Reconciling Set-Valued Policy & Dead-End Discovery in RL: An Empirical Analysis

Sixing Wu[1], **Shengpu Tang**[1]
{sixing.wu, shengpu.tang}@emory.edu

[1]**Department of Computing Science, Emory University, United States**

Abstract

Standard reinforcement learning aims to learn a policy that maps each state to a single optimal action. Recent work has proposed alternative formulations inspired by healthcare applications, including Set-Valued Policies (SVP), which maps each state to multiple near-optimal actions to support clinician-in-the-loop decision making, as well as Dead-End Discovery (DeD), which eliminates high-risk actions in order to avoid undesirable outcomes. While SVP and DeD appear complementary–in that the actions not chosen by SVP could correspond to the same actions eliminated by DeD, and vice versa–the consistency of their recommendations has not been systematically studied. In this work, we empirically evaluate the consistency of SVP and DeD in a clinically inspired grid-world domain, analyzing how their consistency varies across different hyperparameter settings. Our results reveal the complexity of this problem, where seemingly reasonable heuristics on hyperparameter values or action set sizes fail to guarantee consistency. We demonstrate a method to visualize consistency patterns across hyperparameter configurations, highlight conditions under which consistency is more likely achieved, and explore possible reasons for divergence between the two approaches. Our findings underscore the importance of empirically analyzing potential inconsistencies of SVP and DeD before they are deployed and used together on real-world applications.

1 Introduction

Reinforcement learning (RL) has been framed as the problem of solving the optimal policy, which identifies *one action* for each state that yields the highest discounted expected cumulative reward (Sutton & Barto, 1998; 2018). This is done both for simplicity of algorithm implementation and because such deterministic policies are sufficient in fully observable environments (Puterman, 1994). However, under this standard formulation, the non-optimal actions (per state) are entirely ignored. Among recent developments, Set-Valued Policies (SVP) by Tang et al. (2020) and Dead-End Discovery (DeD) by Fatemi et al. (2021) are two notable approaches that proposed different paradigms, both motivated by applications in healthcare.

SVP produces a set of near-optimal actions for each state that lead to similar returns in the worst case rather than a single best action that leads to the highest return, which enables the clinician-in-the-loop decision making (Fard & Pineau, 2011). In contrast, DeD adopts a risk-sensitive perspective by proactively identifying irreversible states (e.g., severe deterioration or mortality) and eliminating actions with a high possibility of leading to such states (Fatemi et al., 2019; Killian et al., 2023).

While both SVP and DeD move beyond the rigid single-action policy of standard RL, they each tackle the problem from a different angle. SVP compares all non-optimal actions with the optimal action by their Q-values (for a particular state) and provides a set of actions that are within a predefined "near-optimality margin" (e.g., 5%) of the optimal action. On the other hand, DeD eliminates potentially high-risk actions through a "death threshold" (e.g., actions that lead to a probability of death $> 95\%$ will be flagged as unsafe) and thereby retaining a set of viable actions. Both SVP and DeD can be seen as a mechanism to partition the action space into a desirable action set and an undesirable action set for each state, and they share a similar goal of communicating action recommendations to a more user-friendly format. However, it is unclear whether the recommendations from SVP and DeD are guaranteed to be consistent with each other. Due to mis-estimation of Q-values with insufficient data and depending on choices of the hyperparameters (near-optimality margin and death threshold), SVP may include unsafe actions whereas DeD may discard viable actions. Thus, it may be possible that the same action is flagged as high-risk by DeD but also included in the SVP's near-optimal set. Understanding whether and how these two approaches may produce conflicting recommendations can guide us to better leverage their strengths and provide a consistent output to end-users receiving these recommendations. For instance, with a consistent policy combination, SVP can recommend a set of treatments that maintain decent therapeutic effect while DeD can filter out treatments with significant side-effect or high uncertainty.

