A Group And Individual Aware Framework For Fair Reinforcement Learning

Alexandra Cimpean
Vrije Universiteit Brussel
Brussels, Belgium
ioana.alexandra.cimpean@vub.be

Catholijn Jonker
Technische Universiteit Delft
Delft, The Netherlands
c.m.jonker@tudelft.nl

Pieter Libin
Vrije Universiteit Brussel
Brussels, Belgium
pieter.libin@vub.be

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@vub.be

ABSTRACT

Real-world sequential decision problems can be approached using a reinforcement learning approach. When these problems impact fairness across groups or individuals, considering fairness-aware techniques is crucial. Therefore, we require algorithms that can make suitable trade-offs between performance and the desired fairness notions. As the desired performance-fairness trade-off is difficult to specify a priori, we propose a framework where multiple trade-offs can be explored. As such, insights provided by the reinforcement learning algorithm, regarding the obtainable performance-fairness trade-offs, can be used by stakeholders to select the best policy for the problem at hand. To capture the appropriate fairness notions, we define an extended Markov decision process, fMDP, that explicitly encodes individuals and groups. Given this fMDP, we formalise fairness notions in the context of sequential decision problems. We formulate a fairness framework, that allows us to compute fairness notions over time. We evaluate our framework in two scenarios, each with distinct fairness requirements. The first is a job hiring setting, where strong teams must be composed, while providing equal treatment to the applicants. The second setting concerns fraud detection, where fraudulent transactions must be detected, while ensuring the burden for customers is distributed fairly.

KEYWORDS

reinforcement learning, automated decision support, fairness framework, trustworthy AI

1 INTRODUCTION

A wide range of real-world decision problems may risk discrimination when solved, including job hiring [45, 46], epidemic mitigation [20, 32], finance [33] and fraud detection [48]. Therefore, they require solutions that focus on fairness. Moreover, real-world problems are often sequential, requiring automated decision support systems to continuously maintain the desired performance and adapt as needed to unseen situations. This demonstrates the need for automated decision support systems that learn policies that can balance performance with fairness over the impacted people.

Obtaining a policy with an appropriate performance-fairness trade-off is context-specific and typically relies on multiple fairness notions to guarantee fairness [35]. Therefore, defining appropriate performance-fairness trade-offs requires input from stakeholders and domain experts. Moreover, even with stakeholders' preferences and domain expertise, the boundary between fair and unfair and other ethical considerations is difficult to decide on a priori. Consequently, selecting a suitable policy is challenging. To this end, we need a framework capable of dealing with multiple fairness notions simultaneously. Furthermore, it is crucial to provide stakeholders with an overview of which trade-offs are possible, such that an informed decision can be made on which policy is most suited for the problem at hand.

Previous work related to fairness mainly focused on the supervised learning setting that operates on a given dataset, such as machine learning [18, 19, 22, 35, 36] and data mining [8, 21, 28]. In contrast, automated decision support problems are typically sequential and can evolve over time. This requires dealing with the impact of short term and long term decisions [15]. As such, reinforcement learning (RL) is a suitable approach to enable an agent to learn a decision support policy by interacting with an environment [50]. 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 balances between exploration and exploitation to reach an optimal policy [50].

RL approaches have focused on fairness in application-specific solutions [10, 26, 27, 44, 47, 53]. However, these solutions focus on a single fairness notion and rely on reward shaping to define the performance-fairness trade-off [10, 34]. However, as the desired performance-fairness trade-off cannot be described a priori by stakeholders, these approaches do not suffice for most real-world problems. Furthermore, dealing with the performance and one or multiple, possibly conflicting, fairness notions requires a multi-objective approach to explore the uncertainty over the obtainable trade-offs [24]. To this end, we propose a formal framework that can learn performance-fairness trade-offs regarding multiple fairness notions. We evaluate this framework in two settings that have distinct fairness requirements: job hiring and credit card fraud detection.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2024), Avalos, Müller, Wang, Yates (eds.), May 6-7, 2024, Online, https://ala2024.github.io/. 2024.

2 RELATED WORK

As reinforcement learning approaches are well suited to deal with sequential processes, new research has focused on multi-armed bandit approaches [27, 40]. To enforce fairness in job hiring, multi-armed bandits [45, 46] as well as generalisations towards MDP approaches [26] 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 [13, 14, 31], which are retrained as labelled data becomes available with a delay. Soemers et al. [48] 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 [20, 32]. Reymond et al. [43] 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. [29] 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.

3 FAIRNESS FRAMEWORK

To define the fairness framework, we first 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 qualified candidates while limiting bias towards sensitive features. The second setting involves fraud detection, where fraudulent transactions must be efficiently flagged, taking into account that verification requires a costly human effort. Moreover, 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. In this paper we make use of simulated data based on real data distributions.

