, policies learned under $\mathcal{M}_+$ outperform those from $\mathcal{M}_-$, and is equivalent to that of $\mathcal{M}_\pm$.

Secondly, we discover that under Assumption 1, 2 and 3, value functions act as probabilistic indicators for reaching their respective terminal states: $V_+(s)$ equals the probability of survival, while $V_-(s)$ equals the negative probability of death. Together, $V_\pm(s)$ faithfully reflects the net directional risk of a state, as it encodes the expected outcome for both survival and death.

On the other hand, our empirical results show that as these assumptions are violated, this interpretability deteriorates. The value functions no longer maintain the probabilistic meaning, and $V_\pm(s)$ begins to exhibit bias, favoring survival or death which is inconsistent with the true risk. This distortion can lead to severe consequences in real-world applications, especially in clinical settings, misrepresentation of risk levels may cause inappropriate decision making.

Additionally, our experiments reveal a subtle phenomenon: in some cases, the value function identity continues to hold empirically even when certain assumptions are violated. In real-world clinical applications, value function–particularly $V(s)$ tables–are often interpreted as proxies for risk or prognosis. However, the complexity of practical systems may obscure underlying distortions when assumptions like guaranteed absorption or no discounting are violated.

We also find that the sufficient conditions ensuring the value identity vary in strength. Among them, Assumption 3 proves to be the most strict: even a slight deviation such as $\gamma = 0.99$ breaks the identity, thereby undermining the interpretability. Secondly, for Assumption 2, our results show that introducing a third absorbing terminal state breaks the identity. In practical environments that combine episode length constraints with cycles, whether escapable or not, can replicate a similar effect. Finally, the Assumption 1 appears to be somewhat more flexible. When intermediate rewards are symmetrically balanced between outcomes, the identity often remains intact; however, introducing asymmetric intermediate rewards (i.e., penalties only for entering dead-end states or rewards discounted by distance to terminal outcomes in an unbalanced DAG) disrupts the identity.

Data and Code Availability

Our code to replicate the analyses and figures presented in this paper, including our implementation of the LifeGate environment and DAG-based MRP/MDP generators, is available at https://github.com/ROLFFFX/Examining-Sparse-Reward-Designs-in-Healthcare-RL.

References

Kartik Choudhary, Dhawal Gupta, and Philip S. Thomas. ICU-Sepsis: A benchmark MDP built from real medical data. *Reinforcement Learning Journal*, 4:1546–1566, 2024.

Mehdi Fatemi, Taylor W. Killian, Jayakumar Subramanian, and Marzyeh Ghassemi. Medical dead-ends and learning to identify high-risk states and treatments. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=4CRpaV4pYp.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019. DOI: 10.1038/s41591-018-0310-5. URL https://doi.org/10.1038/s41591-018-0310-5.

Russell Jeter, Christopher Josef, Supreeth Shashikumar, and Shamim Nemati. Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care? *arXiv preprint arXiv:1902.03271*, 2019. URL https://arxiv.org/abs/1902.03271.

Christina X Ji, Michael Oberst, Sanjat Kanjilal, and David Sontag. Trajectory inspection: A method for iterative clinician-driven design of reinforcement learning studies. *AMIA Summits on Translational Science Proceedings*, 2021:305, 2021.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. ISSN 1546-170X. DOI: 10.1038/s41591-018-0213-5. URL https://doi.org/10.1038/s41591-018-0213-5.

Dayang Liang, Huiyi Deng, and Yunlong Liu. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach. *Applied Intelligence*, 53(9):11034–11044, 2023.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *NeurIPS workshop on Machine Learning for Health (ML4H)*, 2017a. URL https://arxiv.org/abs/1711.09602.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68, pp. 147–163. PMLR, 2017b. URL https://proceedings.mlr.press/v68/raghu17a.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Shengpu Tang, Aditya Modi, Michael W. Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: reinforcement learning with near-optimal set-valued policies. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Rui Tu, Zhipeng Luo, Chuanliang Pan, Zhong Wang, Jie Su, Yu Zhang, and Yifan Wang. Offline safe reinforcement learning for sepsis treatment: Tackling variable-length episodes with sparse rewards. *Human-Centric Intelligent Systems*, 5(1):63–76, 2025.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Summary of reward designs in recent healthcare RL studies

Table 2: Summary of reward designs in recent healthcare RL studies.