In this paper, we begin to answer this question via an empirical evaluation on a grid-world domain, investigating the consistency between SVP and DeD across different hyperparameters. Specifically, we study two seemingly reasonable heuristics for ensuring consistency: (1) where the number of actions flagged as unsafe by DeD plus the number of actions recommended by SVP is smaller than the total number of available actions; (2) when the near-optimal margin and death threshold sums to less than 1. We also examine what kind of states are more vulnerable to conflicts across different hyperparameter combinations. Our results show that SVP and DeD can indeed produce conflicting recommendations, and there does not appear to be simple, generalizable conditions for guaranteeing consistency. Our findings highlight the challenging nature of the problem, motivating further mathematical study into their theoretical foundations and new strategies for unifying them.

2 Background

Reinforcement learning formalizes sequential decision-making as a Markov decision process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is a set of actions, $P(s'|s, a)$ is the transition probability from state s to s' under action a, $R(s, a)$ is the reward function computing the expected reward under action a given state s, and $\gamma \in [0, 1)$ is the discount factor. In the standard RL formulation, the goal is to learn a policy $\pi(a|s)$ that produces the highest expected discounted return. State-action value function $Q^\pi(s, a)$ estimates the expected return from taking action a in state s and following policy π thereafter.

In this work, we consider two approaches that move beyond standard RL policy formulation, namely SVP (Tang et al., 2020) and DeD (Fatemi et al., 2021). To aid our explanation, we will use a unified notation for the "policies" produced by both approaches, $\pi_{\text{SVP}}(s) \subseteq A$ and $\pi_{\text{DeD}}(s) \subseteq A$.

2.1 Set-Valued Policies (SVP)

An SVP policy maps each state to a set of near-optimal actions. Formally,

$$\pi_{\text{SVP}}(s) = \{a \in \mathcal{A} \,|\, Q^{\pi_{\text{SVP}}}(s, a) \geq (1 - \zeta)V^*(s)\}\,.$$

where $\zeta \in [0, 1]$ is a hyperparameter that specifies the acceptable margin of near-optimality.

A larger ζ produces a more inclusive policy that considers a broader set of actions as near-optimal, whereas a smaller ζ produces a more conservative policy with very few action choices, and in the limit of $\zeta = 0$, only the optimal action is included. Importantly, $Q^{\pi_{\text{SVP}}}(s, a)$ is the *worst-case* Q-value (details in Tang et al. (2020)) that considers the worst future trajectory possible under the SVP's recommendation such that, no matter which action the end-user chooses from the near-optimal set, the long-term reward is still ζ-close to the optimal value $V^*(s)$.

2.2 Dead-End Discovery (DeD)

A DeD policy maps each state to a set of risky/unsafe actions. Formally,

$$\pi_{\text{DeD}}(s) = \{a \in \mathcal{A} \,|\, -Q_D(s,a) \geq \theta_D\}\,.$$

DeD estimates a separate value function $Q_D(s,a) \in [-1,0]$, which is the value function resulting from a specific MDP and reward structure (the "death MDP", details in Fatemi et al. (2021)) such that $-Q_D(s,a)$ represents the probability of eventual death after taking action a in state s. $Q_D(s,a) = -1$ indicates a 100% probability of reaching a terminal death state.

The main hyperparameter of DeD is the *death threshold* $\theta_D \in [0,1]$. A larger threshold (closer to 1) makes DeD more tolerant of risky actions (only eliminating actions that are almost certainly leading to death), while a smaller threshold (closer to 0) results in a more conservative, risk-averse policy.

3 Consistency between SVP and DeD

While SVP recommends the best few actions with the highest expected cumulative rewards, DeD eliminates the worst few actions with the highest possibility of death. Ideally, the two approaches should produce consistent recommendations, where the treatments eliminated by DeD should not overlap with those selected as near-optimal by SVP. Understanding where and why inconsistencies or conflicts occur is key to designing a unified framework for combining these two approaches. To quantify the inconsistency, we propose both a state-level and a policy-level measure.

State-Level Inconsistency. Policies π_{SVP} and π_{DeD} are said to be inconsistent at state $s \in \mathcal{S}$ if $\pi_{\text{SVP}}(s) \cap \pi_{\text{DeD}}(s) \neq \emptyset$, i.e., there exists at least one action that is both near-optimal (according to SVP) and risky (according to DeD) for state s.

Policy-Level Inconsistency. Given π_{SVP} and π_{DeD}, we can quantify their extent of inconsistency by calculating the fraction of states that are inconsistent. If there are any inconsistent states, we say π_{SVP} and π_{DeD} are inconsistent.