3.1 fMDP and the fairness history

A sequential decision process can be formally described as a Markov Decision Process (MDP) [50], 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. To ensure that existing fairness notions can be defined in this sequential process, we extend this default MDP to an fMDP. 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. To this end, the fMDP adds a feedback signal f_t, that concerns an indication whether the chosen action a_t was correct at time t. In the fraud detection setting, this signal would indicate for a fraudulent transaction that the correct action would have been to request a reauthentication. Note that this feedback is optional and can be partial, sparse or delayed.

As existing fairness notions typically concern fair treatment between individuals or groups, we introduce the following notation. $\mathcal{I}_t$ refers to the set of individuals involved in the decision process at time t and we use $i_t \in \mathcal{I}_t$ 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) should be made. In the fraud detection setting, $\mathcal{I}_t$ refers to all customers at time t to be considered for verification. We use

$$\mathcal{I}^t = \bigcup_{t'=0}^{t} \mathcal{I}_{t'} \tag{1}$$

to refer to the set of all individuals involved in the decision process from the start $t' = 0$ up to time t. We define all individuals of $\mathcal{I}_t$ belonging to group g as $\mathcal{G}_{g,t} \subseteq \mathcal{I}_t$. We refer to all individuals of group g, involved in the decision process until time t, as:

$$\mathcal{G}_g^t = \bigcup_{t'=0}^{t} \mathcal{G}_{g,t'} \tag{2}$$

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 not discriminate when flagging transactions.

To define individuals and groups in terms of the fMDP, we use the following operator [ ]. Concretely, a state $\mathbf{s}_t$ encodes the individuals $\mathcal{I}_t$ and groups $\mathcal{G}_t$ involved in the decision at time t. We use $\mathcal{I}_t[\mathbf{s}_t]$ and $\mathcal{G}_t[\mathbf{s}_t]$ to refer to the individuals and groups from the state, respectively. Furthermore, the action a_t encodes the decision impacting the involved individuals and groups ($\mathcal{I}_t[a_t]$ and $\mathcal{G}_t[a_t]$), and the feedback f_t specifies the correctness of that decision ($\mathcal{I}_t[f_t]$ and $\mathcal{G}_t[f_t]$).

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

$$\mathcal{H}^t = \{\mathbf{s}_{t'}, a_{t'}, r_{t'}, f_{t'}\}_{t'=0}^{t} \tag{3}$$

The history encodes the individuals and groups encoded in each interaction it stores. Given the operator [ ], we use $\mathcal{I}^t[\mathcal{H}]$ and $\mathcal{G}^t[\mathcal{H}]$ to refer to individuals and groups from all interactions until time t in the history $\mathcal{H}$, respectively.

For ease of notation, we define the encountered states and selected actions from history $\mathcal{H}$ respectively as $\mathcal{H}_S$ and $\mathcal{H}_A$. We refer to feedback regarding the correctness of the action as $\mathcal{H}_f$. In the job hiring setting, the history consists of the encountered 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.

3.2 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 $\mathcal{H}_S$, chosen actions $\mathcal{H}_A$ and feedback $\mathcal{H}_f$ until time t:

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

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

Concretely, each fairness notion $\mathscr{F}$ is defined as the negative absolute difference in treatment between groups or individuals. Therefore, when $\mathscr{F} = 0$, the agent has achieved exact fairness with respect to the given fairness notion. An exact definition of $\mathscr{F}$ can be intractable due to limitations of defining exact fairness [26]. In such cases, 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 concerning the correctness of an action 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* [18], where the probability of receiving the preferable treatment of the agent ($\mathcal{H}_A = 1$) should be the same across groups g and h:

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

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*:

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

where $\mathcal{H}_f = 1$ 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, to compute 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 individuals 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 [18]. Given two individuals i_t and j_t, we assume a distance metric $d(i_t, j_t)$ between the individuals. Note that also a similarity metric could be used. Given the probability distributions M_i and M_j of the agent's policy 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:

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

Individual fairness notions assume that an appropriate distance metric is chosen based on domain expertise [35]. For the purpose of evaluating our framework, we have chosen a distance metric where we exclude the sensitive features, such that similarity is defined only in terms of non-sensitive features. However, we emphasise that for deploying algorithms in the real-world, this decision must be made by stakeholders. To distinguish between nominal features (e.g., ability to speak a language) and numerical features (e.g., years of experience), we employ the Heterogeneous Manhattan-Overlap Metric (HMOM) [54]. Concretely, we define a distance metric using HMOM in the interval $[0, 1]$ as:

$$d(i_t, j_t) = e^{-\lambda HMOM(i_t, j_t)} \tag{9}$$

where $\lambda > 0$ is a smoothing parameter. We use an exponential function to output values in the range $[0, 1]$, such that all compared individuals have the same maximum impact on the outcome of the individual fairness notion.

