A Multi-objective Framework For Fair Reinforcement Learning

Alexandra Cimpean[a], Catholijn Jonker[b], Pieter Libin[a] and Ann Nowé[a]

[a]Vrije Universiteit Brussel
[b]Technische Universiteit Delft

Abstract. Automated decision support systems, based on reinforcement learning, are increasingly useful to complex problem settings in order to optimize a primary objective. When these systems affect individuals or groups, it is essential to reflect on fairness. As absolute fairness is in practice not achievable, we propose a framework which allows to incorporate and balance distinct fairness notions along with the primary objective. To this end, we formulate sequential fairness notions in terms of groups and individuals. First, we present a Markov decision process that is explicitly aware of individuals and groups. Next, we formalize fairness notions in terms of this extended Markov decision process which allows us to evaluate the primary objective along with the fairness notions the user cares about, taking a multi-objective reinforcement learning approach. To investigate our framework, we consider two scenarios that require distinct aspects of the performance-fairness trade-off: job hiring and fraud detection. On the one hand, fairness in job hiring requires composing a strong team, while providing equal treatment of applicants as individuals as well as groups. On the other hand, fraud detection necessitates the detection of fraudulent transactions, while distributing the burden of checking customers fairly. We also highlight key research challenges regarding fairness notions as these need to include parts of the history in order to be calculated, while being impacted by the exploration strategy.

1 Introduction

Fair and balanced automated decision support is essential, to avoid discrimination or favouritism towards individuals and groups. This is crucial in a wide array of applications, such as finance [36], job hiring [49, 50], epidemic mitigation [35, 22, 13] and fraud detection [42]. Fair decision support systems allow stakeholders to make informed decisions, taking into account an appropriate performance-fairness trade-off. This is important, as advice that is proposed by a decision support system potentially impacts individuals and groups. Therefore, it is vital to study this matter to enable a wider acceptance of algorithms that support decision makers. As fairness requirements depend on the problem context and the decision maker's preferences, a framework should be capable of dealing with multiple fairness notions, that encompass the ethical considerations of the problem domain. Consequently, it is important to develop a framework that considers fairness based on sensitive features (e.g., race and gender) and their combinations.

Previous work mainly focused on supervised learning techniques that operate on a given dataset, such as machine learning [38, 20, 39, 24, 21] and data mining [9, 30, 23]. However, automated decision problems are typically sequential. Furthermore, such settings typically evolve over time and as such a reinforcement learning (RL) approach is warranted [17]. This means that we must deal with the impact of short-term decisions on long term performance. RL enables an agent to learn a policy by interacting with an environment [55]. At each time t, the agent observes the state $\mathbf{s}_t$ of the environment and decides which action a_t to take, for which it receives a reward r_t and observes the next state $\mathbf{s}_{t+1}$. The agent learns through trial and evaluation by repeatedly interacting with the environment, where it must carefully balance between exploration and exploitation to reach an optimal policy [55]. Additionally, the agent may need to deal with stochastic and non-stationary environments where it must adapt its behaviour to maintain its performance.

In a supervised classification setting, the ground truth is known and used to train the model. Based on this ground truth, a confusion matrix is computed to reflect on the correctness of the model's predictions. By definition, reinforcement learning agents do not have a priori access to a ground truth, as the agent collects data while interacting with an environment. Therefore, actions taken by the agent cannot be classified to be correct or false, which impedes the use of fairness notions that rely on a confusion matrix. As most fairness notions rely on the ground truth, they are only applicable when feedback regarding this ground truth can be collected from the environment [38].

We emphasise that this ground truth is different from the reward signal in a reinforcement learning setting. While the reward signal may indicate how suitable an action is given a state, it does not conclusively specify whether the action was correct or false. When feedback on the ground truth is available, it may concern a sparse or delayed signal. To illustrate this, consider the example of job hiring, where we receive delayed feedback as the candidate can only be evaluated after working for some time. Moreover, candidates can only be evaluated if they are hired and not when they are declined.

Recent work on fairness in RL has focused on single fairness notions in application-specific solutions [29, 28, 58, 51, 11, 48] and typically relies on reward shaping [37, 11]. However, such approaches do not suffice for real-world decision support problems, as the desired performance-fairness trade-off cannot be described upfront by stakeholders. Furthermore, real-world problems typically require multiple, possibly conflicting, fairness notions [38]. To this end, a multi-objective approach is essential to manage the main objective and to consider multiple fairness notions simultaneously [26]. We propose a formal fairness framework that is capable of dealing with multiple fairness notions. We experimentally evaluate this framework in two distinct settings: job hiring and credit card fraud detection.

2 Fairness framework

We define the fairness framework and highlight its requirements and suitability regarding distinct problem settings. To introduce fairness notions in an RL context, we illustrate them based on two real-world settings. The first setting concerns job hiring, where the aim is to hire highly qualified candidates while limiting bias towards sensitive features. As such, it is crucial that the agents recommend qualified applicants, while rejecting unsuitable ones. The second setting involves fraud detection, where fraudulent transactions must be efficiently flagged, taking into account that verification requires human effort. It is important that the agent targets real fraudulent transactions to not displease genuine customers. Additionally, fraudulent transactions constitute anomalies, rendering them challenging to detect.

We highlight that RL can be used both directly or indirectly in the context of real-world problems. On the one hand, in the fraud detection setting, a detailed simulator is used to train an agent, after which the learned policies can be studied by domain experts. On the other hand, the agent may learn directly in the real world to flag suspicious transactions. For the purpose of validating our results against a variety of scenarios, we make use of simulated data based on real data distributions.

2.1 fMDP *and the fairness history*

A sequential decision process can be formally described as a Markov Decision Process (MDP) [55], consisting of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a set of rewards $\mathcal{R}$ and a transition function $p : \mathcal{S}\times\mathcal{R}\times\mathcal{S}\times\mathcal{A} \rightarrow [0, 1]$ describing the probability of a next state $\mathbf{s}_{t+1}$ and reward r_t given the current state $\mathbf{s}_t$ and action a_t. We extend this standard MDP to an fMDP to encode a feedback signal f_t, that concerns an indication whether the chosen action a_t was correct at time t. Note that this feedback is optional and can be partial, sparse or delayed. As the presence of the ground truth is required for some fairness notions, it must be either obtained through feedback or approximated based on previous interactions.

Existing fairness notions typically concern fair treatment between individuals or groups. We introduce the following notation regarding individuals and groups. $\mathcal{I}_t$ refers to the set of individuals involved in the decision process at time t and we use i_t to refer to an individual of that set. In the job hiring setting, $\mathcal{I}_t$ refers to the set of candidates who applied for the job at time t and for which a decision (i.e., hire or reject the applicant) should be made. In the fraud detection setting, $\mathcal{I}_t$ refers to all customers at time t when deciding whose transactions to verify. We refer to the set of all individuals involved in the decision process from the start $t = 0$ up to time T as $\mathcal{I}^T$.

We define $\mathcal{G}_{g,t}$ as the individuals of $\mathcal{I}_t$ that make up group g. We refer to all individuals involved in the decision process until time T, that belong to group g, as $\mathcal{G}_g^T$. For ease of notation, we assume that groups are predefined and can be empty. In the job hiring setting, $\mathcal{G}_g^T$ refers to the group of men or women, who applied for a job until time T. For the fraud detection setting, $\mathcal{G}_g^T$ refers to a continent for which the RL agent must decide whether or not to flag transactions.