4 Experimental Setup

We conduct our experiments using the LifeGate environment (Fatemi et al., 2021), a synthetic grid world domain designed to simulate the clinical setting. An agent's action set comprises moving up, down, right, left, or staying in place. For both SVP and DeD, we set the discount factor $\gamma = 1$ to ensure that SVP properly account for eventual recovery as the final objective without introducing any temporal preferences, and that DeD accurately estimates the probability of reaching death zones. The LifeGate environment (Figure 1) includes the following key components:

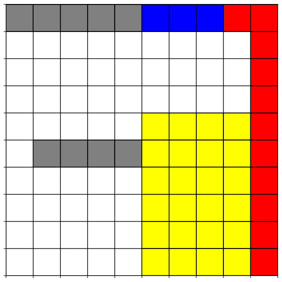


Figure 1: LifeGate environment

Barriers: The gray areas in the figure. If the agent attempts to move into a barrier, it will be redirected back to stay at its current position without receiving any reward or penalty.

Dead-ends: The yellow areas in the figure. States that represent irreversible failure conditions. When entering a dead-end, the agent faces a 70% probability of remaining in the same position (*no-move*) and a 30% probability of being pushed one cell to the right (*death-drag*). This simulates deterioration that cannot be avoided even with intervention in clinical practice.

Death zones: The red areas in the figure. Reaching a death zone immediately terminates the simulation, indicating that the agent has failed to recover (death).

Recovery zones (the "life gates"): The blue areas in the figure. Reaching a recovery zone also ends the simulation, signifying that the agent has successfully recovered.

Neutral states: The white areas in the figure. These comprise the majority of the environment and allow the agent to continue navigating without immediate reward or penalty. However, there is a natural drag mechanism that pushes the agent to the right with a probability of 40%, regardless of the chosen action. This simulates unexpected patient deterioration in the medical domain.

We implemented the environment with two different reward functions tailored to the goals of SVP and DeD. For SVP, the agent receives a reward of $+1$ when reaching a recovery zone and -1 when reaching a death zone, with zero reward for all intermediate transitions, so as to ensure the value function reflects both long-term success and failure. Since all actions in recovery zones, death zones, and dead-end states produce identical rewards, SVP would trivially include every action in the recommendation, conflicting with DeD's exclusions. To enable a meaningful consistency check, we therefore restrict SVP to provide recommendations within non-terminal states. For DeD, the agent receives a reward of -1 only when entering a death zone, while all other transitions yield zero reward, following the setup in Fatemi et al. (2021). This ensures the value function reflects the probability of reaching of eventual death.

5 Experimental Results

Our analysis aims to investigate the consistency of SVP and DeD across various hyperparameter settings. Specifically, we trained 101 versions of SVP policy with ζ from 0 to 1 in increments of 0.01, and 101 versions of DeD policy with θ_D from -1 to 0 in increments of 0.01, and compared the consistency of SVP and DeD for each pair of hyperparameters (ζ, θ_D). The results are organized in two levels of insights: we first consider *global trends*, looking at how policy consistency changes across the hyperparameter space; we then consider *local trends*, investigating which specific states are more prone to conflict between SVP and DeD, and under what hyperparameter settings.

5.1 Global Trends

First, we investigate how ζ affects SVP size and how θ_D affects DeD size by plotting the average number of actions per state across the hyperparameter space (Figure 2). As expected, higher ζ tends to produce a larger SVP set, since more actions are considered near-optimal. However, the relationship between SVP size and ζ is not monotonic. On the other hand, increasing θ_D results in fewer actions being eliminated by DeD, as the threshold for risk becomes more tolerant.

Next, we visualize the SVP-DeD policy inconsistency across different hyperparameter combinations (ζ, θ_D) as a heatmap. In Figure 3, dark green regions represent full consistency between the SVP and DeD for that hyperparameter pair. The top left corner of the heatmap corresponds to a conservative setting, where DeD only eliminates actions that have 100% probability of death, and SVP only includes the optimal actions, and so there is no conflict between the two. The bottom right corner corresponds to the situation where DeD retains actions that have 0% probability of death and ends up eliminating all actions, and SVP includes all the actions, and this is where we see the highest level of inconsistency. For a particular ζ, as θ_D decreases down the heatmap, the average DeD size increases monotonically, which results in more inconsistency. Similarly, for a particular θ_D, as ζ increases towards the right of the heatmap, the average SVP size tends to increase, leading to more inconsistency, but the trend here is less pronounced and not monotonic.

