Fair Distributional Reinforcement Learning

Zequn Chen[0009-0000-0693-1250] and Wesley J. Marrero[0000-0002-7092-2292]

Thayer School of Engineering at Dartmouth College, Hanover NH 03755, USA
zequn.chen.th@dartmouth.edu
wesley.marrero@dartmouth.edu

Abstract. Distributional reinforcement learning (DRL) extends traditional reinforcement learning by modeling the distribution of returns rather than focusing solely on their expectation, enabling more nuanced decision making. However, existing DRL approaches may not be appropriate for settings where equitable outcomes are essential, such as medical decision making. To address this limitation, we propose fair distributional reinforcement learning (FDRL). This algorithm finds policies that approach near-optimal returns and promote fairness by ensuring similar performance among individuals from different contextual subgroups.

The proposed algorithm strikes a balance between maximizing expected returns and minimizing inequalities across population subgroups by augmenting the DRL loss function to address group-level disparities in return distributions. To this end, we construct a loss function with two components: the first quantifies the discrepancy between the predicted and target return distributions, representing the loss for precision; the second component measures the difference between the target return distributions of vulnerable and resilient subgroups, serving as a fairness penalty. If subgroups in a population cannot be identified from a problem context, our approach stratifies individuals based on their probability of experiencing outcomes of interest. The effectiveness of the FDRL framework is evaluated in the context of hypertension management. Experimental results demonstrate that FDRL significantly improves fairness while maintaining near-optimal policy performance. Notably, individuals across diverse demographic groups achieve comparable long-term health outcomes, underscoring the algorithm's ability to ensure equitable treatment without sacrificing overall efficiency.

Keywords: Reinforcement learning · Fairness · Return distributions.

1 Introduction

Discrete-time Markov decision process (MDP) models with finite states, actions, and horizons have been used to inform decisions in a variety of applications, including medicine, transportation, and energy. In the standard MDP setting, a decision maker aims to find a sequence of actions that maximizes the expected return over the planning horizon. Expected return alone may not provide sufficient information to capture uncertainty, and decision-makers in practical applications may instead prioritize optimizing other return properties, such as utilities, conditional value at risk, or entire distributions [3,19]. Recognizing that returns cannot be characterized by their expectation alone, distributional reinforcement learning (DRL) has emerged as a promising approach to consider different return properties by operating on probability distributions [?]. However, as DRL expands to real-world contexts—ranging from healthcare to finance—the potential for biased decision-making becomes a pressing concern.

Although the ideas of DRL can be traced back to Howard and Matheson or Sobel [14,28], it has gained more attention recently after the introduction of the C51 algorithm [3]. Unlike traditional Q-learning, which focuses on estimating the expected return, the C51 algorithm employs a Deep Q-Network (DQN) to approximate the full distribution of returns, offering a richer representation of uncertainty [20]. Researchers have explored various aspects of DRL, including its robustness and error reduction capabilities [18,8]. Other studies delve into nonparametric methods for approximating return distributions, providing alternative approaches to capture the variability in returns [21]. Nonetheless, no method has addressed the potential consequences of inequitable return distributions. Addressing this gap is essential to ensure DRL promotes fair decision-making and does not inadvertently exacerbate outcome inequalities.

The need for fairness in the distribution of returns is inspired by the medical decision making setting, where a healthcare provider must determine an optimal therapy for patients with different perceptions and risk tolerances. Within medicine, we focus on the management of high blood pressure (BP), a key controllable risk factor of atherosclerotic cardiovascular disease (ASCVD) [36]. Our approach may be beneficial in this setting as patient belief has been associated with appropriate disease management and there have been concerns of racial and gender outcome disparities [15,1,37,23,33].

This research initiates the study of fairness in DRL by exploring cases where the environment can be represented with finite MDP models. We present a new algorithm that promotes fairness across contextual groups, which we will refer to as fair distributional reinforcement learning (FDRL). Our approach incorporates a fairness regularizer to the C51 algorithm that penalizes it if distinct groups achieve different return distributions [3,4]. In addition, we apply our FDRL to personalized hypertension treatment planning using a large population in the United States. Through this case study, we compare our approach to the C51 algoritm [3], Q-learning [34], and the most recent clinical guidelines [36].

The remainder of this paper is organized as follows. In Section 2, we specify the setting of our work. Section 3 presents our FDRL algorithm. In Section 4, we exhibit the hypertension management case study. Finally, conclusions and potential research directions are discussed in Section 5.