Given the fMDP, we assume that a state $\mathbf{s}_t$ provided to the RL agent encodes the individuals $\mathcal{I}_t$ and groups $\mathcal{G}_t$ involved in the decision at time t. Furthermore, the agent's action a_t encodes the decision impacting the involved individuals and groups, and the feedback f_t specifies the correctness of that decision. We use the following notation to connect $\mathcal{I}_t$ and $\mathcal{G}_t$ to $\mathbf{s}_t$, a_t and f_t:

$$\mathcal{I}_t[\mathbf{s}_t],\ \mathcal{I}_t[a_t],\ \mathcal{I}_t[f_t] \tag{1}$$

$$\mathcal{G}_t[\mathbf{s}_t],\ \mathcal{G}_t[a_t],\ \mathcal{G}_t[f_t] \tag{2}$$

To define fairness over time, a history of encountered states and chosen actions needs to be maintained, concerning the impacted individuals and groups. Given an fMDP, we define a history $\mathcal{H}^T$ until time T of past interaction tuples and their feedback regarding the ground truth:

$$\mathcal{H}^T = \{\mathbf{s}_t, a_t, r_t, f_t\}_{t=0}^T \tag{3}$$

We define the encountered states and selected actions from history $\mathcal{H}^T$ until time T respectively as $\mathcal{H}_S^T$ and $\mathcal{H}_A^T$. We refer to feedback regarding the correctness of the action as $\mathcal{H}_f^T$. Following from the definitions in Equations 1 and 2, $\mathcal{H}_S^T$, $\mathcal{H}_A^T$ and $\mathcal{H}_f^T$ are defined in terms of groups $\mathcal{G}^T$ and individuals $\mathcal{I}^T$. In the job hiring setting, the history consists of the encountered job applicants and their corresponding decision, indicating whether or not they were hired. In the fraud detection setting, the history consists of all observed transactions, along with their checking status. In both settings, the history is used to define fairness over time.

2.2 *Fairness notions*

We formally define a fairness notion $\mathscr{F}$ as a power set $\mathscr{P}$ over $\mathcal{G}^T$ groups (Equation 4) and $\mathcal{I}^T$ individuals (Equation 5), given the history of encountered states $\mathcal{H}_S^T$, chosen actions $\mathcal{H}_A^T$ and feedback $\mathcal{H}_f^T$ until time T:

$$\mathscr{F} : \mathscr{P}(\mathcal{G}^T) \times \mathscr{P}(\mathcal{H}_S^T) \times \mathscr{P}(\mathcal{H}_A^T) \times \mathscr{P}(\mathcal{H}_f^T) \hookrightarrow \mathbb{R}^- \tag{4}$$

$$\mathscr{F} : \mathscr{P}(\mathcal{I}^T) \times \mathscr{P}(\mathcal{H}_S^T) \times \mathscr{P}(\mathcal{H}_A^T) \times \mathscr{P}(\mathcal{H}_f^T) \hookrightarrow \mathbb{R}^- \tag{5}$$

The fairness notion $\mathscr{F}$ is defined as the negative absolute difference in treatment between groups or individuals. The closer $\mathscr{F}$ is to zero, the smaller the difference in treatment is between the groups or individuals. When $\mathscr{F} = 0$, the agent has achieved exact fairness with respect to the given fairness notion. While $\mathscr{F}$ may be intractable due to limitations of defining exact fairness [28], we propose to approximate it with $\hat{\mathscr{F}}$. For a future fairness objective, $\mathscr{F}$, and by extension its approximation $\hat{\mathscr{F}}$ provide a foundation for a reward signal that can be used with a multi-objective RL approach.

The availability of a ground truth and as a consequence the confusion matrix impacts which fairness notions can be calculated for a given scenario. The confusion matrix is defined as a two-dimensional table comparing predictions of a model to the actual values. In the case of binary actions (e.g., hire or reject an applicant) it specifies the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Consider the group fairness notion *statistical parity* [20], where the probability of receiving the preferable treatment of the agent ($\mathcal{H}_A^T = 1$) should be the same across groups g and h:

$$\begin{aligned}\mathscr{F} = -|&\mathrm{P}(\mathcal{G}_g^T[\mathcal{H}_A^T] = 1|\mathcal{G}_g^T[\mathcal{H}_S^T]) \\ &- \mathrm{P}(\mathcal{G}_h^T[\mathcal{H}_A^T] = 1|\mathcal{G}_h^T[\mathcal{H}_S^T])|\end{aligned} \tag{6}$$

Statistical parity requires that $(TP+FP)/(TP+FP+FN+TN)$ is equal for both groups g and h. Because this fairness notion focuses on equal acceptance rate across groups, it can be expressed without knowledge of the ground truth. Other fairness notions require that the ground truth is (partially) known, such as *equal opportunity*:

$$\begin{aligned}\mathscr{F} = -|&\mathrm{P}(\mathcal{G}_g^T[\mathcal{H}_A^T] = 1|\mathcal{G}_g^T[\mathcal{H}_f^T] = 1, \mathcal{G}_g^T[\mathcal{H}_S^T]) \\ &- \mathrm{P}(\mathcal{G}_h^T[\mathcal{H}_A^T] = 1|\mathcal{G}_h^T[\mathcal{H}_f^T] = 1, \mathcal{G}_h^T[\mathcal{H}_S^T])|,\end{aligned} \tag{7}$$

where $\mathcal{H}_f^T = 1$ is the correct action as specified by the feedback regarding the ground truth. Equal opportunity requires that the recall

$TP/(TP+FN)$ is equal across groups and is consequently independent of FP. However, in order to calculate it, we require a (partial) ground truth that informs us about TP and FN. In the job hiring setting, this requires knowing how qualified a job candidate is to calculate the confusion matrix. In the fraud detection setting, the partial ground truth is available, where transactions flagged as fraudulent are manually verified, which provides the number of TP and FP. In contrast, there is no information on unflagged transactions unless random checks are performed, or when individuals complain about fraud cases in their experience. Ensuring people are treated fairly, with regard to all groups they are a part of, is achieved by ensuring all their groups are treated fairly with regard to each other. If the interest is that each individual receives fair treatment, then individual fairness notions should be used.

Individual fairness notions aim to treat similar individuals similarly [20]. Given two individuals i_t and j_t, we assume a distance $d(i_t, j_t)$ between the individuals. Given the probability distributions M_i and M_j over the actions for i_t and j_t respectively, and a distance metric $D(M_i||M_j)$ between these probability distributions, individual fairness requires that:

$$\forall i_t, j_t \in \mathcal{I}_t : \quad D(M_i||M_j) \leq d(i_t, j_t) \tag{8}$$

As group fairness notions aim to similarly treat groups that differ by a set of sensitive features, they cannot detect unfairness at an individual level, as all attributes except the sensitive ones are ignored [20]. Similarly, individual fairness notions lack the ability to ensure fairness between groups. Ideally, an RL agent conforms to a collection of both group and individual fairness notions to manage this trade-off, which can be managed using a multi-objective learning approach [26].

2.3 *Fairness in sequential decision making*

Defining fairness in a sequential setting requires knowledge of how fairness notions can be defined given the agent-environment interactions. Consider the fraud detection setting, where an agent must decide how to efficiently flag transactions each day for a credit card company [60]. Throughout the day, each individual client may decide to make transactions. The agent aims to flag suspicious transactions, in a way that everyone in the population is subject to a similar amount of re-authentication requests.

Suppose in our fraud detection setting, that each hour the agent encounters the continents from where customers attempt transactions. Each hour, the agent chooses how to flag transactions for the respective continents. Then at each time t, given an observed state $\mathbf{s}_t$ and chosen action a_t, given $\mathcal{G}_t$ groups, a group fairness notion can be defined if $\mathbf{s}_t$ contains all respective groups $\mathcal{G}_t[\mathbf{s}_t]$ and the chosen action a_t represents the action taken towards each group $\mathcal{G}_t[a_t]$. Figure 1a visualises the possible scenarios with regards to the available action, which can be an action over all groups $\mathcal{G}_t$, or a specific action for each group g. Note that if individuals are defined within the state representation, then Equation 2 can be defined by grouping individuals in $\mathcal{I}_t$ under their respective groups $\mathcal{G}_t$.