Is consistency guaranteed when SVP size and DeD size sum to fewer than all actions? Since DeD eliminates the worst actions and SVP recommends the best actions, a conflict is inevitable when their combined set sizes are larger than the total number of actions; by the pigeonhole principle, at least one action must be flagged as unsafe by DeD while also being recommended by SVP. Therefore, one reasonable heuristic to guarantee consistency is if the combined size of the SVP set and DeD set is less than the total number of available actions. At the state-level, we found that among 429,852 states across all hyperparameter combinations satisfying the criterion $|\pi_{\text{SVP}}(s)| + |\pi_{\text{DeD}}(s)| < |\mathcal{A}|$, 247 were inconsistent, corresponding to a consistency rate of 99.94%. At the policy-level, however, our results show otherwise. Among 9,677 hyperparameter combinations satisfying the criterion that the average combined size of the SVP set and DeD set is less than the size of the action set, only 2,447 were consistent, corresponding to a consistency rate of 25.29%.

Does $\zeta + \theta_D < 1$ imply consistency? Another possible heuristic is whether the near-optimal margin and death threshold add up to less than 1. Intuitively, since ζ controls the tolerance of SVP for near-optimal actions and θ_D controls the tolerance of DeD for high-risk actions, if $\zeta + \theta_D = 1$ then we might reasonably expect the SVP and DeD action sets to be the exact complement of each other. The hyperparameter combinations where $\zeta + \theta_D < 1$ correspond to the lower-left triangular region of the heatmap in Figure 3. Unfortunately, it is clear that this heuristic does not guarantee consistency either. Out of 5,050 parameter combinations satisfying $\zeta + \theta_D < 1$, only 496 resulted in consistent policies, corresponding to a consistency rate of merely 9.82%.

When do we get consistency? Based on Figure 3, we can identify a dense band of hyperparameter pairs that always yield consistent policies. Specifically, when $\zeta \in [0.00, 0.44]$ and $\theta_D \in [0.69, 1.00]$, all tested combinations result in full consistency. This consistency extends to slightly higher ζ values if θ_D is set to be near 0.79 or higher. Looking at Figure 2, we observe that SVP recommends more than one treatment on average when $\zeta \in (0.00, 0.44]$ and DeD always eliminates at least one treatment on average when $\theta_D \in [0.69, 1.00]$, which means that both policies remain non-trivial within this hyperparameter range. Given the irregular contours we observed in the heatmap, we believe it could be challenging to provide a generalizable rule that guarantees consistency between SVP and DeD. Therefore, in practice, it would be essential to empirically test and retrieve a range of (ζ, θ_D) combinations that yield consistenct policies, treating the two methods as a form of cross-validation to enhance recommendation reliability.

5.2 Local Trends

Which states are more vulnerable to policy conflicts? To empirically investigate state-level vulnerability to policy conflict, we plotted the frequency that each state receives an inconsistent action recommendation across all (ζ, θ_D) combinations (Figure 4). The dark red region in the heatmap corresponds to the states with the most conflicts, whereas the white region represents the least vulnerable state. The resulting heatmap reveals distinct spatial trends in policy vulnerability. States located near the boundary of the death zone (the rightmost column) and the boundary of dead-end corridors (the bottom right region) exhibit the highest normalized conflict frequencies. In contrast, all terminal states (recovery, death, dead-end states) and states far from the death zone consistently

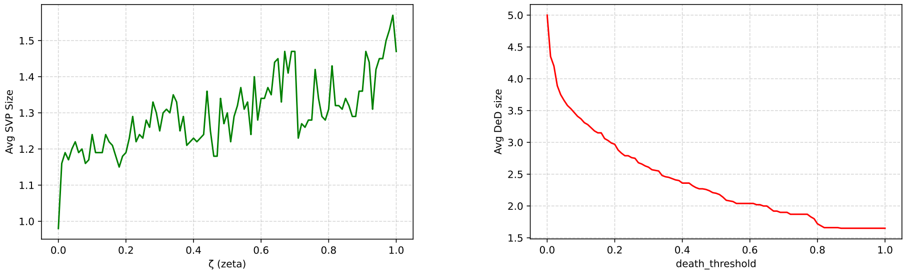


Figure 2: Left - Average SVP policy size vs. ζ. Right - Average DeD policy size vs. θ_D.