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 [18]. 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 [24].

3.3 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 [56]. Throughout the day, each individual client may decide to make transactions. The agent aims to flag suspicious transactions, in a way that every continent is subject to a similar proportion of re-authentication requests.

Suppose in our fraud detection setting, that each hour the agent encounters transactions from different continents. Each hour, the agent chooses how to flag transactions for these 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 all individuals $\mathcal{I}_t$ can be grouped 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 would be the case due to time zone differences. Then a sufficiently long time horizon must be considered to encounter all continents. Concretely, if the state $\mathbf{s}_t$ contains only information on a strict subset $\mathcal{B}_t \subset \mathcal{G}_t$ of the respective groups impacted by the decision at time t, a fairness notion can only be defined over the history $\mathcal{H}$ until t, to contain sufficient information about all impacted $\mathcal{G}_t$ groups for time t:

$$\mathcal{G}_t[\mathbf{s}_t] = \mathcal{G}^t[\mathcal{H}_S] \tag{10}$$

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

$$\mathcal{G}_t[a_t] = \mathcal{G}^t[\mathcal{H}_A] \tag{11}$$

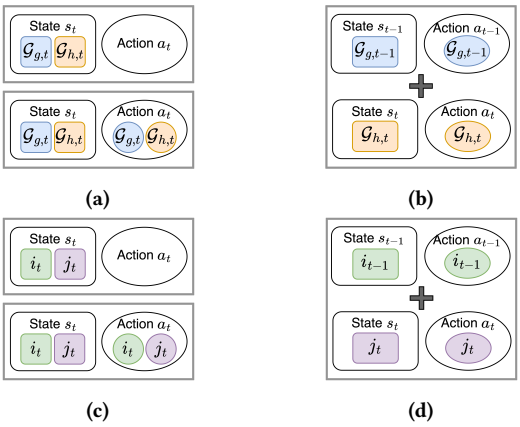


Figure 1: (a) (b) Scenarios where group fairness can be calculated. (a) All groups $\mathcal{G}_t$ are encountered at each time t. Top: action a_t is an action over all groups $\mathcal{G}_t$. Bottom: action a_t encodes a specific action for each group g. (b) All groups $\mathcal{G}_t$ are encountered 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.

If individuals are defined within the state representation of the environment, Equations 10 and 11 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.

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 strict subset $C_t \subset \mathcal{I}_t$, an individual fairness notion can be defined over a history $\mathcal{H}$ until t when subsets of the individuals are encountered:

$$\mathcal{I}_t[\mathbf{s}_t] = \mathcal{I}^t[\mathcal{H}_S] \tag{12}$$

$$\mathcal{I}_t[a_t] = \mathcal{I}^t[\mathcal{H}_A] \tag{13}$$

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.

We consider each fairness notion by computing its approximation $\hat{\mathcal{F}}$, 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 [26]. 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 [38] 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 [52]. We consider such approaches as future work to learn better approximations for fairness notions over a full history.

3.4 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. To support decision makers, policies can be

learned in simulated environments or directly in the real-world. This choice depends on the problem at hand, and particularly how the agent's actions would impact the groups and individuals. Furthermore, the availability of a simulated environment may provide insights on which performance-fairness trade-offs are possible 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.

4 SCENARIOS

In this section, we introduce a job hiring and fraud detection scenario, that we use in our experiments, along with their distinct fairness implications.

4.1 Job hiring

Job hiring is a reoccurring process 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 [35]. 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 [4, 51] to occupational-specific characteristics [1, 2, 12, 25, 30]. Ethnic discrimination has been studied from an immigration perspective [57] and is based on implicit interethnic attitudes [6]. Moreover, combinations of sensitive features are known to cause discrimination [3, 17, 41].

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 <country> population [Omitted for anonymity reasons]. 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 3.2). 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 $\mathcal{G}^t_{men}$ or women $\mathcal{G}^t_{women}$. 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 between applicants as in Equation 8. We set $\lambda = 0.1$ for the heterogeneous distance metric.