Next in the fraud detection setting, consider that the agent only encounters certain continents on an hourly basis, which could be the case due to different time zones. Then a sufficiently long time horizon must be considered to encounter all age groups. Concretely, if the state $\mathbf{s}_t$ contains only information on a subset $\mathcal{B}_t \subset \mathcal{G}_t$ of the respective groups, a fairness notion can only be defined when considering multiple timesteps of encountered groups $\mathcal{B}^T$ to contain sufficient information about all impacted $\mathcal{G}_t$ groups for time t:

$$\mathcal{G}_t[\mathbf{s}_t] = \bigcup_{g \in \mathcal{B}^T} \mathcal{G}_g^T[\mathcal{H}_S^T] \tag{9}$$

Similarly, we require multiple timesteps if the action a_t does not define the action for all groups:

$$\mathcal{G}_t[a_t] = \bigcup_{g \in \mathcal{B}^T} \mathcal{G}_g^T[\mathcal{H}_A^T] \tag{10}$$

If individuals are defined within the state representation of the environment, Equations 9 and 10 can be extended to consider cases where a subset of individuals is encountered. Figure 1b visualises the scenario where only a subset of the groups is available at each time t, requiring a history of timesteps in order to express group fairness notions.

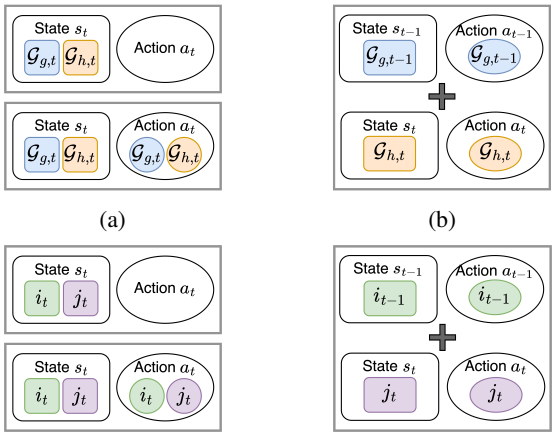


Figure 1: (a) (b) Scenarios where group fairness can be calculated. (a) All groups $\mathcal{G}_t$ are encountered at each time t. Top: action a_t is an action over all groups $\mathcal{G}_t$. Bottom: action a_t encodes a specific action for each group g. (b) All groups $\mathcal{G}_t$ are encountered over a time horizon until time T. The $+$ symbol indicates a union over states and actions. (c) (d) Scenarios where individual fairness can be calculated. (c) All individuals $\mathcal{I}_t$ are encountered at each time t. Top: action a_t is an action over all individuals $\mathcal{I}_t$. Bottom: action a_t encodes a specific action for each individual i_t. (d) All individuals $\mathcal{I}_t$ are encountered over a time horizon until time T. The $+$ symbol indicates a union over states and actions.

Following up on the same fraud detection setting, when the agent encounters all customers each hour, then individual fairness notions can be calculated for the transactions. To define an individual fairness notion for $\mathcal{I}_t$ individuals at time t, given an observed state $\mathbf{s}_t$ and a chosen action a_t, we require that $\mathcal{I}_t[\mathbf{s}_t]$ and $\mathcal{I}_t[a_t]$ are defined. Figure 1c visualises the scenarios where individual fairness can be calculated at each time t. Note that the action can be fine-grained for each individual or coarse-grained over their respective countries or continents.

When only a portion of the individuals is encountered at each time step, then we can only calculate individual fairness notions when we maintain a history of interactions. An example for fraud detection is checking a subset of all customers for a given continent at different times during the day, to monitor suspicious transactions based on their local time. In this case, a fair agent should balance over time which

continents are checked more often to not cause certain customers to re-authenticate more than others. If state $\mathbf{s}_t$ does not contain all $\mathcal{I}_t$ individuals but rather a subset $\mathcal{C}_t \subset \mathcal{I}_t$, an individual fairness notion can be defined over multiple time steps so that all affected individuals $\mathcal{C}^T$ are encountered:

$$\mathcal{I}_t[\mathbf{s}_t] = \bigcup_{i \in \mathcal{C}^T} i^T[\mathcal{H}_S^T] \quad (11)$$

$$\mathcal{I}_t[a_t] = \bigcup_{i \in \mathcal{C}^T} i^T[\mathcal{H}_A^T] \quad (12)$$

Figure 1d visualises the scenario where individual fairness can be expressed over multiple time steps. Note how both group and individual fairness notions can be expressed if the encountered states contain all necessary information about the respective groups and individuals. Regardless of whether the action was specifically assigned to them, their group, or the entire population, we can compare the action which affects them to calculate fairness notions. In this paper, we consider the scenarios from Figures 1b and 1d, where all groups and individuals are encountered over a history, respectively.

Depending on the setting, it could be more important to check fairness notions against the impact of the agent's action rather than against the action itself. We discussed actions with regard to the applicability of fairness notions, however, both the immediate and estimated effect follow similar rules as information about them must also be available in the agent-environment interactions. An example in the context of fraud detection, where the impact of the action is considered important, concerns the need to detect fraudulent customers to prevent that the company makes losses due to fraud. If the impact of the action is more important, the agent should avoid burdening genuine customers with frequent checks to not lower customer satisfaction.

We consider each fairness notion by computing its approximation $\hat{\mathscr{F}}$, through a history with a sliding window of the most recent interactions. Note that this approximation is necessary due to the intractability of fairness notions to achieve exact fairness over the full history. On the one hand, we require enough interactions to guarantee exact fairness [28]. On the other hand, considering the full history makes computing the fairness notions intractable. Individual fairness notions in particular become intractable due to the pairwise comparisons needed between each new individual and all those previously encountered. In the context of data mining, approaches focus on over-representing minorities or rare events [41] in the training data. Similarly, recommender systems suffer from uneven data distributions which impacts fairness and as such requires re-distributing the data to appropriately compare groups and individuals [57]. We consider such approaches as future work to learn better approximations for fairness notions over a full history.

2.4 Learning and exploration

In the previous sections, we assume that the states in the history encompass all groups $\mathcal{G}^T$ and individuals $\mathcal{I}^T$ necessary to compute the relevant fairness notions. However, to meet this assumption, the relevant states need to be encountered, which is highly dependent on how the agent interacts with the environment. To establish this, we need an appropriate exploration strategy that ensures that sufficient information is collected about all groups $\mathcal{G}^T$ and individuals $\mathcal{I}^T$. On the one hand, to guarantee optimality, this exploration strategy will need to collect information on groups and individuals as broadly as possible. On the other hand, to keep the process computationally tractable, the exploration strategy should be effective and targeted. Note that, as we aspire to settings that aim to support decision makers, we can learn and evaluate policies in simulated environments, prior to deploying them in the real world. This facilitates a model-based reinforcement learning loop that could mitigate the hurdle of computationally intensive exploration strategies.

3 Related work

As reinforcement learning approaches are well suited to deal with sequential processes, new research has focused on multi-armed bandit approaches [29, 44]. To enforce fairness in job hiring, multi-armed bandits [49, 50] as well as generalisations towards MDP approaches [28] have been explored. However, current solutions do not employ the multi-objective approach that is necessary for learning an appropriate performance-fairness trade-off. Approaches for fraud detection often rely on offline trained algorithms [16, 15, 34], which are retrained as labelled data becomes available with a delay. Soemers et al. [42] propose a contextual bandit implementation that is able to adapt to changes in fraudulent behaviour. In the field of epidemic control, mitigation strategies have been explored in RL [35, 22, 13]. Reymond et al. [47] present a multi-objective approach for minimising infections and hospitalisations, taking into account the social burden of lost contacts. While this work does not focus on fairness explicitly, it highlights how other real-world problems have a critical fairness component to consider.