Reference	Survival Reward	Death Penalty	Intermediate Rewards?
Komorowski et al. (2018)	$+100$	-100	No
Tang et al. (2020) (Frozen Lake)	$+1$	0	Yes
Tang et al. (2020) (Sepsis)	$+100$	-100	No
Ji et al. (2021)	$+100$	-100	No
Liang et al. (2023)	$+15$	-15	Yes
Choudhary et al. (2024)	$+1$	0	No
Tu et al. (2025)	$+1$	-1	Yes
Fatemi et al. (2021) (LifeGate)	$\{+1, 0\}$	$\{0, -1\}$	No
Fatemi et al. (2021) (Sepsis)	$\{+1, 0\}$	$\{0, -1\}$	No
Jeter et al. (2019)	$+100$	100	No
Raghu et al. (2017a)	$+1$	-1	No
Raghu et al. (2017b)	$+15$	-15	No

B Dictionary

- MRP: $\mathcal{M}^0$
- MRP variants: $\mathcal{M}^0_+$, $\mathcal{M}^0_-$, $\mathcal{M}^0_\pm$
- MDP: $\mathcal{M}$
- MDP variants: $\mathcal{M}_+$, $\mathcal{M}_-$, $\mathcal{M}_\pm$
- State-Value Function for MRP: V_+, V_-
- State-Value Function for MDP: V^π_+, V^π_-
- State-Action Value Function for MDP: Q^π_+, Q^π_-
- Terminal Absorbing states: $\mathcal{S}_\infty$
- Non-terminal states: $\mathcal{S} \setminus \mathcal{S}_\infty$
- Next State is Terminal: $s' \in \mathcal{S}_\infty$
- Next State is Non-Terminal: $s' \in \mathcal{S} \setminus \mathcal{S}_\infty$

C Proofs

C.1 Relationship between $V_+(s)$ and $V_-(s)$ in MRP

We now provide a formal proof that, under the assumptions mentioned above, the value functions of MRPs $\mathcal{M}_+^0$ and $\mathcal{M}_-^0$ satisfy the identity:

$$V_+(s) - V_-(s) = 1 \quad \text{for all } s \in \mathcal{S} \setminus \{s_{\text{survival}}, s_{\text{death}}\}$$

Setup. Let $(\mathcal{S}, P, R_+)$ and $(\mathcal{S}, P, R_-)$ define two MRPs with shared state space $\mathcal{S}$, transition probabilities $P(s'|s)$, and no discounting. Their only difference lies in the reward functions Assumption 1:

$$R_+(s,s') = \begin{cases} 1 & \text{if } s' = s_{\text{survival}}, \\ 0 & \text{otherwise} \end{cases}, \quad R_-(s,s') = \begin{cases} -1 & \text{if } s' = s_{\text{death}}, \\ 0 & \text{otherwise} \end{cases}$$

Together, it gives:

$$R_+(s,s') - R_-(s,s') = \begin{cases} 1 & \text{if } s' \in \mathcal{S}_\infty, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Then the Bellman equations are:

$$V_+(s) = \sum_{s'} P(s'|s) \left[R_+(s,s') + V_+(s')\right], \quad V_-(s) = \sum_{s'} P(s'|s) \left[R_-(s,s') + V_-(s')\right].$$

Subtracting yields:

$$V_+(s) - V_-(s) = \sum_{s'} P(s'|s) \left[R_+(s,s') - R_-(s,s') + (V_+(s') - V_-(s'))\right].$$

Inductive Proof. We prove that $V_+(s) - V_-(s) = 1$ for all non-terminal states s on the number of steps to absorption.

