Lorenz Conditioned Networks for Fair Multi-Objective Reinforcement Learning

Dimitris Michailidis[a,*], **Willem Röpke**[b], **Sennay Ghebreab**[a], **Diederik M. Roijers**[b,c] **and Fernando P. Santos**[a]

[a]Civic AI Lab, Socially-Intelligent Artificial Systems, Informatics Institute, University of Amsterdam
[b]AI Lab, Vrije Universiteit Brussel
[c]Urban Innovation and R&D, City of Amsterdam

Abstract. Multi-objective Reinforcement Learning (MORL) algorithms need to effectively scale to a large number of objectives to be practical for real-world applications. We introduce Lorenz-conditioned networks (LCNs), a novel multi-policy algorithm designed to provide Lorenz-optimal sets of policies, even when scaling up to 10 different objectives. LCN uses Lorenz optimality to learn policies that ensure a fair distribution of rewards among different objectives. Additionally, we address the lack of real-world MORL benchmarks, by introducing a large-scale, multi-objective environment for real-world transportation network design. Our experiments in the city of Xi'an in China demonstrate LCN's ability to learn fair policies in high-dimensional state-action and reward spaces.

1 Introduction

Reinforcement Learning (RL) is a powerful method for identifying optimal policies in sequential decision-making problems [49]. In these settings, agents learn to take actions in an environment in order to maximize expected long-term rewards [46].

Rewards are commonly formalized by combining different criteria (or objectives) into a scalar value [17]. Real-world problems can, however, involve multiple, often conflicting objectives. Formalizing a scalar reward before training can result in a biased decision-making process that overlooks desirable solutions that differ primarily in the weighting of the objectives [48].

To address this challenge, Multi-Objective Reinforcement Learning (MORL) uses vector-valued (instead of scalar) rewards [17]. Rather than learning a single policy maximizing a scalar reward, MORL learns multiple policies, which can later be used by decision-makers according to their preferences. MORL has revealed promising results in decision-making under unknown preferences [38, 3], human-value alignment [32, 37, 19], multi-agent games [42, 41], and others [48].

A key motivation for MORL is uncertainty about the decision-maker's utility function before training. Multi-policy methods, wherein agents learn a set of possibly optimal policies, alleviate this issue by assuming a monotonically increasing utility function and optimizing simultaneously for all objectives [26, 17, 34]. Optimizing for all possible utility functions, however, can be computationally intractable when the objective space is large [30]. Additionally, by not factoring in the result of combining objectives, multi-policy methods might waste resources exploring policy spaces leading to outcomes that, down the line, will be superfluous to a decision-maker.

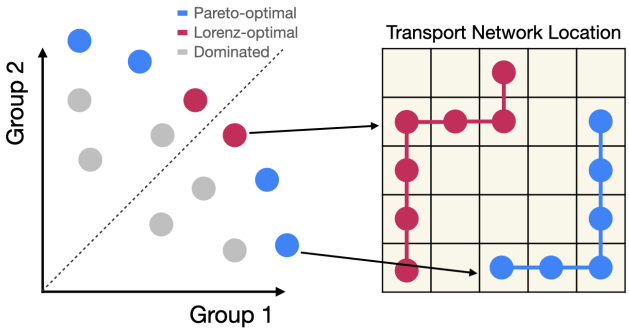

Figure 1: Lorenz-optimal policies lead to a more equal distribution of rewards across different objectives (i.e., closer to the diagonal in the example above). Pareto-optimal policies, instead, include policies that can lead to unequal rewards.

Algorithmic fairness suggests, precisely, problems, where some policies identified by MORL might be undesirable [20, 9] and ideally filtered out at training time. In domains where objectives represent the interests of different societal groups, policies leading to unfair distributions of rewards may generally be undesirable. We can thus take advantage of this knowledge and restrict the set of possible solutions, providing fairness guarantees to the decision-maker, and increasing computational tractability.

We present an example of such a problem in urban planning: designing new transportation lines. Reinforcement learning can be used to optimize the location of new transportation lines [52, 33, 28]. Planners must ensure that new lines are both efficient and that their benefits are fairly distributed among groups in a city [27]. Thus, a solution set that includes all optimal trade-offs (like the Pareto front), contains undesirable policies, whose distribution of rewards is disproportionate among different groups.

To solve this problem, we introduce **Lorenz Conditioned Networks (LCN)**, a multi-policy method that encourages fair policy discovery at training time, while alleviating the computation burden of multi-policy algorithms. LCN is based on Lorenz optimality, a refinement of Pareto optimality that ensures a fair distribution of rewards between objectives [44, 31]. As shown in Figure 1, Lorenz optimality leads to a subset of the Pareto front only composed of fair solutions.

Furthermore, we address the lack of real-world benchmarks for MORL [12] by introducing a new large-scale, multi-objective envi-

* Corresponding Author. Email: d.michailidis@uva.nl

ronment for designing transportation networks in real-world cities. Through experiments in Xi'an, we demonstrate that LCN can learn fair policy sets in large environments: as Lorenz-optimal sets tend to be smaller than Pareto-optimal ones [31], LCN shows great scalability, especially in high-dimensional objective spaces.

2 Related Work

Our work lies at the intersection of multi-policy methods for MORL [17] and algorithmic fairness in sequential decision-making problems [16].

2.1 Multi-Policy MORL

Several multi-policy methods have been proposed to tackle MORL problems, involving learning multiple policies that specialize in different trade-offs [34].

Initial multi-policy methods used Pareto Q-Learning and were limited to small-scale environments [29, 40]. To scale up multi-policy methods to high-dimensional state and action spaces, many works assume that decision-makers have linear preferences, resulting in a simpler solution set called the convex coverage set [1, 4, 34, 38, 39]. Our work does not make such an assumption and instead focuses on obtaining policies with a fairly distributed reward between the different objectives. In multi-objective optimization, GFlowNets-based methods have been proposed to generate diverse optimal solution candidates [21]. Our work relates to these as it can be used to generate diverse candidates; however, our model learns policies instead of solutions and thus can be used beyond static optimization problems (e.g. demand-responsive transport [5]). Lastly, the most relevant method to our approach are the so-called Pareto Conditioned Networks (PCNs) [34, 35, 11], a scalable approach to multi-policy learning. In this method, a single network is trained to learn all optimal policies in a Pareto front. Our method is inspired by PCNs, yet here we focus on fairness considerations, allowing scalability to higher dimensions while learning a set of fair policies. We demonstrate its effectiveness in the real-world problem of transportation network design.