In the context of fairness, group fairness notions often rely on predefined groups. As such, these group notions do not guarantee fairness amongst any further subgroup divisions. Therefore, it is possible for an algorithm to learn a fair policy for the given groups, while being unfair for subgroups. Kearns et al. [31] propose a technique to deal with this phenomenon, which is known as gerrymandering. They highlight the need for a more extensive fairness evaluation when it comes to group fairness by enforcing fairness for the subgroups as well. This work aligns with our argument for a multi-objective approach to enforce multiple fairness constraints with regards to existing fairness notions.

4 Scenarios

In this section, we introduce the two scenarios used for our experiments, along with their distinct fairness implications.

4.1 Job hiring

Job hiring is a reoccurring process as it is repeated multiple times throughout the company's lifetime. This allows companies to use previous data when training algorithms. However, the training data may be subject to historical bias, which is then further exacerbated by the algorithm [38]. Additionally, the job hiring process is sequential, typically consisting of multiple decision stages, i.e., resume screenings and possibly multiple rounds of interviews [7], which warrants a sequential approach. Moreover, unfairness at one stage may be propagated to consecutive stages. In job hiring, gender-based discrimination ranges from stereotypes and employer beliefs [56, 4] to occupational-specific characteristics [27, 33, 2, 14, 1]. Ethnic discrimination has been studied from an immigration perspective [61] and is based on implicit interethnic attitudes [6]. Moreover, combinations of sensitive features are known to cause discrimination [45, 3, 19].

Job hiring fMDP We define the job hiring setting as an fMDP, where an agent must learn to build a well-performing team of employees, when presented applicants sampled from the Belgian population [52]. Given an applicant and the current team composition, the agent must decide on the appropriate action a_t, i.e., to hire or reject the applicant, based on their estimated qualifications. To calculate the qualification of each applicant, we define an objective but noisy goodness score $G \in [-1, 1]$, that quantifies how much the applicant is estimated to improve the company based on their skills. We define this goodness score as the ground truth for our experiments based on which the fMDP classifies applicants. Using a threshold $\epsilon = 0.5$, the ground truth action $\hat{a}_t$ says to hire the applicant if $G_t >= \epsilon$, otherwise reject. We provide additional details in Appendix A on the job hiring fMDP and the applicant generation.

Fairness notions in job hiring In this work, we consider fairness concerns in job hiring based on discrimination grounded in two sensitive features: gender and nationality. As the agent observes an applicant in the state $\mathbf{s}_t$ at each timestep t, both individual and group fairness notions are applicable (Section 2.2). We consider the context of unfairness based on gender, where an applicant $i_t \in \mathcal{I}^T$ can belong to the group of men $\mathcal{G}^T_{gmen}$ or women $\mathcal{G}^T_{gwomen}$. For job hiring, we consider the group fairness notions statistical parity (Equation 6) and equal opportunity (Equation 7) as objectives in addition to the main reward. We define individual fairness (Equation 8) between applicants using the Bray-Curtis distance $d \in [0, 1]$ [8], which corresponds to the Manhattan distance divided by the sum of the applicants' features.

4.2 Fraud detection

Fraudulent credit card transactions result in significant losses when undetected [18]. While manual investigations can accurately detect fraud, it is unfeasible for the large number of transactions without suffering delays. Moreover, fraudsters are known to change their behaviour over time to avoid detection [16], requiring an online approach to continuously adapt to new fraud behaviours. As customers perform multiple transactions over a certain time period, the credit card company must deal with customer satisfaction and patience when requiring authentication steps to process a transaction [60]. As transactions typically include personal and location data, algorithms may learn to discriminate based on sensitive features. For example, countries with higher base rates (i.e., proportions of fraudulent transactions) than others may have customers checked more often based only on their location [38]. To this end, fraud detection requires fairness notions which take into account this difference in base rate to accurately flag transactions.

Fraud detection fMDP The fraud detection setting concerns online credit card transactions where multi-modal authentication is used to identify and reject fraudulent transactions. To simulate customer behaviour, we use the MultiMAuS simulator [60], which is based on a database of real-world credit card transactions. We extend this simulator to a fMDP, by providing the current company's fraudulent transactions percentage and customer satisfaction along with the transaction in the state at each time step. The feedback signal f is defined based on the gain or loss in reward, indicating if revenue was lost due to fraud. Concretely, the agent receives a positive reward of $+1$ for every succesful genuine transaction and -1 for uncaught fraudulent transactions and cancelled transactions. We provide additional details on the MultiMAuS simulator and the fMDP in Appendix B.

Fairness notions in fraud detection We investigate unfairness in fraud detection based on the continent of the customers. As the agent observes a new transaction in state $\mathbf{s}_t$ at timestep t, both individual and group fairness notions are applicable. For simplicity, we consider two continents, C_a and C_b, with the most transactions. We define group fairness notions as follows: Given transactions $i_t \in \mathcal{I}^T$, where transaction i_t can belong to continent C_a or C_b, all group fairness notions require that the difference in treatment between the groups $\mathcal{G}^T_{C_a}$ and $\mathcal{G}^T_{C_b}$ is minimised. For the group fairness notion overall accuracy equality [5], the accuracy of the agent should be the same across the continent groups C_a and C_b.

$$\begin{aligned}\mathscr{F} = -|&\mathrm{P}(\mathcal{G}^T_{C_a}[\mathcal{H}^T_A] = \mathcal{G}^T_{C_a}[\mathcal{H}^T_f]|\mathcal{G}^T_{C_a}[\mathcal{H}^T_S]) \\ &- \mathrm{P}(\mathcal{G}^T_{C_b}[\mathcal{H}^T_A] = \mathcal{G}^T_{C_b}[\mathcal{H}^T_f]|\mathcal{G}^T_{C_b}[\mathcal{H}^T_S])|\end{aligned} \tag{13}$$

Predictive parity [12] requires that the probability of being fraudulent, given that the agent requested a re-authentication, is the same across groups C_a and C_b.

$$\begin{aligned}\mathscr{F} = -|&\mathrm{P}(\mathcal{G}^T_{C_a}[\mathcal{H}^T_f] = 1|\mathcal{G}^T_{C_a}[\mathcal{H}^T_A] = 1, \mathcal{G}^T_{C_a}[\mathcal{H}^T_S]) \\ &- \mathrm{P}(\mathcal{G}^T_{C_b}[\mathcal{H}^T_f] = 1|\mathcal{G}^T_{C_b}[\mathcal{H}^T_A] = 1, \mathcal{G}^T_{C_b}[\mathcal{H}^T_S])|\end{aligned} \tag{14}$$

In fraud detection, we define individual fairness between transactions using the complement of the consistency score [59]:

$$\mathscr{F} = -\frac{1}{||\mathcal{I}^T||}\sum_{i \in \mathcal{I}^T}\frac{1}{k}|a^i - \sum_{j \in kNN(i)} a^j| \tag{15}$$

given action a^i for an individual i, where k is the number of nearest neighbours to consider, given a k-nearest neighbour algorithm kNN [40]. To compare this notion to the individual fairness notion from job hiring, we assume the same Bray-Curtis distance for kNN.

5 Experiments

As both the job hiring and fraud detection scenario deal with a reward and multiple fairness objectives, the number of policies with suitable trade-offs can scale exponentially. To find all policies would therefore be a computationally expensive task. To this end, we use Pareto Conditioned Networks (PCN) [46]. PCN trains a single neural network to approximate all non-dominated policies, by applying supervised learning techniques to improve the policy. We provide additional details on PCN in Appendix C.