2 Setting

In this paper, we represent the interactions between a decision maker and a fully observable system through an MDP model. We consider a system with a finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, and horizon $\mathcal{T} \coloneqq \{0, 1, \ldots, T\}$. At every decision epoch $\mathcal{T} \backslash \{T\}$, the decision maker observes the state of the system $s_t \in \mathcal{S}$ and chooses action $a_t \in \mathcal{A}$ according to policy $\pi_t : \mathcal{S} \mapsto \mathcal{A}$. After action a_t is selected, the decision maker receives a reward $r_t(s_t, a_t) \in \mathbb{R}_{\geq 0}$, and a new state $s_{t+1} \in \mathcal{S}$ is realized with probability $p_t(s_{t+1}|s_t, a_t)$. Upon reaching $s_T \in \mathcal{S}$ at time T, the decision maker receives a terminal reward of $r_T(s_T) \in \mathbb{R}_{\geq 0}$. Future rewards are discounted at a rate of $\gamma \in (0, 1]$.

An MDP is defined by the tuple $\mathcal{M} \coloneqq (\mathcal{T}, \mathcal{S}, \mathcal{A}, P, r, \gamma)$. Given an initial state s, we aim to find a policy $\pi \coloneqq (\pi_t(s_t) : t \in \mathcal{T} \setminus \{T\}, s_t \in \mathcal{S})$ that maximizes function f of the return $Z^\pi(s, a) \coloneqq \sum_{t=0}^{T-1} \gamma^t r_t(s_t, \pi_t(s_t)) + \gamma^T r_T(s_t)$ [?]:

$$J_f(\pi) \coloneqq \sup_\pi f\left(Z^\pi(s, a)\right),$$

where $a = \pi_0(s)$. The function f must be *Bellman Optimizable* [19], such as the expected value $f(Z) = \mathbb{E}[Z]$ or mean-variance $f = \mathbb{E}[Z] - \alpha \text{Var}(Z)$ for $\alpha > 0$.

3 Fair Distributional Reinforcement Learning

We now present our FDRL algorithm to learn policies that generate fair returns based on a finite MDP model. Our goal is to learn a near-optimal policy $\hat{\pi}$ based on a function of returns $f(Z^{\hat{\pi}}(s, a))$ updated by exploring actions according to an ϵ-greedy behavior policy b. Our method is outlined in Algorithm 1.

3.1 Contextual Groups

We assume there are N agents (e.g., patients) in the system of interest which can be categorized into groups $g_1, \ldots, g_K$. If agent groups can be identified from the problem context (e.g., sex or race groups), our algorithm can use them directly. Otherwise, we rank the agents $1, ..., N$ based on their probability of experiencing outcomes of interest (e.g., no ASCVD events). We calculate an aggregate ranking for each agent as the average of their rankings across all actions $a \in \mathcal{A}$. Based on the aggregate ranking, we cluster the N agents into K groups using K-means and order the groups from resilient to vulnerable [12]. The clusters can be interpreted such that for any $k < k'$, any agent $j \in g_k$ is more privileged than agent $j' \in g_{k'}$.

3.2 Algorithm

In DRL, the goal is to model the entire distribution of returns, rather than just the expected return as in traditional reinforcement learning. The return distribution is modeled through a DQN parameterized by θ [20], which outputs a set of probabilities or values representing the distribution over discrete support points or quantiles [3,11]. We initialize the DQN $h_\theta(\cdot)$ with no agents. As long as the DQN has used less than N agents, we sample agents from groups $\mathcal{G} \coloneqq \{g_1, \ldots, g_K\}$ with equal probability. Once agent j is sampled, we generate transitions (s_t, a_t, r_t, s_{t+1}) in each decision epoch $t \in \mathcal{T} \setminus \{T\}$ based on their

MDP $\mathcal{M}_j$. For computational reasons, we then sample the MDP of m agents in the bottom $\underline{q}$ and top $\bar{q}$ quantiles of the outcome probabilities in groups g_1 and g_K to compute the average MDP of the $2m$ agents in each group. These average MDP models reflect the centrality and tail characteristics of the rewards and transition probabilities in each group [25,27]. We regard these averages as representatives of groups g_1 and g_K and denote them by $\overline{\mathcal{M}}_1$ and $\overline{\mathcal{M}}_K$, respectively. Note that in each time step, agent j as well as the representative agents from g_1 and g_k have the same starting state s_t and action a_t.