2.2 Fairness in Reinforcement Learning

Research in Fairness in RL can be categorized along two main themes [16]: fairness in domains where individuals belong to protected groups (societal bias) and fairness in resource allocation problems (non-societal bias). Our work aligns more closely with the first theme, focusing on the fair distribution of benefits, such as public transport, among different societal groups.

Fairness inherently involves balancing multiple objectives. In this theme, many works combine diverse objectives into single fairness-based rewards. This is achieved through linear combinations [36, 8], generalizations to non-linear combinations, and welfare functions like the Generalized Gini Index [45, 18, 13], and other reward-shaping mechanisms [55, 53, 25, 22]. Alternatively, some methods adjust the reward function during training to adhere to fairness constraints [7]. These approaches require encoding the desirable principles into reward functions beforehand, whereas our approach makes no such assumptions.

Our work therefore relates closely to Cimpean et al. [9]'s formal MORL fairness framework, which encodes six fairness notions as objectives. They then train PCNs to identify Pareto-optimal trade-offs between these fairness notions. While our method can be applied within this framework, it does not pre-calculate any fairness notion. Instead, it optimizes for all group objectives directly, allowing the decision-maker to define their criteria after training.

3 Preliminaries

3.1 Multi-Objective Reinforcement Learning

When multiple optimal policies can exist, optimizing a single policy to maximize the expected return is not possible. Thus, Multi-Objective Reinforcement Learning problems are modeled as Multi-Objective Markov Decision Processes (MOMDPs) and are represented as a tuple $\boldsymbol{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \boldsymbol{\mathcal{R}}, \gamma \rangle$ consisting of a set of states $\mathcal{S}$, set of actions $\mathcal{A}$, transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, vector-based reward function $\boldsymbol{\mathcal{R}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, with $d \geq 2$ the number of objectives, and a discount factor γ.

When the reward is vector-based, a common solution set to optimize for is the *Pareto front*, a set of non-Pareto-dominated policies. Let $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^d$ be d dimensional vectors, and v_i the ith coordinate of $\boldsymbol{v}$, where $i \in \{1, 2, ..., d\}$. We say that vector $\boldsymbol{v}$ Pareto dominates vector $\boldsymbol{v}'$, if and only if [17]:

$$\boldsymbol{v} \succ_P \boldsymbol{v}' \iff (\forall i : v_i \geq v'_i) \wedge (\exists i : v_i > v'_i) \tag{1}$$

In essence, $\boldsymbol{v}$ Pareto dominates $\boldsymbol{v}'$ when it is at least equal for all objectives and better in at least one. The Pareto front contains vectors that cannot be improved on one objective without deteriorating another. When objectives correspond to outcomes for different groups, not all vectors in the Pareto front adhere to fairness criteria.

3.2 Fairness in MOMDPs

Various definitions of fairness exist in decision-making and machine learning [6], aiming at encapsulating notions of fairness through a single metric. However, this approach requires specifying preferences over objectives, a challenge not commonly tackled in MOMDPs.

Here, we focus on the distribution of the benefits of a decision among individuals or groups, each represented by a different objective. While this introduces a general preference for fairness, it does not explicitly encode it into a scalarization function (e.g., relative weights are not set).

As shown in Figure 1, Pareto dominance may lead to sets that include policies with undesired outcomes in terms of fairness. Therefore, our methods rely on Lorenz dominance, a refinement of Pareto dominance that considers the distribution of values within the vector [31], and has traditionally been used in economics to assess income inequality [44]. First, we define a Lorenz vector $L(\boldsymbol{v})$ of a vector $\boldsymbol{v}$ as $L(\boldsymbol{v}) = \left(v_{(1)}, v_{(1)} + v_{(2)}, ..., \sum_{i=1}^{d} v_{(i)}\right)$, where $v_{(1)} \leq v_{(2)} \leq ... \leq v_{(d)}$ are the values of the vector $\boldsymbol{v}$, sorted in increasing order. Lorenz dominance is equivalent to the Pareto dominance of one Lorenz vector $L(\boldsymbol{v})$ over another Lorenz vector $L(\boldsymbol{v}')$ [31, 17].

$$\begin{aligned} \boldsymbol{v} \succ_L \boldsymbol{v}' &\iff L(\boldsymbol{v}) \succ_P L(\boldsymbol{v}') \\ &\iff (\forall i : L(\boldsymbol{v})_i \geq L(\boldsymbol{v}')_i) \wedge (\exists i : L(\boldsymbol{v})_i > L(\boldsymbol{v}')_i) \end{aligned} \tag{2}$$

In MORL, we define vectors $\boldsymbol{v}^{\pi}, \boldsymbol{v}^{\pi'}$ as the expected return of the policies π, π', across all objectives of the environment respectively. Therefore, a policy will Lorenz dominate another policy, if and only if its Lorenz vector $L(\boldsymbol{v}^{\pi})$ is greater than or equal on all objectives

than $L(\boldsymbol{v}^{\pi'})$, and there exists at least one objective where it is better [17].

Lorenz-Conditioned Networks search for non-Lorenz dominated policies that form the Lorenz set. The set of non-dominated value vectors is called a *coverage set*, wherein the optimal coverage set is represented by the Pareto front [17]. A Lorenz coverage set is usually (but not necessarily) significantly smaller than a Pareto coverage set [31].

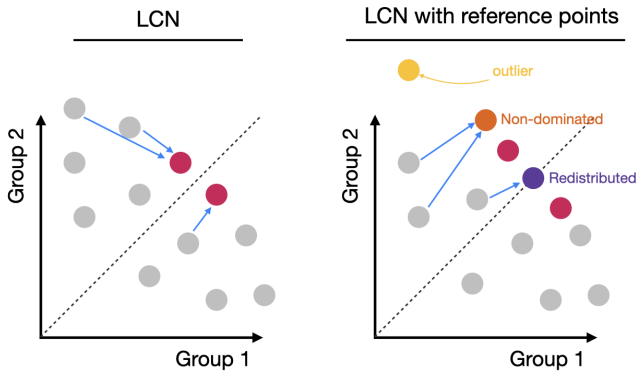

Figure 2: We introduce Lorenz Conditioned Networks (LCNs), a multi-policy method that offers fair trade-offs to decision makers (left). We also introduce reference points for enhancing the training process by filtering the Experience Replay buffer (right).

4 Methodology

We introduce Lorenz-Conditioned Networks (LCN), a single, multi-policy model for fair policy search. Here we describe its main components and introduce the LCN reference points extension (Section 4.4). Algorithm 1 shows an overview of how LCN works.

4.1 Reward Conditioned Networks

LCNs fall under the category of conditioned networks, where a single neural network is trained through supervised learning to learn multiple policies [23, 34]. These policies map states and desired rewards to probability distributions over actions, by collecting experiences, using actions as labels and state/rewards as input.