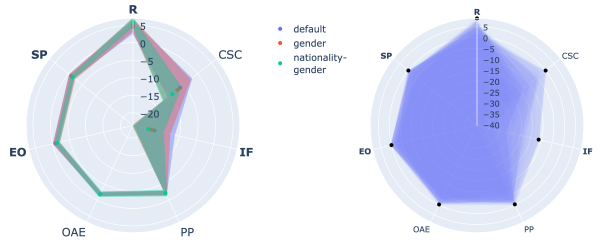
For both experiments, we report the learned non-dominated coverage sets for all fairness notions and the reward [26]. Therefore, the reward vector consists of: the main reward (R), statistical parity (SP), equal opportunity (EO), overall accuracy equality (OAE), predictive parity (PP), individual fairness (IF), consistency score complement (CSC). We assume that the (objective) ground truth is known for both environments, to investigate all 6 fairness notions. Specifically, the fairness notions EO, OAE and PP require access to the ground truth. We explore the impact of multiple sliding windows on the fairness notions in the experiments below.

5.1 Job hiring

For the job hiring scenario, we train a PCN agent to hire and maintain a well-performing team of employees. We compare the task for building teams of two different sizes: Building a team of 20 or 100 employees, where each episode lasts for a maximum of 200 or 1000 timesteps, respectively. We report the results for all objectives, but

ask the agent to optimise on four: {R, SP, EO, IF}. We consider the Belgian population, from the STATBEL statistics [52], and apply 2 permutations two skew the population distribution. For one permutation, we skew the originally equal proportions for gender such that 70% of applicants are men and 30% women. The other permutation focuses on the combination of nationality and gender, such that foreign women become a minority.[1]

Team-20 Figure 2 displays the results for the different population distributions, with regards to the learned non-dominated coverage sets for building a team of 20 employees. We implement the fairness history as a sliding window of 100 timesteps. Note how the best learned policies are close to 0 for all group fairness notions, indicating the agent has learned policies which can satisfy more fairness notions than initially requested. As most group fairness notions require the confusion matrix to compute, there are overlaps with regards to the involved true and predicted actions. In contrast, due to the impact of how the individual fairness notions are defined, it is possible for the agent to find larger differences in non-dominated values among them. Concretely, when comparing similar individuals, IF considers the probability distributions over the actions, while CSC considers the action only. Overall, the agent ensures group fairness in all populations. However, we note a slightly lower individual fairness for the third population with a minority of foreign women. While the agent is able to be fair with regards to gender, it does so without considering the sensitive attribute nationality. This further highlights the need for individual fairness notions to detect feature intersectionality [38].



(a) (b)

Figure 2: Team-20. Non-dominated coverage sets after 100 000 timesteps for a desired team of 20 employees, with requested objectives in bold. (a) Mean and standard deviation of the non-dominated value for each objective across 10 seeds. (b) Trade-off area reached by the non-dominated coverage set for a single seed of the default population. Note how the notions IF and CSC are harder to maximise, indicating the agents are better at dealing with group fairness notions relying on statistical measures than individual fairness notions relying on a chosen similarity metric between individuals.

Figure 3 shows the results for an agent optimising for one objective at a time. It is noteworthy that in all three populations, the agent is more unfair when it optimises for the reward. In contrast, an agent focusing on any fairness notion provides a higher amount of fairness, at the cost of the reward. We argue that applying a multi-objective approach is crucial in these scenarios, to learn policies which reach meaningful objective trade-offs.

[1] We skew the true proportions to $\{(Belgian, man) : 30\% \rightarrow 40\%, (Belgian, woman) : 31\% \rightarrow 40\%, (foreign, man) : 20\% \rightarrow 15\%, (foreign, woman) : 19\% \rightarrow 5\%\}$

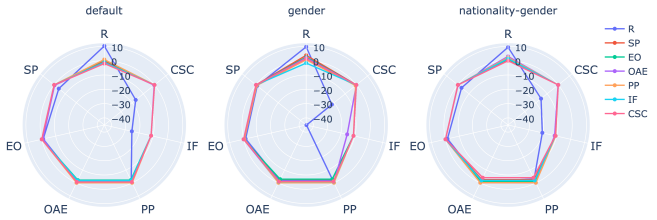
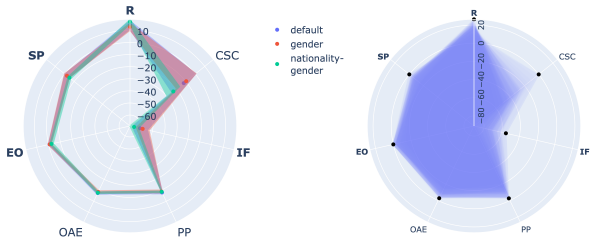


Figure 3: Team-20. Best learned policy when optimising a single objective per population for a single seed, for a history of 100 timesteps.

Team-100 Figure 4 displays the reached non-dominated coverage sets, for building a team of 100 employees. We observe similar results, compared to the teams of 20 employees, across the fairness notions for the different populations. The agent learns policies where all group fairness notions are close to 0, while the individual fairness notions IF and CSC prove more difficult to satisfy.

Figure 5 shows the results for an agent optimising for one objective at a time when building a team of 100 employees. As for team-20, when optimising for the reward only, we observe lower fairness values, with individual fairness notions being particularly low. In contrast to team-20, there is a noticeable difference between optimising a group fairness notion or an individual fairness notion across the populations. This further highlights how only optimising for group fairness does not guarantee individual fairness and vice versa. Furthermore, note how all populations result in less similar policies across the fairness notions than the default population, indicating there is an influence of the observed state distributions on fairness objectives.



(a) (b)

Figure 4: Team-100. Non-dominated coverage sets after 500 000 timesteps for a desired team of 100 employees, with requested objectives in bold. (a) Mean and standard deviation of the non-dominated value for each objective across 10 seeds. (b) Trade-off area reached by the non-dominated coverage set for a single seed of the default population. The individual fairness notion IF is lowest, as it compares the probability distributions over individuals, which proves more difficult to keep similar across individuals. Note that CSC focuses on the chosen actions only, making it easier to satisfy.

5.2 *Fraud detection*

For the fraud detection scenario, we assume the default parameters of the MultiMAuS simulator [60], but increase the frequency of fraudulent transactions to ensure enough genuine and fraudulent transactions are encountered for both continents. This results in approximately 10% fraudulent transactions. We let the agent check transactions for a week, resulting in at most 1000 transactions per episode. The results for all objectives are presented, but we ask the agent to optimise on only four: {R, OAE, PP, CSC}. Due to the different transaction

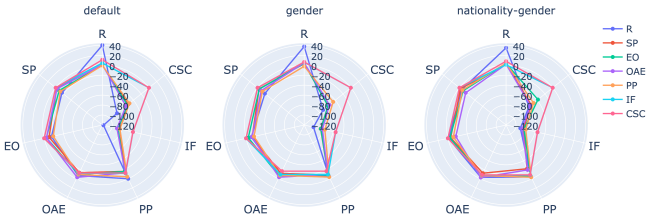


Figure 5: Team-100. Best learned policy when optimising a single objective per population for a single seed, for a history of 100 timesteps.

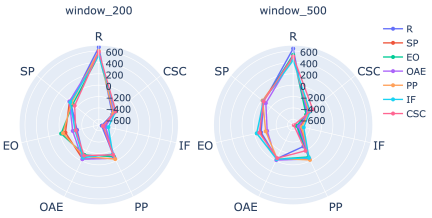


Figure 7: Best learned policy for fraud detection when optimising a single objective. Note how when optimising for each objective separately, the agent learns different policies for both window sizes.

frequencies across continents C_a and C_b, we require a larger sliding window for the history to compare fairness for both continents. In the following experiments, we explore a history with a sliding window of 200 and 500 timesteps.