Algorithm 1: Fair distributional reinforcement learning (FDRL).

```
Input : Let $\mathcal{M}_{1:N}$ denote the MDP models of $N$ agents categorized
        into $K$ groups $g_1, \ldots, g_K$. Set episode $i \leftarrow 0$ and initialize, $b$,
        $\underline{q}$, $\bar{q}$, $h_\theta(\cdot)$, $\lambda$, and $\mathcal{D}^j = 0$ for $j = 1, ..., N$.
while $i < N$ do
    Sample group $g_k \in \mathcal{G}$ and agent $j \in g_k$ randomly and set $i \leftarrow i + 1$;
    Initialize $s_0$;
    for $t = 0$ to $T$ do
        Choose $a_t \sim b_t(s_t)$ and generate $(r_t^j, s_{t+1}^j)$ from $\mathcal{M}_j$;
        Calculate $\overline{\mathcal{M}}_1$ and $\overline{\mathcal{M}}_K$ based on quantiles $\underline{q}$ and $\bar{q}$;
        Generate $(r_t^1, s_{t+1}^1)$ from $\overline{\mathcal{M}}_1$ and $(r_t^K, s_{t+1}^K)$ from $\overline{\mathcal{M}}_K$;
        for $n$ in $\{g_1,\ j,\ g_K\}$ do
            Use DQN $h_\theta(\cdot)$ to estimate $\mathcal{D}^n(s_{t+1}^n, a)$ for all $a$;
            Compute $a_n^* = \arg\max_a f(\mathcal{D}^n(s_{t+1}^n, a))$;
            for $d = 0$ to $D - 1$ do
                Update projection $\widehat{T}_{z_d} \leftarrow [r_t^n + \gamma z_d]_{V_{\min}}^{V_{\max}}$;
                Update position $b_d \leftarrow (\widehat{T}_{z_d} - V_{\min})/\Delta z$;
                Update $\mathcal{D}^n_{\lfloor b_d \rfloor} \leftarrow \mathcal{D}^n_{\lfloor b_d \rfloor} + P_d^n(s_{t+1}, a^*)(\lceil b_d \rceil - b_d)$;
                Update $\mathcal{D}^n_{\lceil b_d \rceil} \leftarrow \mathcal{D}^n_{\lceil b_d \rceil} + P_d^n(s_{t+1}, a^*)(b_d - \lfloor b_d \rfloor)$;
            end for
        end for
        Calculate $\mathcal{L}_1 = -\sum_d \mathcal{D}_d^j \log(P_d^j(s_t, a_t))$;
        Compute $\mathcal{L}_2 = -\sum_d \mathcal{D}_d^{g_1}(s_t, a_t) \log \mathcal{D}_d^{g_K}(s_t, a_t)$;
        Calculate $\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2$;
        Update $h_\theta(\cdot)$ using total loss $\mathcal{L}$ and set $s_t \leftarrow s_{t+1}^j$;
    end for
end while
Output: $h_\theta(\cdot)$ and $\hat{\pi}_t(s_t) := \arg\max_a f(\mathcal{D}^j(s_t, a))$ for all $j$, $s_t$, and $t$.
```

Once we generate transitions from $\mathcal{M}_j$, $\overline{\mathcal{M}}_1$ and $\overline{\mathcal{M}}_K$, we use the trained network $h_\theta(\cdot)$ to compute the return distributions $\mathcal{D}^j(s_{t+1}^j, a)$, $\mathcal{D}^{g_1}(s_{t+1}^1, a)$, and $\mathcal{D}^{g_k}(s_{t+1}^K, a)$ for all $a \in \mathcal{A}$, where s_{t+1}^n denotes the next state according to MDP n. Subsequently, we choose the action at state s_{t+1}^n for all three MDP models based on a Bellman optimizable function f.

The algorithm then proceeds to estimate the return distribution $\mathcal{D}^j$ for agent j by iterating over each support point d from 0 to $D-1$ as in the C51 algorithm [3]. We first project the sample Bellman update $\widehat{T}z_d$ for each atom z_d between bounds $V_{\min}$ and $V_{\max}$. The position b_d of $\widehat{T}z_d$ relative to the predefined support points is then computed based on Δz spacing between consecutive support points. To determine the two closest support points, the lower and upper bounds are identified as $\lfloor b_d \rfloor$ and $\lceil b_d \rceil$, respectively. The probability mass of $\widehat{T}z_d$ is distributed proportionally to support points $\lfloor b_d \rfloor$ and $\lceil b_d \rceil$, based on atom probability $P_d^j(s_{t+1}, a^*)$ given by $h_\theta(\cdot)$ at support point d, action a^*, and state s_{t+1}. By the end of the iteration, the target distribution $\mathcal{D}^j$ is constructed, reflecting the desired probability distribution over the return values for agent j.