Specifically, LCN learns policies $\pi_\theta(a_t|s_t, \hat{h}_t, \hat{\boldsymbol{G}}_t)$, where a_t is the next action the agent will take, s_t is the current state, $\hat{h}_t$ is the desired horizon (steps until the episode ends — this is for environments with varied episode length) and $\hat{\boldsymbol{G}}_t$ is the desired return of the episode, starting from time step t. Note that $\hat{\boldsymbol{G}}_t$ is a vector, with dimension d equal to the number of objectives.

The network maps an input tuple $\langle s_t, \hat{h}_t, \hat{\boldsymbol{G}}_t \rangle$ to an output probability distribution over the next actions $\pi_\theta(a_t|s_t, \hat{h}_t, \hat{\boldsymbol{G}}_t)$. Assuming the agent takes discrete actions, the training process is similar to multi-label classification, with the actions as the classes. The network updates its parameters using a cross-entropy loss function and the Adam optimizer, in accordance with previous works [23, 34].

$$H = -\sum_{a \in A} y_a \log \pi(a_t|s_t, h_t, \boldsymbol{G}_t) \tag{3}$$

where $y_{a,t} = 1$ if $a_t = a$ and 0 otherwise. Note that in Equation 3 we do not use the desired horizon $\hat{h}_t$ and return $\hat{\boldsymbol{G}}_t$. That is because supervised learning occurs on previously collected experiences (explained in Section 4.2). The agent learns independently without imitation learning or expert trajectories; it relies on its own collected experiences. Therefore, it learns sub-optimal policies, which however are optimal for the conditioned desired return $\hat{\boldsymbol{G}}_t$ [23]. Essentially, given a sufficient number of good experiences, the agent will learn good policies. Neural network capabilities can lead to generalization to better policies by modifying the condition [34].

4.2 Constructing the Experience Replay Buffer

Conditioned networks follow a training process to learn a policy π_θ, through two alternating steps: collecting and storing experiences in the Experience Replay (ER) buffer by interacting with the environment and training the policy on previously collected experiences using supervised learning [23, 34]. They are thus off-policy methods, with π_θ starting in a random state.

The experiences are stored as $\langle s_t, \hat{h}_t, \hat{\boldsymbol{G}}_t \rangle$ in the ER buffer. The quality of trajectories and collected experiences is crucial for the effectiveness of supervised learning training. To ensure policy improvement, the ER buffer is constantly updated, with new and better experiences replacing older ones.

The primary mechanism for achieving this improvement in conditioned networks is through continuous improvement of the condition. This is achieved by randomly selecting a non-dominated return from the current coverage set and increasing its value by a sample from a uniform distribution $U(0, \sigma_o)$, where σ_o represents the standard deviation of all non-dominated points in the ER buffer [34]. The resulting updated return is then used as the input $\hat{\boldsymbol{G}}_{\boldsymbol{t}}$ in the policy network. The next step to improve the buffer is its filtering mechanisms.

Algorithm 1 LCN Algorithm

$\theta \leftarrow$ random initial parameters
$\mathcal{B} \leftarrow$ sample `BufferSize` random $< s_t, a_t, h_t, G_t >$
$\mathcal{B}_{\mathcal{ND}} \leftarrow$ non-Lorenz-dominated $G_t \in \mathcal{B}$
for *step* $\leq$ `TotalTimesteps` **do**
  $\hat{\boldsymbol{G}}_t \leftarrow G_t + U(0, \sigma_0)$ ▷ G_t sampled from $\mathcal{B}_{\mathcal{ND}}$
  Generate episodes $\{s_t, a_t, h_t, G_t\}_{t=0}^{T}$ using $\pi_\theta(\hat{\boldsymbol{G}}_t, \hat{h}_t)$
  for timestep t **do**
    Add $< s_t, a_t, h_t, G_t >$ to $\mathcal{B}$
    Filter $\mathcal{B}$ according to Equation 4
    $\mathcal{B}_{\mathcal{ND}} \leftarrow$ non-Lorenz-dominated $G_t \in \mathcal{B}$
  end for
  Train Model after every `TrainModelSteps`
  Sample batch $b \sim \mathcal{B}$ of size `BufferSize`, predictions $\pi(a_t|s_t, h_t, \boldsymbol{G}_t)$
  Compute H_b using Equation 3
  $\theta \leftarrow \theta - \alpha \nabla_\theta H_b$
end for

Return Trained parameters θ

4.3 Filtering experiences via Lorenz-dominance

During exploration, new experiences replace older ones via a distance-based mechanism. In a previous method (PCNs), experiences in the buffer are evaluated based on their distance from the closest non-Pareto-dominated point in the buffer. In addition, a *crowding distance* is calculated for each point, measuring its distance

to its closest neighbors [10]. Points that are too close to other neighbors have a large crowding distance and are penalized, increasing the likelihood of being replaced compared to sparser points. This ensures that ER experiences are distributed across the objective space, facilitating faster attainment of Pareto coverage sets [34].

Lorenz Conditioned Networks (LCN), on the other hand, seek fairly distributed policies using Lorenz dominance. Thus, the evaluation of each experience e_i in the Experience Replay buffer $\mathcal{B}$ is determined by its proximity to the nearest *non-Lorenz dominated* point $l_j \in \mathcal{L}(\mathcal{B}) \subseteq \mathcal{B}$. In Figure 2 (left) we show an example of this distance calculation.

We denote the distance between an experience e_i and a reference point t_i as $d_{e_i,t_i} = ||e_i - t_i||_2$, where $t_i = \arg\min ||e_i - l_j||_2$ is the nearest non-Lorenz-dominated point, to be referred to as *reference point*. In Section 4.4 we devise mechanisms for reference points. We formalize the final distance for the evaluation with the dominance score $ds_{\text{Lorenz},i}$ as follows:

$$ds_{\text{Lorenz},i} = \begin{cases} d_{e_i,t_i} & \text{if } d_{cd,i} > \tau_{cd} \\ 2(d_{e_i,t_i} + c) & \text{if } d_{cd,i} \leq \tau_{cd} \end{cases} \tag{4}$$

Where d_{cd} is the crowding distance of i and τ_{cd} is the crowding distance threshold. A constant penalty c is added to the points below the threshold, whose distance is also doubled [34]. The points in the ER buffer are sorted based on ds_{Lorenz} and those with the highest get replaced first when a better experience is collected.

4.4 Improving Training with Reference Points

The nearest-point filtering method has two drawbacks. Firstly, during exploration, stored experiences undergo significant changes as the agent discovers new and improved trajectories. This leads to a volatile ER buffer and moving targets, posing stability challenges during supervised learning. Secondly, given fairness considerations, it is known in advance that certain experiences, even if non-dominated, are undesirable due to their unfair distribution of rewards.