Figure 6 shows the learned trade-offs for both history sizes. The policies learned by the agent across both window sizes follow similar trade-offs with regards to the reward and the fairness notions. While the policies improve the requested group fairness notions OAE and PP, there is a notable difference with the EO fairness notion. This is caused by the similarity in treatment required by these fairness notions. Concretely, OAE requires that the agent has the same accuracy across continents, while PP requires that re-authentication requests lead to the same probability of catching fraudulent transactions for these continents. In contrast, EO requires that fraudulent transactions are flagged with the same probability across continents. Note that individual fairness is low for both IF and CSC, indicating the reward is conflicting with the agent's fairness. The largest contributor to this effect is the different base rates for fraudulent transactions between individuals, indicating the agent has mostly focused on improving the requested group fairness notions, at the cost of individual fairness.

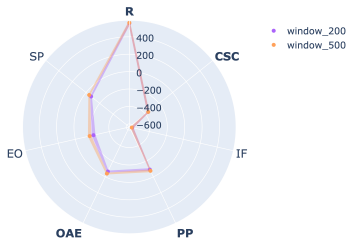


Figure 6: Fraud detection, with requested objectives in bold. Mean and standard deviation of the non-dominated value for each objective across 10 seeds after 500 000 timesteps.

Figure 7 shows the best found single-objective policies for both window sizes. We observe a difference in learned policies when using a different window size. Most notably, the group fairness notions EO and PP cannot be maximised equally by optimising for one of the other objectives. We attribute this difference to the reliance of group fairness notions on statistics over the history. By using a different history size, choosing the appropriately fair action depends on the actions chosen previously to provide similar treatment over the groups.

6 Discussion

We propose a framework for exploring the use of fairness notions in RL. In this framework, we establish a formulation of fairness notions that can be used as additional reward signals following a multi-objective learning approach. Based on this formulation, we classify distinct fairness settings grounded in real-world problems. We highlight the need of multiple fairness notions, particularly ensuring both group and individual fairness simultaneously. Due to the context dependency of fairness, we show how requested fairness notions can be conflicting with the main objective to optimise, as in the fraud detection scenario. As such, we argue the multi-objective aspect is crucial in the development of the fairness framework.

By formulating fairness notions in terms of the history defined, we establish a formal way to reason about fairness notions as reward functions. Yet, as maintaining the full history will prove computationally intractable for most real-world applications, a major challenge remains to construct approximate fairness notions. Individual fairness notions in particular require pair-wise comparisons of all individuals, in contrast to group fairness notions that rely on the statistical measures of each group. One research direction is to consider a sliding window approach, where the history is kept for a fixed or dynamic number of steps [43]. Another path is to explore the use of distinct neural sub-networks for the different fairness notions.

Generalising fairness notions to continuous actions presents an interesting venue to extend fairness to a wider array of problem settings. In the field of regression, algorithms produce a scalar value rather than a discrete action from a predefined set. Consequently, regression compares actions based on how much they differ and can detect correlations between the action and one or more sensitive features [32], which makes it an interesting approach for comparing actions in RL.

Within the overarching topic of ethics, work on explainable AI focuses on making algorithms interpretable and provides explanations for their decisions [25]. While explainability aims to provide transparency with regards to an agent's decisions and policy, fairness focuses on whether or not the agent makes decisions which conform to expected impartial treatment. We argue that fairness is an equally important aspect to focus on to work towards ethical AI. To truly build a fair decision support system, we envision the need to combine fairness notions with explainable reinforcement learning, such that fairness can be taken into account when explaining policies to the decision maker.

Acknowledgements

Alexandra Cimpean receives funding from the Fonds voor Wetenschappelijk Onderzoek (FWO) via fellowship grant 1SF7823N. Pieter Libin gratefully acknowledges support from FWO postdoctoral fellowship 1242021N, FWO grant G059423N, and the Research council of the Vrije Universiteit Brussel via grant number OZR3863BOF. Catholijn Jonker's work is supported by the National Science Foundation (NWO) under grant number 1136993. Ann Nowé acknowledges support from FWO grant G062819N. All experiments were performed on the VSC high performance computing infrastructure [10].

References

[1] Mladen Adamovic and Andreas Leibbrandt, 'A large-scale field experiment on occupational gender segregation and hiring discrimination', *Industrial Relations*, **62**(1), 34–59, (2023).

[2] Ali Ahmed, Mark Granberg, and Shantanu Khanna, 'Gender discrimination in hiring: An experimental reexamination of the Swedish case', *PLOS ONE*, **16**(1), 1–15, (2021).

[3] Stijn Baert, *Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005*, 63–77, Springer International Publishing, Cham, 2018.

[4] Kai Barron, Ruth Ditlmann, Stefan Gehrig, and Sebastian Schweighofer-Kodritsch, 'Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment', *SSRN Electronic Journal*, (9731), (2022).

[5] Richard A. Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, 'Fairness in criminal justice risk assessments: The state of the art', *Sociological Methods & Research*, **50**, 3 – 44, (2018).

[6] Lieselotte Blommaert, Frank van Tubergen, and Marcel Coenders, 'Implicit and explicit interethnic attitudes and ethnic discrimination in hiring', *Social Science Research*, **41**(1), 61–73, (2012).

[7] Miranda Bogen and Aaron Rieke, 'Help wanted: An examination of hiring algorithms, equity, and bias', Technical report, (2018).

[8] J Roger Bray, 'An ordination of the upland forest communities of southern wisconsin', *Ecological monographs*, **27**(4), 326–349, (1957).

[9] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney, 'Optimized pre-processing for discrimination prevention', in *31st International Conference on NIPS*, p. 3995–4004, (2017).

[10] Vlaams Supercomputing Center. Hydra hardware, 2023. https://www.vscentrum.be.

[11] Jingdi Chen, Yimeng Wang, and Tian Lan, 'Bringing fairness to actor-critic reinforcement learning for network utility optimization', in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, p. 1–10, Vancouver, BC, Canada, (2021). IEEE Press.

[12] Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big data*, **5**(2), 153–163, (2017).

[13] Alexandra Cimpean, Timothy Verstraeten, Lander Willem, Niel Hens, Ann Nowé, and Pieter Libin, 'Evaluating covid-19 vaccine allocation policies using bayesian m-top exploration', *arXiv preprint arXiv:2301.12822*, (2023).

[14] Clara Cortina, Jorge Rodríguez, and M. José González, 'Mind the Job: The Role of Occupational Characteristics in Explaining Gender Discrimination', *Social Indicators Research*, **156**(1), 91–110, (2021).

[15] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi, 'Credit card fraud detection and concept-drift adaptation with delayed supervised information', in *2015 international joint conference on Neural networks (IJCNN)*, pp. 1–8. IEEE, (2015).

[16] Andrea Dal Pozzolo, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi, 'Learned lessons in credit card fraud detection from a practitioner perspective', *Expert systems with applications*, **41**(10), 4915–4928, (2014).

[17] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern, 'Fairness is not static: Deeper understanding of long term fairness via simulation studies', in *Conference on Fairness, Accountability, and Transparency*, pp. 525–534, Barcelona, Spain, (2020). Association for Computing Machinery.

[18] Linda Delamaire, Hussein Abdou, and John Pointon, 'Credit card fraud and detection techniques: a review', *Banks and Bank systems*, **4**(2), 57–68, (2009).

[19] Eva Derous and Roland Pepermans, 'Gender discrimination in hiring: Intersectional effects with ethnicity and cognitive job demands.', *Archives of Scientific Psychology*, **7**, 40–49, (2019).

[20] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, 'Fairness through awareness', in *3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, p. 214–226, New York, NY, USA, (2012). Association for Computing Machinery.

[21] Cynthia Dwork, Christina Ilvento, Guy N. Rothblum, and Pragya Sur, 'Abstracting Fairness: Oracles, Metrics, and Interpretability', in *1st Symposium on Foundations of Responsible Computing*, Cambridge, MA, USA, (2020). Curran Associates, Inc.