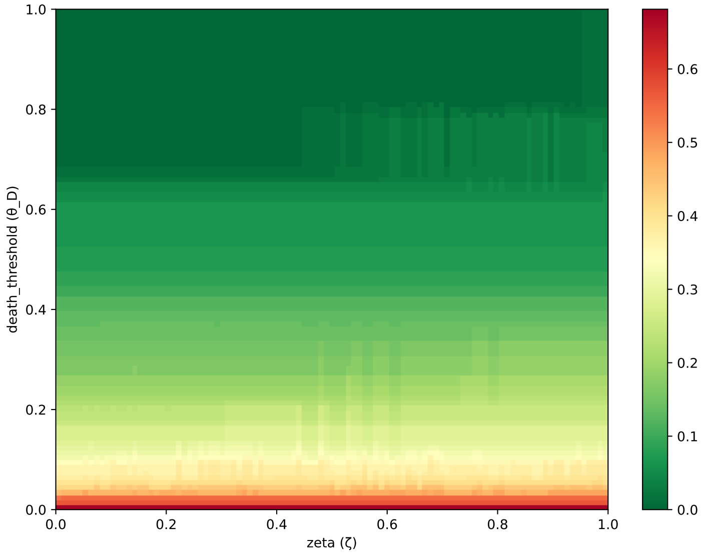


Figure 3: Heatmap of SVP-DeD inconsistency (measured as fraction states that are inconsistent) under different hyperparameter combinations of (ζ, θ_D).

demonstrate negligible conflict rates. This suggests that these regions are largely insensitive to variations in ζ and θ_D, and that policy disagreements are most likely to happen in regions where ambiguity exists near critical transitions.

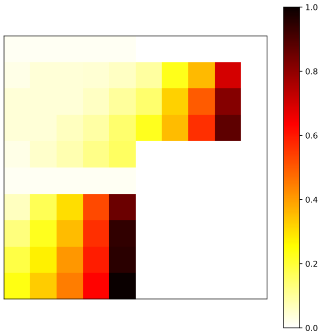
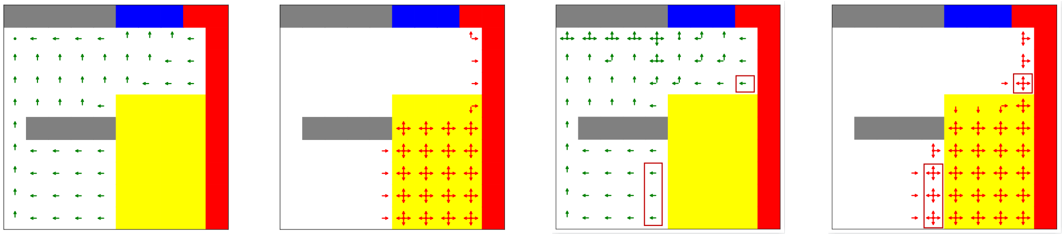


Figure 4: Heatmap of normalized frequencies that each state receives inconsistent recommendations from SVP and DeD across all (ζ, θ_D) pairs.

Counterexample: conflict under low ζ and moderate θ_D. While our previous analysis identified a broad hyperparameter range where policy consistency is generally maintained, we now highlight a notable counterexample. Consider $\zeta = 0.10$ and $\theta_D = 0.65$, a configuration where the SVP applies a relatively conservative margin (near-optimal is within 10% of the optimal) and the death-threshold is moderate (actions leading to a death probability exceeding 65% are eliminated). Under this setting, we examine the consistency between action sets derived from SVP and DeD. We visualize the policies produced by SVP and DeD in this scenario, together with another pair of hyperparameters where no conflict occurs ($\zeta = 0.00$ and $\theta_D = 1.00$) as a frame of reference (Figure 5). Despite the low value of ζ, which cautiously recommends the near-optimal actions, conflicts between the two policies still emerge in several states, particularly those located close to death and dead-end zones (highlighted with red rectangle). This example further emphasizes the importance of incorporating empirical validation when selecting hyperparameters, as certain states still remain vulnerable to conflict even under seemingly conservative settings.