Subsequently, we calculate the cross-entropy loss between $\mathcal{D}^j$ and P^j, $\mathcal{L}_1$, representing the precision of the network prediction. We implement the same procedures to compute the target return distribution for $\overline{\mathcal{M}}_1$ and $\overline{\mathcal{M}}_K$, and calculate the cross-entropy loss over these two distributions. This cross-entropy loss, $\mathcal{L}_2$, serves as a fairness penalty term because it measures the return distribution difference between the resilient and vulnerable subgroups. The total loss is then $\mathcal{L} \coloneqq \mathcal{L}_1 + \lambda\mathcal{L}_2, \lambda \in [0,1]$, which measures prediction precision and fairness. Finally, we use the total loss to update the parameters of $h_\theta(\cdot)$. The loop continues until the agent reaches the terminal stage T.

4 Case Study

This section evaluates the implications of the FDRL algorithm on the fairness and optimality of hypertension treatment plans. As an initial study of fairness in DRL, we focus on risk-neutral optimization (i.e., $f = \mathbb{E}$) and leave the assessment of other functions as future work. We adopt the MDP presented by Garcia and coauthors [13]. In summary, their MDP considers a planning horizon $\mathcal{T}$ of 10 years with decisions made once a year. The state space $\mathcal{S}$ consists of patients' demographic information, clinical observations, and a health condition that accounts for patient's history of ASCVD. The action space $\mathcal{A}$ contains from 0 to 5 antihypertensive medications at a half and standard dosage. Transition probabilities $p_t(s_{t+1}|s_t, a_t)$ are derived from the medical literature, including patients' risk for ASCVD events [37,5,6], the benefit from treatment [31,30,32], and mortality [22,2]. The model rewards $r_t(s_t, a_t)$ are defined as the quality of life weight associated with patients' health condition minus the treatment-related disutility from each medication [16,32]. In contrast to this study, we do not incorporate a terminal reward $r_T(s_T)$ to avoid capturing disparities in patients' life expectancy after the planning horizon. Additionally, we use $\gamma = 1$ to highlight the differences among patients' return [24].

4.1 Analysis

Although we recognize subgroups within the context of ASCVD can be identified based on race or sex [37,23], we identified $K = 3$ groups using the K-means clustering described in Section 3.1. The first group is characterized by the best overall health status, predominately younger White females, with normal BP, and it is referred to as the resilient group. The second, neutral, group captures patients with moderate risk factors such as elevated BP, middle age, and a mix of Black females and White males. Lastly, the third and vulnerable group is composed of patients with older age, hypertension, and a high prevalence of Black males. We identify representative MDP models $\overline{\mathcal{M}}_1$, $\overline{\mathcal{M}}_2$, and $\overline{\mathcal{M}}_3$ using patients ranked in the bottom $\underline{q} = 10$ and top $\bar{q} = 90$ quantiles of each group.

We simulate a total 10,000 episodes. Given that each patient is equally likely to originate from any of the $K = 3$ subgroups, we expect approximately 3,330 episodes per group. We enter the simulated data into a DQN [20], which receives the state as input and outputs the return distributions for all possible actions.

Fairness Evaluation To assess disparities in our population, we evaluate the convergence of the returns achieved by different groups. We then examine the average return difference among subgroups achieved by policy π endowed with fairness penalty $\lambda \in (0, 1]$, which we refer to as the *fairness violation*:

$$\mathrm{FV}_\lambda(\pi) := \frac{1}{\binom{K}{2}} \sum_{g_k \in \mathcal{G}} \sum_{g_{k'} \neq g_k} |\bar{Z}^\pi_{g_k} - \bar{Z}^\pi_{g_{k'}}|,$$