[22] Ezekiel J. Emanuel, Govind Persad, Adam Kern, Allen Buchanan, Cécile Fabre, Daniel Halliday, Joseph Heath, Lisa Herzog, R. J. Leland, Ephrem T. Lemango, Florencia Luna, Matthew S. McCoy, Ole F. Norheim, Trygve Ottersen, G. Owen Schaefer, Kok-Chor Tan, Christopher Heath Wellman, Jonathan Wolff, and Henry S. Richardson, 'An ethical framework for global vaccine allocation', *Science*, **369**(6509), 1309–1312, (2020).

[23] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, 'Certifying and removing disparate impact', in *21st International Conference on Knowledge Discovery and Data Mining*, KDD '15, p. 259–268, (2015).

[24] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness, 2016.

[25] Bryce Goodman and Seth Flaxman, 'European union regulations on algorithmic decision making and a "right to explanation"', *AI Magazine*, **38**(3), 50–57, (2017).

[26] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers, 'A practical guide to multi-objective reinforcement learning and planning', in *AAMAS*, volume 36, p. 26, (2022).

[27] Luke Holman, Devi Stuart-Fox, and Cindy E Hauser, 'The gender gap in science: How long until women are equally represented?', *PLOS Biology*, **16**(4), 1–20, (2018).

[28] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth, 'Fairness in Reinforcement Learning', in *ICML*, eds., Doina Precup and Yee Whye Teh, volume 70 of *Proceedings of Machine Learning Research*, pp. 1617–1626, Sydney, Australia, (06–11 Aug 2017). PMLR.

[29] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth, 'Fairness in Learning: Classic and contextual bandits', *Advances in Neural Information Processing Systems*, **29**, 325–333, (2016).

[30] Faisal Kamiran and Toon Calders, 'Classifying without discriminating', in *2nd International Conference on Computer, Control and Communication*, pp. 1–6, (2009).

[31] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu, 'Preventing fairness gerrymandering: Auditing and learning for subgroup fairness', in *ICML*, eds., Jennifer Dy and Andreas Krause, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, (10–15 Jul 2018).

[32] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao, 'Nonconvex optimization for regression with fairness constraints', in *ICML*, eds., Jennifer Dy and Andreas Krause, volume 80 of *Proceedings of Machine Learning Research*, pp. 2737–2746, Stockholm, Sweden, (10–15 Jul 2018). PMLR.

[33] Dorothea Kübler, Julia Schmid, and Robert Stüber, 'Gender Discrimination in Hiring Across Occupations: A Nationally-Representative Vignette Study', *Labour Economics*, **55**(October), 215–229, (2018).

[34] Bertrand Lebichot, Fabian Braun, Olivier Caelen, and Marco Saerens, 'A graph-based, semi-supervised, credit card fraud detection system', in *Complex Networks & Their Applications V: Proceedings of the 5th International Workshop on Complex Networks and their Applications*, pp. 721–733. Springer, (2017).

[35] Pieter J. K. Libin, Arno Moonens, Timothy Verstraeten, Fabian Perez-Sanjines, Niel Hens, Philippe Lemey, and Ann Nowé, 'Deep reinforcement learning for large-scale epidemic control', in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, eds., Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke, pp. 155–170, Cham, (2021). Springer International Publishing.

[36] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt, 'Delayed impact of fair machine learning', in *ICML*, volume 80, pp. 3150–3158, Stockholm, Sweden, (2018). PMLR.

[37] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng, *Balancing Between Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning*, volume 12084 LNAI, Springer International Publishing, Cham, 2020.

[38] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of ML fairness notions, 2020.

[39] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, 'A survey on bias and fairness in machine learning', *ACM Comput. Surv.*, **54**(6), (jul 2021).

[40] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1 edn., 1997.

[41] Fatemeh Nargesian, Abolfazl Asudeh, and HV Jagadish, 'Tailoring data source distributions for fairness-aware data integration', *Proceedings of*

the VLDB Endowment, **14**(11), 2519–2532, (2021).
[42] Dennis Soemers, Ann Nowé, Tim Brys, Kurt Driessens, and Mark Winands, 'Adapting to concept drift in credit card transaction data streams using contextual bandits and decision trees', *AAAI*, **32**(1), 7831–7836, (2018).
[43] Javier Ortiz Laguna, Angel García Olaya, and Daniel Borrajo, 'A dynamic sliding window approach for activity recognition', in *User Modeling, Adaption and Personalization*, eds., Joseph A. Konstan, Ricardo Conejo, José L. Marzo, and Nuria Oliver, pp. 219–230, Berlin, Heidelberg, (2011). Springer Berlin Heidelberg.
[44] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari, 'Achieving fairness in stochastic multi-armed bandit problem', *arXiv*, (2020).
[45] Pascale Petit, 'The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector', *Labour Economics*, **14**(3), 371–391, (2007).
[46] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé, 'Pareto conditioned networks', in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, p. 1110–1118, Richland, SC, (2022). International Foundation for Autonomous Agents and Multiagent Systems.
[47] Mathieu Reymond, Conor F Hayes, Lander Willem, Roxana Rădulescu, Steven Abrams, Diederik M Roijers, Enda Howley, Patrick Mannion, Niel Hens, Ann Nowé, and Pieter Libin, 'Exploring the pareto front of multi-objective covid-19 mitigation policies using reinforcement learning', *arXiv preprint arXiv:2204.05027*, (2022).
[48] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar, 'Guaranteeing the Learning of Ethical Behaviour through Multi-Objective Reinforcement Learning', *ALA*, (2021).
[49] Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson, 'The Diverse Cohort Selection Problem', *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, **2**, 601–609, (2019).
[50] Candice Schumann, Jeffrey S. Foster, Nicholas Mattei, and John P. Dickerson, 'We need fairness and explainability in algorithmic hiring', *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, **2020-May**(Aamas), 1716–1720, (2020).
[51] Umer Siddique, Paul Weng, and Matthieu Zimmer, 'Learning fair policies in multiobjective (Deep) reinforcement learning with Average and Discounted Rewards', *ICML*, **119**, 8864–8874, (13–18 Jul 2020).
[52] STATBEL. Employment and unemployment, 2023. https://statbel.fgov.be/en/themes/work-training/labour-market/employment-and-unemployment#figures.
[53] STATBEL. Transitions on the labour market, 2023. https://statbel.fgov.be/en/themes/work-training/labour-market/transitions-labour-market#figures.
[54] STATBEL. Volwasseneneducatie, 2023. https://statbel.fgov.be/nl/themas/werk-opleiding/opleidingen-en-onderwijs/volwasseneneducatie#figures.
[55] Richard S. Sutton, Andrew G. Barto, and et al, *Reinforcement Learning : An Introduction*, MIT Press, 2018.
[56] Hannah Van Borm and Stijn Baert, 'Diving in the Minds of Recruiters: What Triggers Gender Stereotypes in Hiring?', *SSRN Electronic Journal*, (15261), (2022).
[57] Jiayin Wang, Weizhi Ma, Jiayu Li, Hongyu Lu, Min Zhang, Biao Li, Yiqun Liu, Peng Jiang, and Shaoping Ma, 'Make fairness more fair: Fair item utility estimation and exposure re-distribution', in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1868–1877, (2022).
[58] Paul Weng, 'Fairness in reinforcement learning', *CoRR*, **abs/1907.10323**, (2019).
[59] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork, 'Learning fair representations', in *ICML*, eds., Sanjoy Dasgupta and David McAllester, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, (17–19 Jun 2013). PMLR.
[60] Luisa M Zintgraf, Edgar A Lopez-Rojas, Diederik M Roijers, and Ann Nowé, 'Multimaus: a multi-modal authentication simulator for fraud detection research', in *29th European Modeling and Simulation Symp.(EMSS 2017)*, pp. 360–370. Curran Associates, Inc., (2017).
[61] Eva Zschirnt and Didier Ruedin, 'Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015', *Journal of Ethnic and Migration Studies*, **42**(7), 1115–1134, (2016).