Consider, for example, vectors: $\boldsymbol{v} = (8, 0), \boldsymbol{w} = (3, 4)$ and their corresponding Lorenz vectors $L(\boldsymbol{v}) = (0, 8), L(\boldsymbol{w}) = (3, 7)$. Both $\boldsymbol{v}$ and $\boldsymbol{w}$ are non-Lorenz dominated, and would typically be used as targets for evaluating other experiences. However, $\boldsymbol{v}$ is not a desirable target due to its unfair distribution of rewards (this is essentially a limitation of Lorenz dominance, when one objective is very large). To address these issues, we propose leveraging fairness constraints to define reference points for distance calculations. We introduce two reference point mechanisms: a **redistribution** mechanism and a **mean** reference point mechanism.

4.4.1 Redistributed Reference Point (LCN-Redist)

This mechanism draws inspiration from the Pigou-Dalton principle in welfare economics [2]. Consider a vector $\boldsymbol{v} \in \mathbb{R}^d$, where $v_i > v_j$ for some i, j. Then, for any ϵ, $0 < \epsilon \leq v_i - v_j$, the vector $\boldsymbol{v}' = \boldsymbol{v} - \epsilon I_{v_i} + \epsilon I_{v_j}$, where I_{v_i} and I_{v_j} are vectors with 1 at the ith and jth elements, respectively, and 0 elsewhere is preferable to $\boldsymbol{v}$, because the transfer resulted in a more desirable distribution of rewards while maintaining the total sum at the same level [31]. In the example we provided above, given $\epsilon = 4$, $\boldsymbol{v}' = (4, 4)$ is more desirable than $\boldsymbol{v} = (8, 0)$, while the sum of rewards remains the same.

Under this axiomatic principle, any experience in the ER buffer can be adjusted to provide a more desirable one. We employ this transfer mechanism to create the most desirable out of all collected experiences by identifying the one with the highest sum of rewards and performing a transfer, evenly distributing the total reward across all dimensions of the vector. This is then assigned as the new *reference point* t for all experiences $e \in \mathcal{B}$:

$$t_{\text{redist}} = \frac{1}{n} \left(\arg\max_{e \in D} \sum_{j=1}^{n} e_j \right) \mathbf{1}, \tag{5}$$

where $\mathbf{1}$ is a vector of ones with dimension d. Note that t is now the same for all $e \in \mathcal{B}$. Subsequently, we measure the distances of all $e \in \mathcal{B}$ to this newly defined reference point and filter out those that are farthest from it, according to Equation 4 (replace t_i with t_{redist}). In Figure 2 (right), we illustrate this transfer mechanism. As the collected experiences get better via exploration and conditioning, so does the reference point t_{redist}, which ensures an equal distribution over the rewards. This in time leads to the promotion of efficient and fair experiences.

4.4.2 Non-Dominated Mean Reference Point (LCN-Mean)

The redistributed reference point we outlined in Section 4.4.1 is generated by focusing solely on experience with the highest sum of rewards. However, this approach disregards other experiences, leading to the creation of unrealistic reference points, which may introduce bias in favor of utopic solutions, disadvantaging other viable options that, while not perfectly equal, are still sufficiently good.

To address this limitation, we propose an alternative reference point mechanism: a straightforward averaging of all non-Lorenz dominated vectors in the experience replay (ER) buffer. This approach provides a simpler and more targeted method for incorporating collected experiences, while simultaneously smoothing out outlier non-dominated points. The reference point, denoted as t_{mean}, is defined as follows: Let $\mathcal{L}(D) = \{l_1, l_2, \ldots, l_j\}$ represent the set of non-Lorenz dominated experiences in the ER buffer,

$$t_{\text{mean}} = \frac{1}{|\mathcal{L}(D)|} \sum_{l_j \in \mathcal{L}(D)} l_j \tag{6}$$

In Figure 2 (right) we show how this approach defines a reference point. It is important to emphasize here that both mechanisms outlined above are used for defining reference points, not the conditioned return the model is trained on, which continues to be the collected experiences.

5 Experiment Setup

Existing MORL benchmarks are often small-scale, with small state-action spaces or low-dimensional objective spaces [47, 24]. We introduce a novel and modular MORL environment, named the Multi-Objective Transport Network Design Problem (MO-TNDP). MO-TNDP is inspired by the real-world challenge of transportation design, with a large state-action space and adaptability to high-reward dimensions.

5.1 Multi-Objective Transport Network Design Problem (MO-TNDP)

Built on MO-Gymnasium [15], the MO-TNDP environment [1] simulates the design of public transportation networks in cities of varying size and morphology, addressing TNDP, an NP-hard optimization

[1] https://github.com/sias-uva/mo-tndp

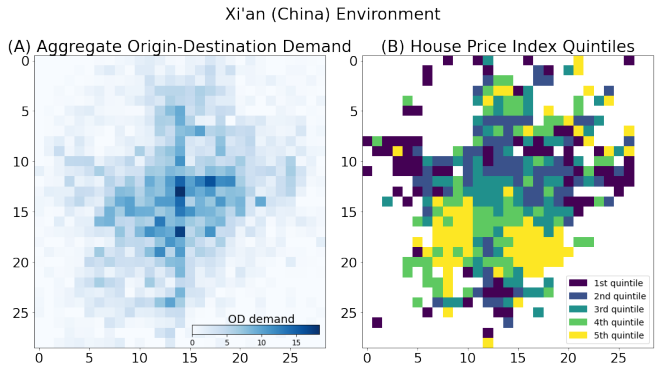

Figure 3: An instance of the MO-TNDP environment for designing transportation networks in real-world cities, illustrated using data from the city of Xi'an, China, provided by Wei et al. [52]. Panel A displays the aggregate Origin-Destination Demand per cell, representing the sum of all incoming and outgoing flows. Panel B categorizes each cell into quintiles based on the house price index, showing the group membership of each cell.

problem aiming to generate a transportation line that maximizes the satisfied travel demand [14].

In MO-TNDP, a city is represented as $H^{m\times n}$, a grid with equally sized cells. The mobility demand forecast between cells is captured by an Origin-Destination (OD) flow matrix $OD^{|H|\times|H|}$. Each cell $h \in H^{n\times m}$ is associated with a socioeconomic group $g \in \mathcal{G}$, defined by indicators such as income or the development index, which determine the dimensionality of the reward function. In this paper, we scale it from 2 to 10 groups (objectives).