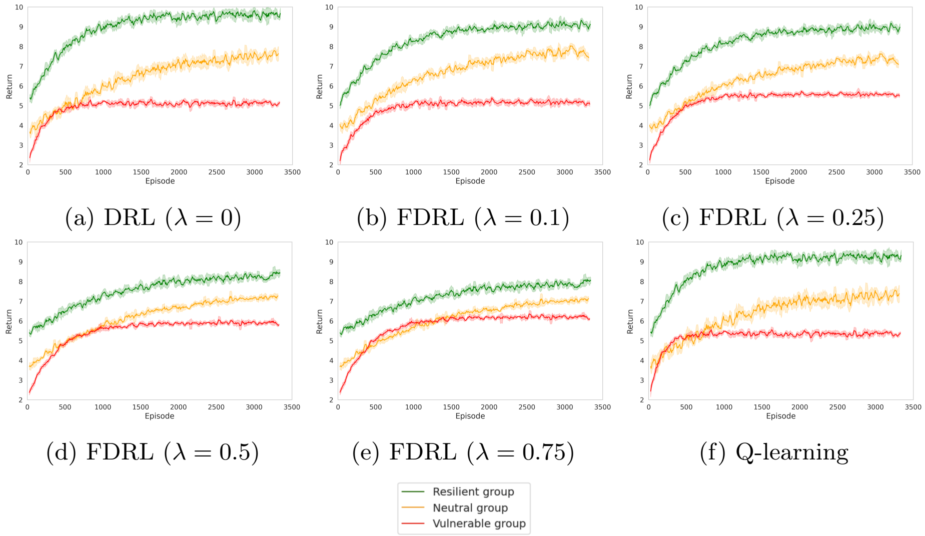
where $\bar{Z}^\pi_{g_k} := N^{-1}_{g_k} \sum_j Z^\pi_j(s, a)$ and N_{g_k} is the number of patients in group k. This definition allows us to calculate the *fairness improvement* of $\lambda \in (0, 1]$ over $\lambda = 0$ as the percentage decrease in fairness violation.

4.2 Numerical Results

Figure 1 illustrates the return over a 10-year horizon for patients categorized into three patient subgroups. These categories reflect varying levels of health and serve as representative groups for analyzing the impact of FDRL.

When the fairness penalty parameter is $\lambda = 0$, our algorithm is equivalent to C51 [3]. The derived policy prioritizes optimality without regard of fairness across subgroups. In this scenario, patients in the resilient group achieve the highest return, with the final value fluctuating around 9.7, closely approaching the theoretical maximum return of 10. However, the disparity in returns between subgroups is pronounced. These gaps underscore the need for incorporating fairness considerations into the policy optimization process.

We evaluate the overall optimality through the average return of the three groups denoted as $\bar{V}_0(s_0)$ in Table 1. As the fairness penalty λ is increased, the results show notable changes in the distribution of returns across the subgroups. For patients in the resilient subgroup, the return decreases as fairness is prioritized. Conversely, the returns for patients in the neutral and vulnerable subgroups exhibit marked improvements, effectively reducing the return disparity between subgroups. This observation demonstrates the ability of the FDRL algorithm to balance the trade-off between fairness and optimality by redistributing returns to promote equity. However, when λ is set to excessively large



(a) DRL ($\lambda = 0$)

(b) FDRL ($\lambda = 0.1$)

(c) FDRL ($\lambda = 0.25$)

(d) FDRL ($\lambda = 0.5$)

(e) FDRL ($\lambda = 0.75$)

(f) Q-learning

Fig. 1: Average return by group across learning episodes.

values, the average returns of all three subgroups decline drastically. Under such conditions, the emphasis on fairness undermines the overall effectiveness of the policy, reducing its capacity to achieve high cumulative returns for any subgroup. For $\lambda = 0.1, 0.25$, the average return of the FDRL algorithm exceeds that of Q-learning and demonstrates improved performance in terms of both fairness and optimality. In addition, we find that our policies outperform the clinical guidelines, which achieved average return of 5.41, 4.51, and 3.83 quality-adjusted life years (QALYs) across the resilient, neutral, and vulnerable groups, respectively.

We use $\overline{\mathcal{M}}_1, \overline{\mathcal{M}}_2, \overline{\mathcal{M}}_3$ to denote the average return for the resilient, neutral, and vulnerable groups in Table 1, respectively. As λ increases to 0.1, 0.25, 0.5, and 0.75, the fairness violation is decreased and the fairness improvement results in $16.18\%, 25.57\%, 44.46\%$, and 60.84%, respectively. The results reveal that when λ is chosen judiciously, the average return under a fair policy remains near-optimal. In such cases, the policy achieves a balance between fairness and performance, addressing the disparities between subgroups without significantly compromising overall returns. In contrast, excessively large values of λ result in a fair but far-from-optimal policy, as the algorithm prioritizes fairness to the detriment of satisfactory returns.

These empirical observations align with the theoretical intuition regarding the influence of fairness penalties on policy optimization. They highlight the importance of selecting an appropriate value for λ to achieve a desirable trade-off between fairness and optimality. Practitioners are advised to carefully choose the fairness weight based on the specific objectives of their application, ensuring that the policy achieves both equity across subgroups and robust performance.