A Job hiring fMDP

For the job hiring setting, we create a simulator for building a team of employees that supplies at each timestep a new candidate to the agent. To apply RL, we define the job hiring setting as a Markov Decision Process (MDP) [55]. The MDP is represented by the tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, p\}$, consisting of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a set of rewards $\mathcal{R}$ and a transition function p.

State Each timestep t, the agent is presented with the current state $\mathbf{s}_t \in \mathcal{S}$, which specifies the company's current composition p_t of hired applicants and a new job applicant c_t to assess. A job applicant c_t is represented by the following set of features: their gender, age, years of experience, degree, extra degree, marital status, nationality and their ability to speak four languages $\{l_{dutch}, l_{french}, l_{english}, l_{german}\}$. For the purpose of this study, we consider gender, nationality and marital status sensitive features, which should not be taken into account when hiring a job applicant. To generate realistic applicants, we sample from the distribution of the Belgian active employed and unemployed population provided by the Belgian federal government [52]. For the context of our job hiring scenario, we exclude individuals younger than 18 years from this data. To assign spoken languages to the candidates, we sample based on the most known foreign languages of adults [54]. We define the maximum experience of each applicant in function of their age and obtained degrees: $max_e = age - 18 - 3 * degree - 2 * extra\_degree$. We assume a linearly increasing probability for each possible year of experience $year \in [0, max_e]$ for the applicant, equal to

$$P(year) = \frac{year + 1}{\sum_{y=0}^{max_e}(y+1)} \tag{16}$$

The company's state p is represented by a set of features focusing on the employees' skills. These features consist of the average employee potential P, the percentage of collected degrees, extra degrees, the combined years of experience and language entropy. We normalise all features based on the desired final team size K, such that each applicant can impact the team as much as they would in a full team. We further normalise the combined years of experience such that all features lie in the interval $[0, 1]$. Based on hired applicants, the company's team composition p_t is implemented as the proportions of skill and diversity features. For example, the language diversity is represented by four values [0.6, 0.4, 0.2, 0.1] indicating 60% of the spoken languages is Dutch, 40% is French, 20% is English and 10% is German. On these values the entropy is calculated for the goodness score and reward. Therefore, the state does not contain a list of all employees, but does contain their contributions to the team's skills such that the agent can decide for a new candidate if they are a good fit. Given K the desired final team size and k the number of employees (i.e., hired applicants), we define the company's potential based on the degree d, extra degree e and experience x each employee holds on average. Concretely, the potential of the employees follows a Gaussian with mean

$$P = \frac{1}{K}\sum_{i=1}^{k}\frac{1}{3}|\{f^i \in \{d, e, x\} : f^i \neq 0\}| \tag{17}$$

and a standard deviation of 0.01. For the estimated company potential given a new applicant, we use the same distribution given the assumption that an applicant's resume based on these features does not perfectly match the applicant's potential once hired for the job.

Goodness score To define how suitable each candidate is for hire, we define an objective goodness score $G_t \in [-1, 1]$ based on how the estimated new company state $\hat{p}_{t+1}$ would differ from the current p_t, should the applicant be hired:

$$G_t = \frac{K}{N}\sum_{f_t \in p_t}(\hat{f}_{t+1} - f_t) \tag{18}$$

with N the number of skill features. Note that this goodness score is also noisy due to the noise in the current company potential and the estimated new potential. Intuitively, the goodness score is higher for applicants who can improve the average potential, have the requested skills and improve the language entropy of the team.

Action and reward At each timestep t, the agent must choose whether to reject or hire the applicant for a given state $\mathbf{s}_t$. Given the chosen action a_t for state $\mathbf{s}_t$, the agent receives a reward r_t based on the goodness score G_t of the presented applicant. Given the goodness score G_t and threshold ϵ, the reward for hiring an applicant is

$$r_{t,hire} = G_t - \epsilon + \mathcal{N}(0, 0.01) \tag{19}$$

We add Gaussian noise to the reward under the assumption that the applicant's qualification may differ slightly from the estimation of the goodness score. This models the employer's uncertainty about the suitability of hired applicants. The reward for rejecting an applicant is the negative reward of hiring the applicant:

$$r_{t,reject} = -r_{t,hire} \tag{20}$$

Transition function We define the transition function $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ as the probability of encountering the next state $\mathbf{s}_{t+1}$ and reward r_t given the current state $\mathbf{s}_t$ and action a_t. To mimic a realistic team composition over time, we allow employees to leave the company based on real job transition probabilities corresponding to their age [53]. This provides the agent with the additional challenge of replacing lost skills of leaving employees to keep the team balanced.

Feedback signal To extend the MDP to an $f\text{MDP}$, we implement the feedback signal f_t as the correct action $\hat{a}_t$ based on the goodness score:

$$f_t = \hat{a}_t \tag{21}$$

B MultiMAuS fMDP

The fraud detection setting concerns online credit card transactions where multi-modal authentication is used to identify and reject fraudulent transactions. We make the following adaptations to the MultiMAuS simulator [60], but keep their default parameters.

State Each hour, a set of customers, both genuine and fraudulent, attempt to make transactions, where each transaction is characterised by the following features: card id, merchant id, amount, currency, country and the date and hour when the transaction is occurring. As the agent must check transactions on an individual basis, we consider a new timestep for every transaction request. At each timestep t, the agent observes the current state $\mathbf{s}_t$ containing information about the current company state, and a new transaction to process. We define two company state features: the proportion of genuine to fraud transactions and the average customer satisfaction.

Reward + action For each transaction, the agent must decide whether or not to request an authentication from the customer. Based on the chosen action a_t, the agent receives a reward

$$r_t = \begin{cases} +1 & \text{if genuine authentication,} \\ -1 & \text{otherwise} \end{cases} \tag{22}$$

Based on this reward, always asking for authentication results in more fraudulent transactions being caught, as fraudsters are assumed to not be able to provide a second authentication [60]. However, asking for authentication too often reduces the customer's patience in completing transactions. Furthermore, too many authentication requests make it more likely for customers to leave the credit card company. Therefore, the agent must carefully select transactions to check to keep customer satisfaction high, while also catching as many fraudulent transactions as possible.

Feedback signal The reward r_t specifies the correctness of the action if the agent requests authentication. Consequently, if the reward is positive the transaction is considered genuine, while a negative reward indicates an unsuccessful transaction, caused by a loss in commission or by stolen money requiring the credit company to repay the losses to the client. To implement a feedback signal f, we infer the correctness when authenticating to observe the amount of true positives and false positives.

C Pareto Conditioned Networks

A Pareto Conditioned Network (PCN) [46] applies supervised learning techniques to approximate all non-dominated policies within a single neural network. PCN takes as input a tuple $\langle \mathbf{s}, \hat{h}, \hat{\mathbf{R}} \rangle$, representing the observed state $\mathbf{s}$, the desired return $\hat{\mathbf{R}}$ to reach at the end of the episode and the desired horizon $\hat{h}$ indicating the number of timesteps that should be executed before reaching $\hat{\mathbf{R}}$. Both $\hat{h}$ and $\hat{\mathbf{R}}$ are chosen by the decision maker at the start of an episode. Consequently, at every timestep t, the desired return is updated by the received reward $\hat{\mathbf{R}} \leftarrow \hat{\mathbf{R}} - r_t$ and the desired horizon is decreased by one timestep $\hat{h} \leftarrow \hat{h} - 1$. PCN learns policies similar to classification techniques, where $\langle \mathbf{s}_t, h_t, \mathbf{R}_t \rangle$ is the input at timestep t and the chosen action a_t is the output. We employ a dense neural network with state, horizon and return embeddings, with each consisting of a hidden layer of 64 neurons and a sigmoid activation function. Their outputs are fed through a fully connected neural network of 2 layers with a RELU activation on the first layer. This last network produces outputs for each action.