4.2 Fraud detection

Fraudulent credit card transactions result in significant losses when undetected [16]. 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 [14], 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 [56]. 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 [35]. 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 [56], 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 successful 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.

$$\begin{aligned}\mathscr{F} = -|&\mathrm{P}(\mathcal{G}^t_{C_a}[\mathcal{H}_A] = \mathcal{G}^t_{C_a}[\mathcal{H}_f] | \mathcal{G}^t_{C_a}[\mathcal{H}_S]) \\ &- \mathrm{P}(\mathcal{G}^t_{C_b}[\mathcal{H}_A] = \mathcal{G}^t_{C_b}[\mathcal{H}_f] | \mathcal{G}^t_{C_b}[\mathcal{H}_S])|\end{aligned} \tag{14}$$

Predictive parity [11] 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.

$$\begin{aligned}\mathscr{F} = -|&\mathrm{P}(\mathcal{G}^t_{C_a}[\mathcal{H}_f] = 1 | \mathcal{G}^t_{C_a}[\mathcal{H}_A] = 1, \mathcal{G}^t_{C_a}[\mathcal{H}_S]) \\ &- \mathrm{P}(\mathcal{G}^t_{C_b}[\mathcal{H}_f] = 1 | \mathcal{G}^t_{C_b}[\mathcal{H}_A] = 1, \mathcal{G}^t_{C_b}[\mathcal{H}_S])|\end{aligned} \tag{15}$$

In fraud detection, we define individual fairness between transactions using the complement of the consistency score [55]:

$$\mathscr{F} = -\frac{1}{||\mathcal{I}^t||} \sum_{i \in \mathcal{I}^t} \frac{1}{k} |a^i - \sum_{j \in kNN(i)} a^j| \tag{16}$$

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 [37]. We assume the same distance metric and $\lambda = 0.1$ as for individual fairness.

5 RESULTS

As both scenarios deal with a reward and multiple fairness objectives, the number of policies with suitable trade-offs can scale exponentially. To learn all policies would therefore be computationally intractable, to explore the entire state space. To this end, we use Pareto Conditioned Networks (PCN) [42]. 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 all experiments, we report the learned non-dominated coverage sets for all objectives [24]. As the number of trade-off policies learned is quite high, we selected a representative subset of policies which provide a good approximation of the Pareto front the figures below. We provide additional details on the visualisation in Appendix D. The reward vector consists of the following objectives: the performance reward (R), statistical parity (SP), equal opportunity (EO), overall accuracy equality (OAE), predictive parity (PP), individual fairness (IF), consistency score complement (CSC). Note that the fairness notions EO, OAE and PP require access to the ground truth or a proxy to be computed. We present results for 10 seeds per experiment of 500 000 timesteps and implement the fairness history as a sliding window of 500 timesteps. We use the HEOM distance metric for the individual fairness notions.

5.1 Job hiring

For the job hiring scenario, we train an agent to hire and maintain a well-performing team of 100 employees, where each episode lasts for a maximum of 1000 timesteps. We consider the Belgian population, informed by the official statistics registry of Belgium, STATBEL [49], where the agent must be fair towards men and women in the hiring process.

First, we consider the cases where the agent optimises for a single objective. Figure 2 shows the results for building a team of 100 employees. When asked to optimise the performance reward, the agent obtains a reward close to the maximum (0). When the agent is requested to learn to optimise one of the group fairness notions, the other group fairness notions also receive a close to zero score. This can be explained as most group fairness notions require the confusion matrix to be computed, there are overlaps regarding the involved true and predicted actions. Or phrased differently, the group fairness notions are, in this use case, quite well aligned, and it is feasible to optimise more than one group fairness notion simultaneously.

In contrast, individual fairness notions make pair-wise comparisons of similar individuals. Concretely, it is possible for the agent to find larger differences in the non-dominated values, as IF considers the probability distributions over the actions, while CSC only considers the action. Optimising for any fairness notion results in a lower performance reward. We observe lower individual fairness when optimising for the reward or some group fairness notions in particular.

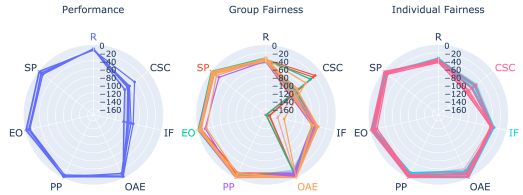


Figure 2: Representative set of learned hiring policies when optimising a single objective, split per type of objective. Left: Optimising for the performance reward. Center: Optimising a group fairness notion. Right: Optimising an individual fairness notion. Different fairness notions are indicated by different colours. Lines in the same colour represent outcomes of different runs.

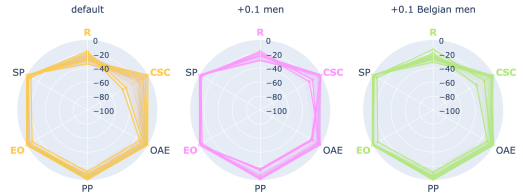


Figure 3: Representative set of learned hiring policies when simultaneously optimising R, EO and CSC with different trade-offs under three different reward configurations. Note that SP, PP and OAE are not used in the optimisation process, but are included in the plots for informative reasons. Due to computational reasons, we only show the requested individual fairness notion CSC. In Appendix E we provide additional results when learning policies that include IF instead.