Policy	**FV**	$\bar{V}_0(s_0)$	$\bar{\mathcal{M}}_1$	$\bar{\mathcal{M}}_2$	$\bar{\mathcal{M}}_3$
DRL ($\lambda = 0$)	3.09	7.59	9.74	7.93	5.11
FDRL ($\lambda = 0.10$)	2.59	7.49	9.18	8.01	5.29
FDRL ($\lambda = 0.25$)	2.30	7.41	9.01	7.66	5.56
FDRL ($\lambda = 0.5$)	1.71	7.26	8.40	7.54	5.84
FDRL ($\lambda = 0.75$)	1.21	7.18	8.01	7.34	6.19
Q-learning	2.61	7.35	9.24	7.50	5.32
Clinical guidelines	1.05	4.58	5.41	4.51	3.83

Table 1: Summary of fairness violation (FV) and average return in overall $\bar{V}_0(s_0)$ and across resilient $\bar{\mathcal{M}}_1$, neutral $\bar{\mathcal{M}}_2$, and vulnerable $\bar{\mathcal{M}}_3$ groups.

5 Conclusions

In this paper, we initiated the study of fairness within DRL. Building upon the initial C51 algorithm and fairness regularization [3,4], we presented a DRL method that penalizes a DQN agent if it treats distinct groups differently. By addressing the potential for biased outcome distributions, the algorithm presented in this paper improves the usability and acceptance of DRL in practice.

Two primary conclusions can be made from our hypertension treatment case study. First, our FDRL achieves better performance than the clinical guidelines and Q-learning with only minor negative consequences compared to the C51 algorithm [3], while achieving more equitable outcomes. Second, the degree of the fairness penalty may greatly influence the optimality and equity of outcomes. This parameter must be chosen carefully, depending on the goals of health providers. Our results make headway in reducing outcome disparities without considerably affecting the health outcomes of resilient populations.

There are opportunities for future work that build upon our initial study of algorithmic fairness in DRL. Our regularization technique may be extended to other DRL algorithms, such as quantile-regression DQN [11,10,26], implicit quantile networks [9], Sinkhorn DRL [29], and distributional policy-gradient [17]. Another extension may be to consider a richer class of policies beyond the expected value. Ideas from risk-sensitive DRL along with expected utilities and distorted means could be used to achieve fairness in general policies [18,19]. In addition, other approaches to promote fairness can be examined, like adversarial learning, reweighting, and constrained optimization [7,35].