Episodes start in a specified (or random/learned) cell and last a predefined number of steps. A single agent traverses the city, connecting grid cells with eight available actions (movement to the neighbor cell in all directions). At each time step, the agent receives a vectorial reward with dimension $d = |\mathcal{G}|$, each corresponding to the percent satisfied demand of each group. We formulate it as an MOMDP $\boldsymbol{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \boldsymbol{\mathcal{R}}, \gamma \rangle$, where $\mathcal{S}$ is the current location of the agent (grid cell), $\mathcal{A}$ is the next direction of movement and $\boldsymbol{\mathcal{R}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is the additional demand satisfied by taking the last action (connecting previously visited cells) for each group. Given the discrete, episodic nature the discount factor $\gamma = 1$. The transition function $\mathbf{P}$ is deterministic and each episode starts from the same state. Note that MO-TNDP is deterministic, but LCN can also be used in stochastic environments.

Additional directional constraints can be imposed on the agent action space. The environment code enables developers to modify the city object, incorporating adjustments to grid size, OD matrix, cell group membership, and directional constraints, making it adaptable to any city. It supports both creating new transportation networks and expanding existing ones, identifying connections, and calculating the additional satisfied demand of new lines.

Here, we focus on the city of Xi'an in China, in a grid that contains a total of 841 cells. The agent is constrained to design typical metro lines, avoiding circular directions. Group membership for each cell is determined by the average house price, which is divided into 2-10 equally sized buckets. Figure 3 illustrates an instance for Xi'an for five objectives. Note that MO-TNDP is a single-agent environment, though our method can be adapted to multi-agent environments where rewards represent payoffs of different agents.

Through Bayesian hyperparameter search of 100 runs, we tuned the batch size, learning rate, ER buffer size, number of layers, and hidden dimension across all reported models, environments, and objective dimensions. We compare LCN with the baseline PCN [34]. We provide the code of the repository, which contains details of the hyperparameters for each different setting [2].

5.2 *Evaluation*

Hypervolume: a widely used metric in multi-objective decision-making [50, 51, 34], assesses the quality of a set of non-dominated solutions by considering its closeness to the Pareto front, diversity, and spread. This metric measures the volume of a set of points relative to a specific reference point and is maximized for the Pareto front.

$$\texttt{HV}(CS, \boldsymbol{V}_{ref}) = \bigcup_{\pi \in CS} \texttt{Volume}(\boldsymbol{V}_{ref}, \boldsymbol{V}^{\pi}), \tag{7}$$

where $\texttt{Volume}(\boldsymbol{V}_{ref}, \boldsymbol{V}^{\pi})$ is the volume of the hypercube spanned by a reference vector $\boldsymbol{V}_{ref}$ and the coverage set vector $\boldsymbol{V}^{\pi}$ [17].

Sen Welfare: a welfare function that combines total efficiency and equality into a single measure. Total efficiency is the cumulative satisfied demand across all groups, while equality is expressed through the Gini coefficient—a quantification of the Lorenz curve that measures reward distribution among groups (with 0 indicating perfect equality and 1 perfect inequality) [43]. The final score is measured by multiplying total efficiency by $(1 - \text{Gini Index})$.

$$\texttt{SW}(\pi) = \sum_i \boldsymbol{v}_i^{\boldsymbol{\pi}} (1 - \texttt{GI}(v^{\pi})), \tag{8}$$

Where $\sum_i \boldsymbol{v}_i^{\boldsymbol{\pi}}$ is the sum of the returns of all objectives in policy π, and $\texttt{GI}(v^{\pi})$ is the Gini index of the return of policy π. Note that here we introduce a welfare function for comparative purposes, reflecting a balanced scenario where both efficiency and equality are considered. Sen welfare has been utilized in economic simulations employing Reinforcement Learning techniques before [54]. However, alternative welfare functions can be employed to better align with the decision-maker's preferences. A higher Sen welfare value signifies increased satisfaction of total demand and a more equitable distribution among groups.

6 Results

The results discussed in this section are based on 5 different random seeded runs. Figure 4 presents a comparison between PCN and LCN in Hypervolume and Sen Welfare across all objectives within the MO-TNDP-Xi'an environment. Table 1 displays the (normalized for each objective) Sen Welfare results for all proposed models across all objectives (both environments). Comprehensive results for all evaluation metrics and environments are available in Table 1.

6.1 *LCN outperforms PCN on high-dimensional reward spaces*

As shown in Figure 4, PCN shows strong performance on hypervolume, outperforming our proposed LCN model in scenarios with 2–4 objectives. This is expected, as PCN is designed to learn diverse, non-Pareto-dominated solutions, which maximize hypervolume. However, for larger objective spaces (more than 5), LCN surpasses PCN even in hypervolume, a metric it is not specifically designed for. This occurs because, given the high state-action space of the environment,

[2] https://github.com/sias-uva/mo-transport-network-design

Table 1: Results of all models, for 1–10 objectives. Underline indicates the best results.