To investigate the strengths and shortcomings of all fairness notions, we consider two additional reward configurations next to the currently objective configuration, which assigns a similar reward to individuals who apply to the same team with the same qualifications. The second configuration assigns a +0.1 bias to men, while the third configuration assigns a +0.1 bias to Belgian men. Figure 3 shows the learned policies when optimising for R, EO and CSC for each reward configuration. As in the previous results, the agent achieves high group fairness for most of the policies. Depending on the (lack of) bias, the agent finds different trade-offs. The policies that maximise the reward in the default unbiased configuration result in the lowest individual fairness. In contrast, the agent is able to obtain higher individual fairness in the biased configurations. As CSC compares individuals directly, it is better suited when dealing with multiple sensitive attributes. Consequently, by optimising for

CSC, the agent is able to learn policies that reach a high individual fairness in all 3 reward configurations. However, there is a notable difference in the group fairness notions PP and OAE in the +0.1 men configuration. Specifically, there is a larger difference in the probability of being a qualified applicant when hired, as well as the accuracy for hiring and rejecting applicants. This difference in treatment is undetected in the +0.1 Belgian men configuration, as the group notions do not consider sub-group implications, allowing the agent to gerrymander [29]. In the +0.1 Belgian men configuration, CSC detects unfairness if the policy prioritises the biased reward, while the group fairness notions cannot. We provide additional results in Appendix E.

Figure 4 shows the learned multi-objective policies, when optimising the performance reward, a group fairness notions and an individual fairness notion simultaneously. In general, all group fairness notions, including the ones who were not initially requested, are easy for the agent to optimise. We observe the largest differences in the learned trade-offs with regards to the reward and either individual fairness notion. Note that the requested group fairness notion does impact which trade-offs can be found, indicating the combination of requested fairness notions influences the overall fairness a policy can provide.

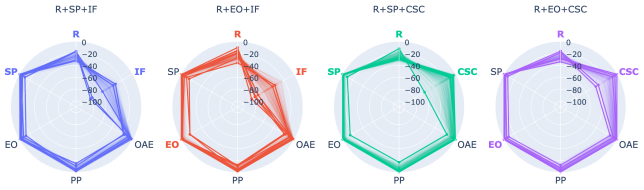
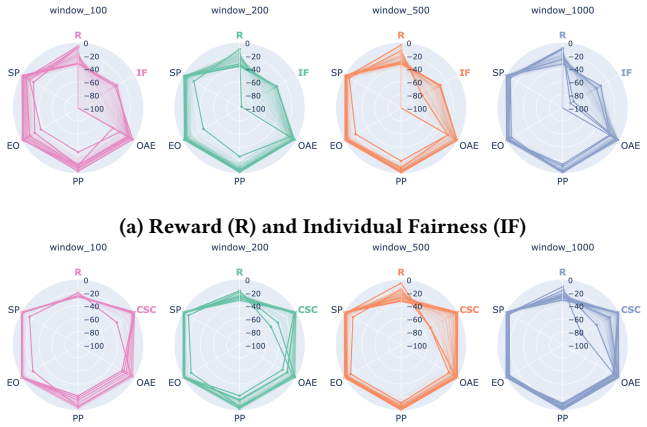


Figure 4: Representative set of learned hiring policies when optimising the reward (R), a group fairness notion (SP or EO) and an individual fairness notion (IF or CSC).

Figure 5 shows the impact of different history sizes on the learned trade-off policies. In general, we observe that group fairness benefits from a larger history. As group fairness notions focus more on statistical measures over groups of individuals, it is easier for the agent to provide equal treatment over the groups. In contrast, individual fairness (IF and CSC) are more impacted by the trade-offs with the performance reward (R). Note that for both individual fairness notions, across all 4 window sizes, the reward can only be improved at the cost of individual fairness and vice versa. This indicates that these objectives may be conflicting.

5.2 Fraud detection

For the fraud detection scenario, we assume the default parameters of the MultiMAuS simulator [56], but increase the frequency of fraudulent transactions to ensure enough genuine and fraudulent transactions are encountered for continents C_a and C_b. This results in approximately 10% fraudulent transactions. Note that continents C_a and C_b have different base rates of fraudulent transactions. We let the agent check transactions for a week, resulting in at most 1000 transactions per episode, where the agent must be fair towards requesting re-authentications from both continents.



(a) Reward (R) and Individual Fairness (IF)

(b) Reward (R) and Consistency Score Complement (CSC)

Figure 5: Representative set of learned hiring policies when optimising the reward and an individual fairness notion. Showing results for histories with different sliding window sizes.