(a) SVP with ζ = 0.00 (b) DeD with θ_D = 1.00 (c) SVP with ζ = 0.10 (d) DeD with θ_D = 0.65

Figure 5: Comparisons between SVP and DeD strategies under different hyperparameter settings. Green arrows in the figure represent actions that are recommended by SVP and red arrows points to actions flagged as high-risk by DeD. No inconsistency is seen between (a) and (b). There are 4 states with conflicting recommendations between (c) and (d). The green arrows represent actions that are recommended by SVP and red arrows points to actions flagged as high-risk by DeD.

6 Discussion and Future Work

In this work, we investigated the consistency between SVP and DeD, two recent approaches that move beyond the standard RL formulation of single-action optimal policies. Although they appear intuitively complementary, with SVP retaining near-optimal actions and DeD eliminating high-risk actions, our findings demonstrate that their recommendations are not always consistent. Empirically, we observed consistency under a subset of hyperparameter configurations, particularly when ζ is small and θ_D is large. This range reflect practical safety and performance consideration, where only actions near the optimal (small ζ) are recommended and actions along with high probability of catastrophic failure (large θ_D) are eliminated. Beyond this region, inconsistencies were observed in some seemingly reasonable settings, especially for states near high-risk regions such as the dead-end corridors or close to death zones. These results suggest that the interaction between SVP and DeD should be carefully evaluated before they can be deployed jointly in real-world applications.

The inconsistencies we observe could be an indication that SVP and DeD are prioritizing different aspects of the action's quality, which is the consequence of them making use of different reward functions and value functions. Specifically, for the "life gate" terminal state, SVP assigns a reward of $+1$ whereas DeD assigns a reward of 0; SVP uses the worst-case value functions whereas DeD uses the standard (best-case) value functions. This highlights an opportunity for hybrid frameworks (Lizotte & Laber, 2016; Harsh et al., 2021) that can unify these two considerations explicitly rather than treating it as a post-processing step. Several promising directions can be further studied. First, developing theoretical conditions where consistency is guaranteed can offer practitioners with more assurances beyond empirical observation. Second, we can enhance the robustness of SVP by integrating an dynamically adaptive margin that can adjust based on state's risk level into the SVP framework. Pursuing these directions may potentially lead to safer and more adaptable decision-making support, particularly in high-risk domains such as healthcare, where both flexibility and risk-awareness are essential (Gu et al., 2022; Sivaraman et al., 2023).

References

M Milani Fard and Joelle Pineau. Non-deterministic policies in markovian decision processes. *Journal of Artificial Intelligence Research*, 40:1–24, 2011.

Mehdi Fatemi, Shikhar Sharma, Harm Van Seijen, and Samira Ebrahimi Kahou. Dead-ends and secure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 1873–1881. PMLR, 2019.

Mehdi Fatemi, Taylor W Killian, Jayakumar Subramanian, and Marzyeh Ghassemi. Medical dead-ends and learning to identify high-risk states and treatments. *Advances in Neural Information Processing Systems*, 34:4856–4870, 2021.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

Satija Harsh, Philip S Thomas, Joelle Pineau, Romain Laroche, et al. Multi-objective spibb: Seldonian offline policy improvement with safety constraints in finite mdps. *Advances in Neural Information Processing Systems*, 34:2004–2017, 2021.

Taylor W. Killian, Sonali Parbhoo, and Marzyeh Ghassemi. Risk sensitive dead-end identification in safety-critical offline reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=oKlEOT83gI.

Daniel J Lizotte and Eric B Laber. Multi-objective markov decision processes for data-driven decision support. *Journal of Machine Learning Research*, 17(210):1–28, 2016.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. Ignore, trust, or negotiate: understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

Shengpu Tang, Aditya Modi, Michael Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. In *International Conference on Machine Learning*, pp. 9387–9396. PMLR, 2020.