Normalized Hypervolume									
	Number of Objectives								
Xi'an	2	3	4	5	6	7	8	9	10
PCN	$\underline{0.92 \pm 0.02}$	$\underline{0.89 \pm 0.03}$	$\underline{0.63 \pm 0.10}$	$\underline{0.57 \pm 0.11}$	0.38 ± 0.09	0.11 ± 0.05	0.02 ± 0.01	0.00 ± 0.00	0.00 ± 0.00
LCN	0.81 ± 0.06	0.54 ± 0.02	0.62 ± 0.02	0.46 ± 0.07	0.50 ± 0.15	$\underline{0.69 \pm 0.08}$	$\underline{0.76 \pm 0.07}$	$\underline{0.70 \pm 0.09}$	$\underline{0.71 \pm 0.16}$
LCN-Redist.	0.76 ± 0.03	0.44 ± 0.03	0.50 ± 0.06	$\underline{0.57 \pm 0.10}$	$\underline{0.86 \pm 0.04}$	0.44 ± 0.16	0.65 ± 0.05	0.39 ± 0.10	0.53 ± 0.22
LCN-Mean	0.77 ± 0.03	0.43 ± 0.03	0.44 ± 0.03	0.32 ± 0.10	0.57 ± 0.04	0.32 ± 0.11	0.38 ± 0.06	0.36 ± 0.13	0.33 ± 0.13
Normalized Sen Welfare									
Xi'an	2	3	4	5	6	7	8	9	10
PCN	0.81 ± 0.02	0.70 ± 0.01	0.64 ± 0.01	0.63 ± 0.01	0.57 ± 0.01	0.49 ± 0.01	0.29 ± 0.01	0.32 ± 0.01	0.35 ± 0.01
LCN	0.84 ± 0.03	0.89 ± 0.01	0.88 ± 0.01	0.78 ± 0.04	0.67 ± 0.04	0.76 ± 0.01	$\underline{0.86 \pm 0.01}$	$\underline{0.82 \pm 0.01}$	$\underline{0.93 \pm 0.01}$
LCN-Redist.	$\underline{0.94 \pm 0.01}$	$\underline{0.91 \pm 0.01}$	$\underline{0.93 \pm 0.01}$	0.80 ± 0.01	0.73 ± 0.01	$\underline{0.77 \pm 0.02}$	0.60 ± 0.01	0.50 ± 0.01	0.55 ± 0.02
LCN-Mean	0.93 ± 0.02	0.85 ± 0.03	0.75 ± 0.02	$\underline{0.84 \pm 0.03}$	$\underline{0.84 \pm 0.03}$	0.71 ± 0.01	0.79 ± 0.01	$\underline{0.82 \pm 0.02}$	0.78 ± 0.03
Gini Index (the lower the better)									
Xi'an	2	3	4	5	6	7	8	9	10
PCN	0.13 ± 0.02	0.23 ± 0.01	0.32 ± 0.01	0.32 ± 0.01	0.36 ± 0.01	0.37 ± 0.01	0.45 ± 0.01	0.45 ± 0.00	0.49 ± 0.01
LCN	0.09 ± 0.02	0.09 ± 0.01	0.12 ± 0.01	0.17 ± 0.02	0.23 ± 0.02	0.28 ± 0.01	$\underline{0.23 \pm 0.01}$	$\underline{0.21 \pm 0.01}$	$\underline{0.19 \pm 0.01}$
LCN-Redist.	$\underline{0.04 \pm 0.01}$	$\underline{0.04 \pm 0.01}$	$\underline{0.08 \pm 0.01}$	$\underline{0.15 \pm 0.01}$	$\underline{0.20 \pm 0.01}$	$\underline{0.20 \pm 0.01}$	0.28 ± 0.01	0.29 ± 0.01	0.35 ± 0.01
LCN-Mean	0.07 ± 0.01	0.10 ± 0.02	0.20 ± 0.01	$\underline{0.15 \pm 0.01}$	$\underline{0.21 \pm 0.02}$	0.29 ± 0.01	$\underline{0.21 \pm 0.01}$	0.23 ± 0.01	0.25 ± 0.02

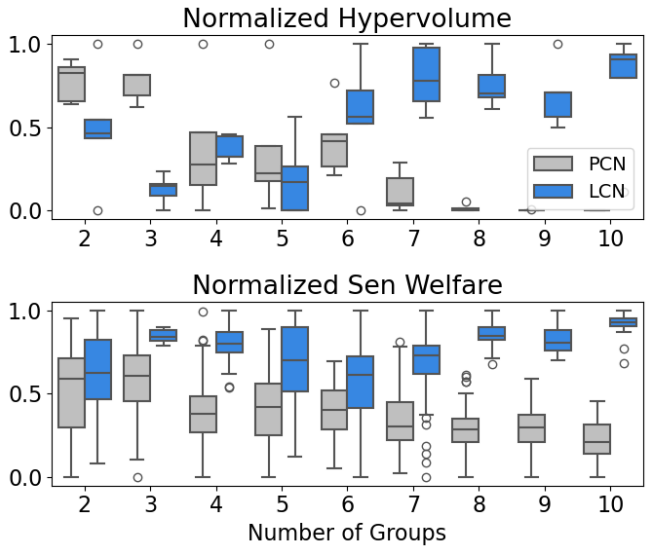

Figure 4: The proposed LCN (no reference points) outperforms PCN across all objectives in the Sen Welfare measure (Xi'an environment). Additionally, LCN outperforms PCN in hypervolume when the number of objectives > 4, showcasing the scalability of LCN in contrast with the limitations of PCN.

the non-Pareto-dominated solutions significantly expand with objectives, making the supervised training of PCN challenging. Furthermore, we observe that for objectives exceeding 7, PCN collapses, failing to achieve good diverse policies. In contrast, LCN effectively scales across the objective space, leveraging the smaller non-Lorenz-dominated set.

LCN consistently outperforms PCN on Sen Welfare across all objectives. The Sen Welfare metric, which promotes solutions balancing efficiency and equality, shows that LCN excels in generating effective policies even when the solution space is constrained. Notably, LCN maintains its superior performance relative to PCN even as the number of objectives increases. LCN additionally outperforms PCN on total efficiency and the Gini coefficient of the proposed policies, as shown in Table 1.

6.2 Reference points can improve training

In Table 1, we compare the reference point mechanisms (LCN-Redist and LCN-Mean) to PCN and LCN across all numbers of objectives on Hypervolume, Sen Welfare and Gini Index. Both reference point mechanisms outperform PCN and mostly outperform LCN.

LCN-Redist is effective when the reward space is small, as redistributing the best over a few objectives leads to a reference point that is closer to the rest of the ER buffer. However, its stability diminishes with an increasing number of objectives, showing very high variance.

LCN-Mean performs great in Sen Welfare across all objectives, offering better stability than LCN-Redist. This result is expected, as LCN is designed to maximize Sen Welfare. LCN-Mean effectively balances outliers and creates reference points that balance efficiency and equality with minimal intervention.

7 Conclusion

We introduced LCN, a novel model to nudge fair solutions, at training time, in multi-objective reinforcement learning. This is accomplished via a single network to generate multiple policies. We developed a new, multi-objective environment for simulating Public Transport Network Design, thereby enhancing the applicability of MORL to real-world scenarios. Our findings demonstrate that LCN outperforms the baseline PCN in fairness and, furthermore, surpasses PCN in hypervolume when the objective space increases. These contributions move the research field toward more realistic and applicable solutions in real-world contexts, thereby advancing the state-of-the-art in algorithmic fairness in sequential decision-making and MORL.

Acknowledgements

DM is supported by the Innovation Center for AI (ICAI, The Netherlands) and the City of Amsterdam. WR is supported by the Research Foundation – Flanders (FWO), grant number 1197622N. This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

References

[1] A. Abels, D. Roijers, T. Lenaerts, A. Nowé, and D. Steckelmacher. Dynamic Weights in Multi-Objective Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 11–20. PMLR, May 2019. URL https://proceedings.mlr.press/v97/abels19a.html. ISSN: 2640-3498.

[2] M. D. Adler. The Pigou-Dalton Principle and the Structure of Distributive Justice, May 2013. URL https://papers.ssrn.com/abstract=2263536.