We hope the FDRL algorithm paves the way for a line of work that enhances the usability of return distributions in high-stake situations. Our work is a step toward policies that promote equitable outcomes across diverse populations. Fair policies have great potential to enable the implementation of DRL-guided recommendations into practice within and beyond healthcare applications.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdalla, M., Bolen, S.D., Brettler, J., Egan, B.M., Ferdinand, K.C., Ford, C.D., Lackland, D.T., Wall, H.K., Shimbo, D., on behalf of the American Heart Association and American Medical Association: Implementation Strategies to Improve Blood Pressure Control in the United States: A Scientific Statement From the American Heart Association and American Medical Association. Hypertension **80**(10) (Oct 2023). https://doi.org/10.1161/HYP.0000000000000232, https://www.ahajournals.org/doi/10.1161/HYP.0000000000000232
2. Arias, E., Xu, J.: United States Life Tables, 2017. National Vital Statistics Reports **68**(7) (2019)
3. Bellemare, M.G., Dabney, W., Munos, R.: A Distributional Perspective on Reinforcement Learning (Jul 2017), http://arxiv.org/abs/1707.06887, arXiv:1707.06887 [cs, stat]
4. Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., Roth, A.: A Convex Framework for Fair Regression (Jun 2017). https://doi.org/10.48550/arXiv.1706.02409, http://arxiv.org/abs/1706.02409, arXiv:1706.02409 [cs]
5. Brønnnum-Hansen, H., Jørgensen, T., Davidsen, M., Madsen, M., Osler, M., Gerdes, L.U., Schroll, M.: Survival and cause of death after myocardial infarction: the Danish MONICA study. Journal of Clinical Epidemiology **54**(12), 1244–1250 (2001). https://doi.org/10.1016/S0895-4356(01)00405-X
6. Burn, J., Dennis, M., Bamford, J., Sandercock, P., Wade, D., Warlow, C.: Long-term risk of recurrent stroke after a first-ever stroke. The Oxfordshire Community Stroke Project. Stroke **25**(2), 333–7 (1994). https://doi.org/http://dx.doi.org/10.1161/01.STR.25.2.333
7. Caton, S., Haas, C.: Fairness in Machine Learning: A Survey. ACM Computing Surveys **56**(7), 1–38 (Jul 2024). https://doi.org/10.1145/3616865, https://dl.acm.org/doi/10.1145/3616865
8. Clavier, P., Allassoniere, S., Pennec, E.L.: Robust reinforcement learning with distributional risk-averse formulation. arXiv preprint arXiv:2206.06841 (2022)
9. Dabney, W., Ostrovski, G., Silver, D., Munos, R.: Implicit Quantile Networks for Distributional Reinforcement Learning. In: Proceedings of the 35th International Conference on Machine Learning. pp. 1096–1105. PMLR (Jul 2018), https://proceedings.mlr.press/v80/dabney18a.html, iSSN: 2640-3498
10. Dabney, W., Rowland, M., Bellemare, M., Munos, R.: Distributional Reinforcement Learning With Quantile Regression. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11791, https://ojs.aaai.org/index.php/AAAI/article/view/11791
11. Dabney, W., Rowland, M., Bellemare, M.G., Munos, R.: Distributional Reinforcement Learning with Quantile Regression (Oct 2017). https://doi.org/10.48550/arXiv.1710.10044, http://arxiv.org/abs/1710.10044, arXiv:1710.10044 [cs, stat] version: 1
12. Distefano, C., Mindrila, D.: Cluster Analysis. In: Teo, T. (ed.) Handbook of Quantitative Methods for Educational Research, pp. 103–122. SensePublishers, Rotterdam (2013). https://doi.org/10.1007/978-94-6209-404-8_5, https://doi.org/10.1007/978-94-6209-404-8_5
13. Garcia, G.G.P., Steimle, L.N., Marrero, W.J., Sussman, J.B.: Interpretable Policies and the Price of Interpretability in Hypertension Treatment Planning. Manufacturing & Service Operations Management **26**(1), 80–94 (Jan 2024). https://

doi.org/10.1287/msom.2021.0373, https://pubsonline.informs.org/doi/10.1287/msom.2021.0373, publisher: INFORMS

14. Howard, R.A., Matheson, J.E.: Risk-Sensitive Markov Decision Processes. Management Science **18**(7), 356–369 (1972), https://www.jstor.org/stable/2629352, publisher: INFORMS
15. Kataria Golestaneh, A., Clarke, J.M., Appelbaum, N., Gonzalvez, C.R., Jose, A.P., Philip, R., Poulter, N.R., Beaney, T.: The factors influencing clinician use of hypertension guidelines in different resource settings: a qualitative study investigating clinicians' perspectives and experiences. BMC Health Services Research **21**(1), 767 (Dec 2021). https://doi.org/10.1186/s12913-021-06782-w, https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-021-06782-w
16. Kohli-Lynch, C.N., Bellows, B.K., Thanassoulis, G., Zhang, Y., Pletcher, M.J., Vittinghoff, E., et al.: Cost-effectiveness of Low-density Lipoprotein Cholesterol Level–Guided Statin Treatment in Patients With Borderline Cardiovascular Risk. JAMA Cardiology **4**(10), 969–977 (2019). https://doi.org/10.1001/jamacardio.2019.2851
17. Liu, Q., Li, Y., Shi, X., Lin, K., Liu, Y., Lou, Y.: Distributional Policy Gradient With Distributional Value Function. IEEE Transactions on Neural Networks and Learning Systems pp. 1–13 (2024). https://doi.org/10.1109/TNNLS.2024.3386225, https://ieeexplore.ieee.org/abstract/document/10508809, conference Name: IEEE Transactions on Neural Networks and Learning Systems
18. Ma, X., Xia, L., Zhou, Z., Yang, J., Zhao, Q.: DSAC: Distributional Soft Actor Critic for Risk-Sensitive Reinforcement Learning (Jun 2020). https://doi.org/10.48550/arXiv.2004.14547, http://arxiv.org/abs/2004.14547, arXiv:2004.14547 [cs]
19. Marthe, A., Garivier, A., Vernade, C.: Beyond Average Return in Markov Decision Processes. Advances in Neural Information Processing Systems **36**, 56488–56507 (Dec 2023)
20. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (Feb 2015). https://doi.org/10.1038/nature14236, https://www.nature.com/articles/nature14236, publisher: Nature Publishing Group
21. Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., Tanaka, T.: Nonparametric return distribution approximation for reinforcement learning. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. pp. 799–806. ICML'10, Omnipress, Madison, WI, USA (Jun 2010)
22. NCHS: Health, United States, 2016: with chartbook on long-term trends in health. Center for Disease Control pp. 314–317 (2017), https://www.cdc.gov/nchs/data/hus/hus16.pdf{#}019
23. Pfohl, S., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., Shah, N.H.: Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk (Jun 2019). https://doi.org/10.48550/arXiv.1809.04663, http://arxiv.org/abs/1809.04663, arXiv:1809.04663 [cs]
24. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons (Aug 2014), google-Books-ID: VvBjBAAAQBAJ
25. Rahimian, H., Mehrotra, S.: Distributionally Robust Optimization: A Review