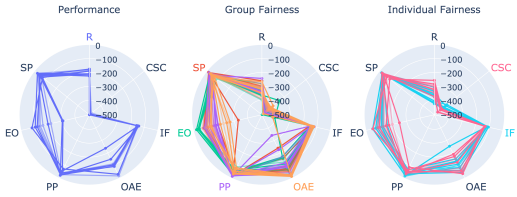


Figure 6: Representative set of learned fraud detection policies optimising a single objective, split per objective type. Left: Optimising for the performance reward. Center: Optimising a group fairness notion. Right: Optimising an individual fairness notion. Different fairness notions are indicated by different colours. Lines in the same colour represent outcomes of different runs.

Figure 6 shows the learned single-objective policies. When optimising for the reward (i.e., detecting fraudulent transactions), small variations in the reward lead to high variations with respect to EO and OAE. This indicates that interesting trade-offs can be found in this use case. Overall, we observe that the agent is unable to find policies which maximise CSC, regardless of the objective. We attribute this to the difficulty of the environment, when it comes to treating similar transactions similarly. We hypothesise that this may be caused by a correlation between the action and the sensitive features of the transaction, including the continent where the transaction originates from.

As the choice of sliding window impacts the calculation of all fairness notions, we consider additional window sizes. For a multi-objective approach, we ask the agent to optimise for R and CSC. In Figure 7, we note that using a smaller window size makes it more difficult to maximise both group and individual fairness notions. Note that CSC is still low compared to the other objectives, indicating the performance reward is conflicting with the agent's

fairness. The largest contributor to this effect is the different base rates for fraudulent transactions between individuals. While the policies improve both R and CSC compared to the single-objective results, we again notice a larger variation with the EO and OAE fairness notions. 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 EO requires that fraudulent transactions are flagged with the same probability across continents. We provide additional results in Appendix F.

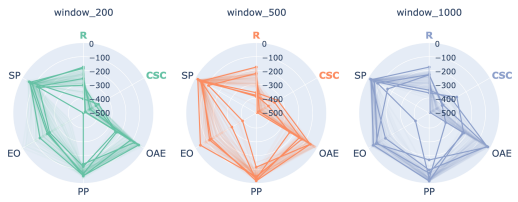


Figure 7: Representative set of fraud detection policies when optimising R and CSC under different history sizes.

In Figure 8, we note that most trade-offs have the largest differences across the reward R, EO and OAE. Note how individual fairness does not change as much regardless of the requested objectives. As such, we opted to present the results on CSC. We emphasise how the combination of requested objectives impact the obtainable trade-offs. This further highlights the need for multiple fairness notions, which should be chosen by stakeholders with the necessary domain expertise.

While the reward (R) and consistence score complement (CSC) were easier to optimise in the job hiring setting, we observe the opposite effect in the fraud detection setting. We hypothesise this is caused by the context of the problem. Concretely, in job hiring it is easier to find applicants with similar attributes that must be treated similarly. For example, individuals with the same qualifications should receive the same decision as their most similar neighbours under CSC. As IF considers the probability distribution over the actions, and compares all individuals, it is more difficult for the agent to provide the appropriate treatment. In contrast, in the fraud detection setting it is not necessarily the case that similar transactions are all fraudulent or all genuine. This makes it more difficult to ensure equal treatment with regards to CSC. Moreover, fraud detection constitutes anomalies, making it more likely that most transactions are ignored, providing better (but possibly misleading) results for IF.

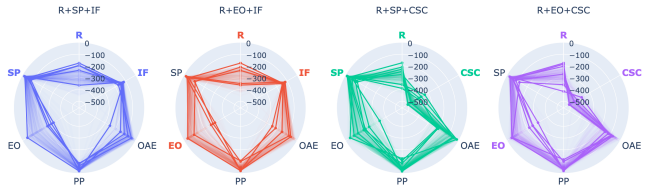


Figure 8: Representative set of learned fraud detection policies when optimising the reward (R), a group fairness notion (SP or EO) and an individual fairness notion (IF or CSC).

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 performance reward. 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 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 varying number of steps [39]. Another path is to explore the use of distinct neural sub-networks for approximating different fairness notions directly. We highlight that the size of the history further influences the fairness, especially when considering multiple fairness notions.

Within the overarching topic of ethics, work on explainable AI focuses on making algorithms interpretable and provides explanations for their decisions [23]. While explainability aims to provide transparency regarding 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.