Base Case: Consider a terminating state s such that all transitions from s lead directly to terminal states. Since terminal states are absorbing, $V_+(s') = V_-(s') = 0$ for terminal s'. Then by the Bellman equations:

$$V_+(s) = \sum_{s'} P(s'|s) R_+(s,s'), \quad V_-(s) = \sum_{s'} P(s'|s) R_-(s,s').$$

Subtracting:

$$V_+(s) - V_-(s) = \sum_{s'} P(s'|s) \left(R_+(s,s') - R_-(s,s')\right).$$

By definition, $R_+(s,s') - R_-(s,s') = 1$ when s' is terminal. Moreover, by Assumption 2, we know that the agent must transition to an absorbing state with probability 1. Therefore, the difference between value functions is proven to be 1:

$$V_+(s) - V_-(s) = \sum_{s'} P(s'|s) \cdot 1 = 1.$$

Inductive Step: Assume that for all successor states s' of a non-terminal state s, we have $V_+(s') - V_-(s') = 1$. Then:

$$V_+(s) - V_-(s) = \sum_{s'} P(s'|s) \left[R_+(s,s') - R_-(s,s') + (V_+(s') - V_-(s')) \right].$$

Partition the sum over s' into terminal and non-terminal states and apply (1) yields:

$$\begin{aligned} V_+(s) - V_-(s) = &\sum_{s' \in \mathcal{S}_\infty} P(s'|s) \left[1 + (V_+(s') - V_-(s')) \right] \\ &+ \sum_{s' \notin \mathcal{S}_\infty} P(s'|s) \left[V_+(s') - V_-(s') \right]. \end{aligned}$$

Apply the inductive hypothesis:

$$V_+(s) - V_-(s) = \sum_{s' \in \mathcal{S}_\infty} P(s'|s)(1+0) + \sum_{s' \notin \mathcal{S}_\infty} P(s'|s)(0+1).$$

This simplifies to:

$$V_+(s) - V_-(s) = \sum_{s' \in \mathcal{S}_\infty} P(s'|s) + \sum_{s' \notin \mathcal{S}_\infty} P(s'|s).$$

Combining the two sums gives:

$$V_+(s) - V_-(s) = \sum_{s'} P(s'|s) = 1.$$

C.2 Relationship between $Q_+^\pi(s,a)$ and $Q_-^\pi(s,a)$ in MDP

We now provide a formal proof that, under the assumptions mentioned earlier, the action-value functions of $\mathcal{M}_+$ and $\mathcal{M}_-$ satisfy the identity:

$$Q_+^\pi(s,a) - Q_-^\pi(s,a) = 1 \quad \text{for all } (s,a) \in \mathcal{S} \setminus \mathcal{S}_\infty \times \mathcal{A}.$$

Setup. Let $(\mathcal{S}, \mathcal{A}, P, R_+)$ and $(\mathcal{S}, \mathcal{A}, P, R_-)$ define two MDPs that share the same state and action spaces, as well as transition model $P(s'|s,a)$. The only difference lies in the reward functions:

$$R_+(s,a,s') = \begin{cases} 1 & \text{if } s' = s_{\text{survival}} \\ 0 & \text{otherwise} \end{cases}, \quad R_-(s,a,s') = \begin{cases} -1 & \text{if } s' = s_{\text{death}} \\ 0 & \text{otherwise} \end{cases}$$

We fix a policy π, and define the corresponding action-value functions:

$$Q_+^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_+(s_t, a_t, s_{t+1}) \,\middle|\, s_0 = s, a_0 = a \right],$$

$$Q_-^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_-(s_t, a_t, s_{t+1}) \,\middle|\, s_0 = s, a_0 = a \right].$$

By assumption, the MDP is episodic with absorbing terminal states, sparse rewards, and no discounting ($\gamma = 1$).

Bellman Equations. The Bellman equations for action-value functions under policy π are:

$$Q_+^\pi(s,a) = \sum_{s'} P(s'|s,a)\left[R_+(s,a,s') + V_\pi^+(s')\right], \tag{1}$$

$$Q_-^\pi(s,a) = \sum_{s'} P(s'|s,a)\left[R_-(s,a,s') + V_\pi^-(s')\right]. \tag{2}$$

We partition each sum by terminal and non-terminal states:

$$Q_+^\pi(s,a) = \sum_{s'\in\mathcal{S}_\infty} P(s'|s,a)R_+(s,a,s') + \sum_{s'\in\mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a)V_\pi^+(s'),$$

$$Q_-^\pi(s,a) = \sum_{s'\in\mathcal{S}_\infty} P(s'|s,a)R_-(s,a,s') + \sum_{s'\in\mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a)V_\pi^-(s').$$