26. Rowland, M., Bellemare, M., Dabney, W., Munos, R., Teh, Y.W.: An Analysis of Categorical Distributional Reinforcement Learning. In: Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. pp. 29–37. PMLR (Mar 2018), https://proceedings.mlr.press/v84/rowland18a.html, iSSN: 2640-3498
27. Sabater, C., Le Maître, O., Congedo, P.M., Görtz, S.: A Bayesian approach for quantile optimization problems with high-dimensional uncertainty sources. Computer Methods in Applied Mechanics and Engineering **376**, 113632 (Apr 2021). https://doi.org/10.1016/j.cma.2020.113632, https://www.sciencedirect.com/science/article/pii/S0045782520308173
28. Sobel, M.J.: The Variance of Discounted Markov Decision Processes. Journal of Applied Probability **19**(4), 794–802 (1982). https://doi.org/10.2307/3213832, https://www.jstor.org/stable/3213832, publisher: Applied Probability Trust
29. Sun, K., Zhao, Y., Liu, W., Jiang, B., Kong, L.: Sinkhorn Distributional Reinforcement Learning (Oct 2023), https://openreview.net/forum?id=aiPcdCFmYy
30. Sundström, J., Arima, H., Jackson, R., Turnbull, F., Rahimi, K., Chalmers, J., Woodward, M., Neal, B.: Effects of blood pressure reduction in mild hypertension: A systematic review and meta-analysis. Annals of Internal Medicine **162**(3), 184–191 (2015). https://doi.org/10.7326/M14-0773
31. Sundström, J., Arima, H., Woodward, M., Jackson, R., Karmali, K., Lloyd-Jones, D., Baigent, C., et al.: Blood pressure-lowering treatment based on cardiovascular risk: A meta-analysis of individual patient data. The Lancet **384**(9943), 591–598 (2014). https://doi.org/10.1016/S0140-6736(14)61212-5, http://dx.doi.org/10.1016/S0140-6736(14)61212-5
32. Sussman, J., Vijan, S., Hayward, R.: Using benefit-based tailored treatment to improve the use of antihypertensive medications. Circulation **128**(21), 2309–2317 (2013)
33. Varga, T.V.: Algorithmic fairness in cardiovascular disease risk prediction: overcoming inequalities. Open Heart **10**(2), e002395 (Nov 2023). https://doi.org/10.1136/openhrt-2023-002395, https://openheart.bmj.com/lookup/doi/10.1136/openhrt-2023-002395
34. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning **8**(3), 279–292 (May 1992). https://doi.org/10.1007/BF00992698, https://doi.org/10.1007/BF00992698
35. Wen, M., Bastani, O., Topcu, U.: Algorithms for Fairness in Sequential Decision Making. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021. vol. 130 (2021), https://github.com/wmgithub/fairness.
36. Whelton, P.K., Carey, R.M., Aronow, W.S., Casey, D.E., Collins, K.J., Dennison Himmelfarb, C., DePalma, S.M., Gidding, S., Jamerson, K.A., Jones, D.W., MacLaughlin, E.J., Muntner, P., Ovbiagele, B., Smith, S.C., Spencer, C.C., Stafford, R.S., Taler, S.J., Thomas, R.J., Williams, K.A., Williamson, J.D., Wright, J.T.: 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. Journal of the American College of Cardiology **71**(19), e127–e248 (May 2018). https://doi.org/10.1016/j.jacc.2017.11.006
37. Yadlowsky, S., Hayward, R.A., Sussman, J.B., McClelland, R.L., Min, Y.I., Basu, S.: Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. Annals of Internal Medicine **169**(1),

20 (2018). https://doi.org/10.7326/M17-3011, http://annals.org/article.aspx?doi=10.7326/M17-3011