[3] L. N. Alegre, A. Bazzan, and B. C. D. Silva. Optimistic Linear Support and Successor Features as a Basis for Optimal Policy Transfer. In *Proceedings of the 39th International Conference on Machine Learning*,

pages 394–413. PMLR, June 2022. URL https://proceedings.mlr.press/v162/alegre22a.html. ISSN: 2640-3498.

[4] L. N. Alegre, A. L. C. Bazzan, D. M. Roijers, A. Nowé, and B. C. da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 2003–2012, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

[5] M. J. Alonso-González, T. Liu, O. Cats, N. Van Oort, and S. Hoogendoorn. The Potential of Demand-Responsive Transport as a Complement to Public Transport: An Assessment Framework and an Empirical Evaluation. *Transportation Research Record*, 2672(8):879–889, Dec. 2018. ISSN 0361-1981. doi: 10.1177/0361198118790842. URL https://doi.org/10.1177/0361198118790842. Publisher: SAGE Publications Inc.

[6] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[7] J. Chen, Y. Wang, and T. Lan. Bringing Fairness to Actor-Critic Reinforcement Learning for Network Utility Optimization. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, May 2021. doi: 10.1109/INFOCOM42981.2021.9488823. URL https://ieeexplore.ieee.org/abstract/document/9488823. ISSN: 2641-9874.

[8] X. Chen, T. Wang, B. W. Thomas, and M. W. Ulmer. Same-day delivery with fair customer service. *European Journal of Operational Research*, 308(2):738–751, July 2023. ISSN 0377-2217. doi: 10.1016/j.ejor.2022.12.009. URL https://www.sciencedirect.com/science/article/pii/S0377221722009365.

[9] A. Cimpean, C. Jonker, P. Libin, and A. Nowé. A Multi-objective Framework For Fair Reinforcement Learning. 2023.

[10] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II. In M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature PPSN VI*, Lecture Notes in Computer Science, pages 849–858, Berlin, Heidelberg, 2000. Springer. ISBN 978-3-540-45356-7. doi: 10.1007/3-540-45356-3_83.

[11] F. Delgrange and M. Reymond. WAE-PCN: Wasserstein-autoencoded Pareto Conditioned Networks. 2023. URL https://cris.vub.be/ws/portalfiles/portal/95875411/WAE_PCN_3.pdf.

[12] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, Sept. 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05961-4. URL https://doi.org/10.1007/s10994-021-05961-4.

[13] Z. Fan, N. Peng, M. Tian, and B. Fain. Welfare and Fairness in Multi-objective Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1991–1999, Richland, SC, May 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-9432-1.

[14] R. Z. Farahani, E. Miandoabchi, W. Y. Szeto, and H. Rashidi. A review of urban transportation network design problems. *European journal of operational research*, 229(2):281–302, 2013.

[15] F. Felten, L. N. Alegre, A. Nowé, A. L. C. Bazzan, E. G. Talbi, G. Danoy, and B. C. da. Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[16] P. Gajane, A. Saxena, M. Tavakol, G. Fletcher, and M. Pechenizkiy. Survey on Fair Reinforcement Learning: Theory and Practice, May 2022. URL http://arxiv.org/abs/2205.10032. arXiv:2205.10032 [cs].

[17] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, Apr. 2022. ISSN 1573-7454. doi: 10.1007/s10458-022-09552-y. URL https://doi.org/10.1007/s10458-022-09552-y.

[18] X. Hu, Y. Zhang, H. Xia, W. Wei, Q. Dai, and J. Li. Towards Fair Power Grid Control: A Hierarchical Multi-Objective Reinforcement Learning Approach. *IEEE Internet of Things Journal*, pages 1–1, 2023. ISSN 2327-4662. doi: 10.1109/JIOT.2023.3314522. URL https://ieeexplore.ieee.org/abstract/document/10247509. Conference Name: IEEE Internet of Things Journal.

[19] M. Hwang, L. Weihs, C. Park, K. Lee, A. Kembhavi, and K. Ehsani. Promptable Behaviors: Personalizing Multi-Objective Rewards from Human Preferences, Dec. 2023. URL http://arxiv.org/abs/2312.09337. arXiv:2312.09337 [cs].

[20] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in reinforcement learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1617–1626. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jabbari17a.html.

[21] M. Jain, S. C. Raparthy, A. Hernández-Garcıa, J. Rector-Brooks, Y. Bengio, S. Miret, and E. Bengio. Multi-Objective GFlowNets. In *Proceedings of the 40th International Conference on Machine Learning*, pages 14631–14653. PMLR, July 2023. URL https://proceedings.mlr.press/v202/jain23a.html. ISSN: 2640-3498.

[22] A. Kumar and W. Yeoh. Fairness in Scarce Societal Resource Allocation: A Case Study in Homelessness Applications. 2023.

[23] A. Kumar, X. B. Peng, and S. Levine. Reward-Conditioned Policies, Dec. 2019. URL http://arxiv.org/abs/1912.13465. arXiv:1912.13465 [cs, stat].

[24] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner. Microscopic Traffic Simulation using SUMO. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582, Nov. 2018. doi: 10.1109/ITSC.2018.8569938. URL https://ieeexplore.ieee.org/document/8569938. ISSN: 2153-0017.

[25] D. Mandal and J. Gan. Socially Fair Reinforcement Learning, Feb. 2023. URL http://arxiv.org/abs/2208.12584. arXiv:2208.12584 [cs].

[26] P. Mannion, F. Heintz, T. G. Karimpanal, and P. Vamplew. Multi-Objective Decision Making for Trustworthy AI. 2021.

[27] K. Martens. *Transport Justice: Designing fair transportation systems*. Routledge, July 2016. ISBN 978-1-317-59958-6. Google-Books-ID: m0yTDAAAQBAJ.

[28] D. Michailidis, S. Ghebreab, and F. P. Santos. Balancing Fairness and Efficiency in Transport Network Design through Reinforcement Learning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, London, United Kingdom, May 2023. IFAAMAS.

[29] K. V. Moffaert and A. Nowé. Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies. *Journal of Machine Learning Research*, 15(107):3663–3692, 2014. ISSN 1533-7928. URL http://jmlr.org/papers/v15/vanmoffaert14a.html.

[30] T. T. Nguyen, N. D. Nguyen, P. Vamplew, S. Nahavandi, R. Dazeley, and C. P. Lim. A multi-objective deep reinforcement learning framework. *Engineering Applications of Artificial Intelligence*, 96:103915, Nov. 2020. ISSN 0952-1976. doi: 10.1016/j.engappai.2020.103915. URL https://www.sciencedirect.com/science/article/pii/S0952197620302475.

[31] P. Perny, P. Weng, J. Goldsmith, and J. Hanna. Approximation of Lorenz-Optimal Solutions in Multiobjective Markov Decision Processes, Sept. 2013. URL http://arxiv.org/abs/1309.6856. arXiv:1309.6856 [cs].