ACKNOWLEDGMENTS

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 EU Horizon 2020 research and innovation programme under GA Numbers 952215 (TAILOR), and 820437 (HumaneAI Net), and by the National Science Foundation (NWO) under Grant Numbers 024.004.022 (Hybrid Intelligence), 024.004.031 (ESDiT), and 024.005.017 (ALGOSOC). Ann Nowé acknowledges support from FWO grant G062819N. All experiments were performed on the VSC high performance computing infrastructure [9].

REFERENCES

[1] Mladen Adamovic and Andreas Leibbrandt. 2023. A large-scale field experiment on occupational gender segregation and hiring discrimination. *Industrial Relations* 62, 1 (2023), 34–59. https://doi.org/10.1111/irel.12318

[2] Ali Ahmed, Mark Granberg, and Shantanu Khanna. 2021. Gender discrimination in hiring: An experimental reexamination of the Swedish case. *PLOS ONE* 16, 1 (2021), 1–15. https://doi.org/10.1371/journal.pone.0245513

[3] Stijn Baert. 2018. *Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005.* Springer International Publishing, Cham, 63–77. https://doi.org/10.1007/978-3-319-71153-9_3

[4] Kai Barron, Ruth Ditlmann, Stefan Gehrig, and Sebastian Schweighofer-Kodritsch. 2022. Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment. *SSRN Electronic Journal* 9731 (2022). https://doi.org/10.2139/ssrn.4097858

[5] Richard A. Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50 (2018), 3 – 44.

[6] Lieselotte Blommaert, Frank van Tubergen, and Marcel Coenders. 2012. Implicit and explicit interethnic attitudes and ethnic discrimination in hiring. *Social Science Research* 41, 1 (2012), 61–73. https://doi.org/10.1016/j.ssresearch.2011.09.007

[7] Miranda Bogen and Aaron Rieke. 2018. *Help wanted: An examination of hiring algorithms, equity, and bias.* Technical Report.

[8] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *31st International Conference on NIPS*. 3995–4004.

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

[10] Jingdi Chen, Yimeng Wang, and Tian Lan. 2021. Bringing fairness to actor-critic reinforcement learning for network utility optimization. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications* (Vancouver, BC, Canada). IEEE Press, Vancouver, BC, Canada, 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488823

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

[12] Clara Cortina, Jorge Rodríguez, and M. José González. 2021. Mind the Job: The Role of Occupational Characteristics in Explaining Gender Discrimination. *Social Indicators Research* 156, 1 (2021), 91–110. https://doi.org/10.1007/s11205-021-02646-2

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

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

[15] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Barcelona, Spain, 525–534. https://doi.org/10.1145/3351095.3372878

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

[17] Eva Derous and Roland Pepermans. 2019. Gender discrimination in hiring: Intersectional effects with ethnicity and cognitive job demands. *Archives of Scientific Psychology* 7 (2019), 40–49. https://doi.org/10.1037/arc0000061

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[19] Cynthia Dwork, Christina Ilvento, Guy N. Rothblum, and Pragya Sur. 2020. Abstracting Fairness: Oracles, Metrics, and Interpretability. In *1st Symposium on Foundations of Responsible Computing*. Curran Associates, Inc., Cambridge, MA, USA, 16. https://doi.org/10.4230/LIPIcs.FORC.2020.8

[20] 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. 2020. An ethical framework for global vaccine allocation. *Science* 369, 6509 (2020), 1309–1312. https://doi.org/10.1126/science.abe2803

[21] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *21st International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 259–268. https://doi.org/10.1145/2783258.2783311

[22] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. , 16 pages. arXiv:1609.07236

[23] Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741 arXiv:1606.08813

[24] 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. 2022. A practical guide to multi-objective reinforcement learning and planning. In *AAMAS* (2022/04/13), Vol. 36. 26.

[25] Luke Holman, Devi Stuart-Fox, and Cindy E Hauser. 2018. The gender gap in science: How long until women are equally represented? *PLOS Biology* 16, 4 (2018), 1–20. https://doi.org/10.1371/journal.pbio.2004956

[26] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in Reinforcement Learning. In *ICML (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, Sydney, Australia, 1617–1626. https://proceedings.mlr.press/v70/jabbari17a.html

[27] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems* 29 (2016), 325–333. arXiv:1605.07139

[28] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*. 1–6. https://doi.org/10.1109/IC4.2009.4909197

[29] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *ICML (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2564–2572.

[30] Dorothea Kübler, Julia Schmid, and Robert Stüber. 2018. Gender Discrimination in Hiring Across Occupations: A Nationally-Representative Vignette Study. *Labour Economics* 55, October (2018), 215–229. https://doi.org/10.1016/j.labeco.2018.10.002

[31] Bertrand Lebichot, Fabian Braun, Olivier Caelen, and Marco Saerens. 2017. 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*. Springer, 721–733.