Reward Definitions. We use the reward specifications from our setup:

$$R_+(s,a,s_{\text{survival}}) = 1, \quad R_+(s,a,s_{\text{death}}) = 0,$$

$$R_-(s,a,s_{\text{survival}}) = 0, \quad R_-(s,a,s_{\text{death}}) = -1.$$

Thus, we simplify:

$$Q_+^\pi(s,a) = P(s_{\text{survival}}|s,a) + \sum_{s'\in\mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a)V_\pi^+(s'), \tag{3}$$

$$Q_-^\pi(s,a) = -P(s_{\text{death}}|s,a) + \sum_{s'\in\mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a)V_\pi^-(s'). \tag{4}$$

Base Case. Suppose $s' \in \mathcal{S}\setminus\mathcal{S}_\infty$ is one step away from a terminal state, i.e., $P(s' \in \mathcal{S}_\infty \mid s,a) = 1$. Then all transitions from s under action a lead to terminal states. Since terminal states are absorbing, $V_\pi^+(s') = V_\pi^-(s') = 0$ for all $s' \in \mathcal{S}_\infty$.

Applying (1) and (2):

$$Q_+^\pi(s,a) = \sum_{s'\in\mathcal{S}_\infty} P(s'|s,a)R_+(s,a,s'),$$

$$Q_-^\pi(s,a) = \sum_{s'\in\mathcal{S}_\infty} P(s'|s,a)R_-(s,a,s').$$

Then,

$$Q_+^\pi(s,a) - Q_-^\pi(s,a) = \sum_{s'\in\mathcal{S}_\infty} P(s'|s,a)\left[R_+(s,a,s') - R_-(s,a,s')\right].$$

Using the reward difference:

$$R_+(s,a,s') - R_-(s,a,s') = \begin{cases} 1, & \text{if } s' \in \mathcal{S}_\infty, \\ 0, & \text{otherwise.} \end{cases}$$

So:

$$Q_+^\pi(s,a) - Q_-^\pi(s,a) = \sum_{s'\in\mathcal{S}_\infty} P(s'|s,a) = 1. \qquad \text{(by Assumption 2)}$$

Inductive Step. Assume the identity holds for all successor states of s, and we now prove it for s. Using (3) and (4), we compute the difference:

$$\begin{aligned} Q_+^\pi(s,a) - Q_-^\pi(s,a) &= P(s_{\text{survival}}|s,a) + \sum_{s' \in \mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a) V_\pi^+(s') \\ &\quad + P(s_{\text{death}}|s,a) - \sum_{s' \in \mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a) V_\pi^-(s') \\ &= P(s_{\text{survival}}|s,a) + P(s_{\text{death}}|s,a) \\ &\quad + \sum_{s' \in \mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a) \left(V_\pi^+(s') - V_\pi^-(s')\right). \end{aligned}$$

Applying Proposition 4:

$$\sum_{s' \in \mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a) \left(V_\pi^+(s') - V_\pi^-(s')\right) = \sum_{s' \in \mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a).$$

Also, from Assumption 2:

$$P(s_{\text{survival}}|s,a) + P(s_{\text{death}}|s,a) + \sum_{s' \in \mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a) = 1.$$

Putting all together:

$$Q_+^\pi(s,a) - Q_-^\pi(s,a) = P(s_{\text{survival}}|s,a) + P(s_{\text{death}}|s,a) + \sum_{s' \in \mathcal{S}\setminus\mathcal{S}_\infty} P(s'|s,a) = 1.$$

C.3 Reward definitions for three terminal-state MRP variants used in this work.

Name	Symbol	Reward for s_{surv}	Reward for s_{death}
Survival MRP	$\mathcal{M}_+^0$	$R_+(s_{\text{surv}}) = +1$	$R_+(s_{\text{death}}) = \ \ 0$
Death MRP	$\mathcal{M}_-^0$	$R_-(s_{\text{surv}}) = \ \ 0$	$R_-(s_{\text{death}}) = -1$
Mixed MRP	$\mathcal{M}_\pm^0$	$R_\pm(s_{\text{surv}}) = +1$	$R_\pm(s_{\text{death}}) = -1$

Table 3: Reward definitions for three terminal-state MRP variants used in this work.