[32] M. Peschl, A. Zgonnikov, F. A. Oliehoek, and L. C. Siebert. MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning, Dec. 2021. URL http://arxiv.org/abs/2201.00012. arXiv:2201.00012 [cs].

[33] G. S. Ramachandran, I. Brugere, L. R. Varshney, and C. Xiong. GAEA: Graph Augmentation for Equitable Access via Reinforcement Learning. *arXiv:2012.03900 [cs]*, Apr. 2021. URL http://arxiv.org/abs/2012.03900. arXiv: 2012.03900.

[34] M. Reymond, E. Bargiacchi, and A. Nowé. Pareto Conditioned Networks, Apr. 2022. URL http://arxiv.org/abs/2204.05036. arXiv:2204.05036 [cs].

[35] M. Reymond, C. F. Hayes, L. Willem, R. Rădulescu, S. Abrams, D. M. Roijers, E. Howley, P. Mannion, N. Hens, A. Nowé, and P. Libin. Exploring the Pareto front of multi-objective COVID-19 mitigation policies using reinforcement learning, Apr. 2022. URL http://arxiv.org/abs/2204.05027. arXiv:2204.05027 [cs, q-bio].

[36] M. Rodriguez-Soto, M. Lopez-Sanchez, and J. A. R. Aguilar. Multi-Objective Reinforcement Learning for Designing Ethical Environments. volume 1, pages 545–551, Aug. 2021. doi: 10.24963/ijcai.2021/76. URL https://www.ijcai.org/proceedings/2021/76. ISSN: 1045-0823.

[37] M. Rodriguez-Soto, M. Serramia, M. Lopez-Sanchez, and J. A. Rodriguez-Aguilar. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1):9, Jan. 2022. ISSN 1572-8439. doi: 10.1007/s10676-022-09635-0. URL https://doi.org/10.1007/s10676-022-09635-0.

[38] D. M. Roijers, S. Whiteson, and F. A. Oliehoek. Computing Convex Coverage Sets for Faster Multi-objective Coordination. *Journal of Artificial Intelligence Research*, 52:399–443, Mar. 2015. ISSN 1076-9757. doi: 10.1613/jair.4550. URL https://www.jair.org/index.php/jair/article/view/10933.

[39] W. Röpke, M. Reymond, P. Mannion, D. M. Roijers, A. Nowé, and R. Rădulescu. Divide and conquer: Provably unveiling the pareto front with multi-objective reinforcement learning. *arXiv preprint arXiv:2402.07182*, 2024.

[40] M. Ruiz-Montiel, L. Mandow, and J.-L. Pérez-de-la Cruz. A temporal difference method for multi-objective reinforcement learning. *Neurocomputing*, 263:15–25, Nov. 2017. ISSN 0925-2312. doi: 10.1016/j.neucom.2016.10.100. URL https://www.sciencedirect.com/science/article/pii/S0925231217310998.

[41] W. Röpke. Reinforcement Learning in Multi-Objective Multi-Agent Systems. 2023.

[42] R. Rădulescu, P. Mannion, D. M. Roijers, and A. Nowé. Equilibria in Multi-Objective Games: a Utility-Based Perspective. 2019.

[43] A. Sen. Poverty: An Ordinal Approach to Measurement. *Econometrica*, 44(2):219–231, 1976. ISSN 0012-9682. doi: 10.2307/1912718. URL https://www.jstor.org/stable/1912718. Publisher: [Wiley, Econometric Society].

[44] A. F. Shorrocks. Ranking income distributions. *Economica*, 50(197): 3–17, 1983. ISSN 00130427, 14680335. URL http://www.jstor.org/stable/2554117.

[45] U. Siddique, P. Weng, and M. Zimmer. Learning fair policies in multi-objective (Deep) reinforcement learning with average and discounted rewards. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8905–8915. PMLR, July 2020.

[46] R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

[47] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, 84(1):51–80, July 2011. ISSN 1573-0565. doi: 10.1007/s10994-010-5232-5. URL https://doi.org/10.1007/s10994-010-5232-5.

[48] P. Vamplew, B. J. Smith, J. Källström, G. Ramos, R. Rădulescu, D. M. Roijers, C. F. Hayes, F. Heintz, P. Mannion, P. J. K. Libin, R. Dazeley, and C. Foale. Scalar reward is not enough: a response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2):41, July 2022. ISSN 1573-7454. doi: 10.1007/s10458-022-09575-5. URL https://doi.org/10.1007/s10458-022-09575-5.

[49] H.-n. Wang, N. Liu, Y.-y. Zhang, D.-w. Feng, F. Huang, D.-s. Li, and Y.-m. Zhang. Deep reinforcement learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 21(12):1726–1744, Dec. 2020. ISSN 2095-9230. doi: 10.1631/FITEE.1900533. URL https://doi.org/10.1631/FITEE.1900533.

[50] W. Wang and M. Sebag. Multi-objective Monte-Carlo Tree Search. In *Proceedings of the Asian Conference on Machine Learning*, pages 507–522. PMLR, Nov. 2012. URL https://proceedings.mlr.press/v25/wang12b.html. ISSN: 1938-7228.

[51] W. Wang and M. Sebag. Hypervolume indicator and dominance reward based multi-objective Monte-Carlo Tree Search. *Machine Learning*, 92(2):403–429, Sept. 2013. ISSN 1573-0565. doi: 10.1007/s10994-013-5369-0. URL https://doi.org/10.1007/s10994-013-5369-0.

[52] Y. Wei, M. Mao, X. Zhao, J. Zou, and P. An. City Metro Network Expansion with Reinforcement Learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2646–2656, Virtual Event CA USA, Aug. 2020. ACM. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403315. URL https://dl.acm.org/doi/10.1145/3394486.3403315.

[53] E. Y. Yu, Z. Qin, M. K. Lee, and S. Gao. Policy Optimization with Advantage Regularization for Long-Term Fairness in Decision Systems, Oct. 2022. URL http://arxiv.org/abs/2210.12546. arXiv:2210.12546 [cs].

[54] S. Zheng, A. Trott, S. Srinivasa, D. C. Parkes, and R. Socher. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607, May 2022. doi: 10.1126/sciadv.abk2607. URL https://www.science.org/doi/10.1126/sciadv.abk2607. Publisher: American Association for the Advancement of Science.

[55] M. Zimmer, C. Glanois, U. Siddique, and P. Weng. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12967–12978. PMLR, July 2021. URL https://proceedings.mlr.press/v139/zimmer21a.html. ISSN: 2640-3498.