[32] Pieter J. K. Libin, Arno Moonens, Timothy Verstraeten, Fabian Perez-Sanjines, Niel Hens, Philippe Lemey, and Ann Nowé. 2021. Deep Reinforcement Learning for Large-Scale Epidemic Control. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke (Eds.). Springer International Publishing, Cham, 155–170.

[33] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *ICML*, Vol. 80. PMLR, Stockholm, Sweden, 3150–3158.

[34] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. 2020. *Balancing Between Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning*. Vol. 12084 LNAI. Springer International Publishing, Cham. 155–167 pages. https://doi.org/10.1007/978-3-030-47426-3_13 arXiv:2106.13386

[35] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the applicability of ML fairness notions. , 32 pages. arXiv:2006.16745

[36] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[37] Tom M. Mitchell. 1997. *Machine Learning* (1 ed.). McGraw-Hill. 432 pages.

[38] Fatemeh Nargesian, Abolfazl Asudeh, and HV Jagadish. 2021. Tailoring data source distributions for fairness-aware data integration. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2519–2532.

[39] Javier Ortiz Laguna, Angel García Olaya, and Daniel Borrajo. 2011. A Dynamic Sliding Window Approach for Activity Recognition. In *User Modeling, Adaption and Personalization*, Joseph A. Konstan, Ricardo Conejo, José L. Marzo, and Nuria Oliver (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 219–230.

[40] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. 2020. Achieving fairness in stochastic multi-armed bandit problem. *arXiv* (2020). https://doi.org/10.1609/aaai.v34i04.5986 arXiv:1907.10516

[41] Pascale Petit. 2007. The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector. *Labour Economics* 14, 3 (2007), 371–391. https://doi.org/10.1016/j.labeco.2006.01.006

[42] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. 2022. Pareto Conditioned Networks. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1110–1118.

[43] 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. 2022. Exploring the pareto front of multi-objective covid-19 mitigation policies using reinforcement learning. *arXiv preprint arXiv:2204.05027* (2022), 25.

[44] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2021. Guaranteeing the Learning of Ethical Behaviour through Multi-Objective Reinforcement Learning. *ALA* (2021), 9.

[45] Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson. 2019. The Diverse Cohort Selection Problem. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 2 (2019),

601–609. arXiv:1709.03441

[46] Candice Schumann, Jeffrey S. Foster, Nicholas Mattei, and John P. Dickerson. 2020. We need fairness and explainability in algorithmic hiring. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 2020-May, Aamas (2020), 1716–1720.

[47] Umer Siddique, Paul Weng, and Matthieu Zimmer. 2020. Learning fair policies in multiobjective (Deep) reinforcement learning with Average and Discounted Rewards. *ICML* 119 (13–18 Jul 2020), 8864–8874. https://proceedings.mlr.press/v119/siddique20a.html

[48] Dennis Soemers, Ann Nowé, Tim Brys, Kurt Driessens, and Mark Winands. 2018. Adapting to Concept Drift in Credit Card Transaction Data Streams Using Contextual Bandits and Decision Trees. *AAAI* 32, 1 (2018), 7831–7836.

[49] STATBEL. 2023. Employment and unemployment. https://statbel.fgov.be/en/themes/work-training/labour-market/employment-and-unemployment#figures

[50] Richard S. Sutton, Andrew G. Barto, and et al. 2018. *Reinforcement Learning : An Introduction*. MIT Press. 526 pages.

[51] Hannah Van Borm and Stijn Baert. 2022. Diving in the Minds of Recruiters: What Triggers Gender Stereotypes in Hiring? *SSRN Electronic Journal* 15261 (2022). https://doi.org/10.2139/ssrn.4114837

[52] Jiayin Wang, Weizhi Ma, Jiayu Li, Hongyu Lu, Min Zhang, Biao Li, Yiqun Liu, Peng Jiang, and Shaoping Ma. 2022. 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*. 1868–1877.

[53] Paul Weng. 2019. Fairness in reinforcement learning. *CoRR* abs/1907.10323 (2019), 5. arXiv:1907.10323

[54] D Randall Wilson and Tony R Martinez. 1997. Improved heterogeneous distance functions. *Journal of artificial intelligence research* 6 (1997), 1–34.

[55] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *ICML (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 325–333.

[56] Luisa M Zintgraf, Edgar A Lopez-Rojas, Diederik M Roijers, and Ann Nowé. 2017. MultiMAuS: a multi-modal authentication simulator for fraud detection research. In *29th European Modeling and Simulation Symp.(EMSS 2017)*. Curran Associates, Inc., 360–370.

[57] Eva Zschirnt and Didier Ruedin. 2016. Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies* 42, 7 (2016), 1115–1134. https://doi.org/10.1080/1369183X.2015.1133279