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) [6]. 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, age 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 [3]. 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 [5]. 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

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

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

$$P = \frac{1}{K}\sum_{i=1}^{k}\frac{1}{3}|\{f^i \in \{d, e, x\} : f^i \neq 0\}| \tag{2}$$

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:

$$G_t = \frac{K}{N}\sum_{f_t \in p_t}(\hat{f}_{t+1} - f_t) \tag{3}$$

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

$$r_{t,hire} = G_t - \epsilon + \mathcal{N}(0, 0.01) \tag{4}$$

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:

$$r_{t,reject} = -r_{t,hire} \tag{5}$$

Transition function. We define the transition function $p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ 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 [4]. 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 fMDP, we implement the feedback signal f_t as the correct action $\hat{a}_t$ based on the goodness score:

$$f_t = \hat{a}_t \tag{6}$$

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 [8], 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

$$r_t = \begin{cases} +1 & \text{if genuine authentication,} \\ -1 & \text{if fraudulent authentication,} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

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 [8]. 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) [2] 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.

D POLICY VISUALISATION

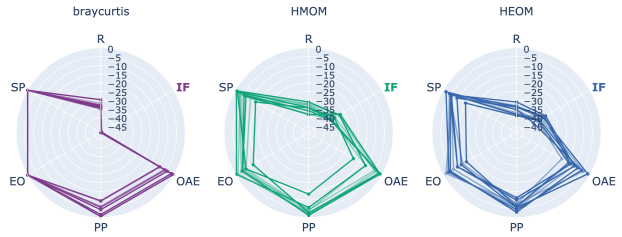
To easily compare the range of possible policy trade-offs across multiple objectives, all objectives are normalised and/or shifted, such that the maximum any objective can reach is 0. As the number of policies is large (more than 30 per seed), we opt to highlight a representative subset of 5-10 policies for each of the figures. This subset contains the policy with the highest performance reward (R), along with policies which differ most from each other and the un-highlighted policies across all policies. Note that we do plot all policies in low opacity. As such, more shaded regions indicate a larger number of policies which obtain similar trade-offs.

E ADDITIONAL JOB HIRING RESULTS

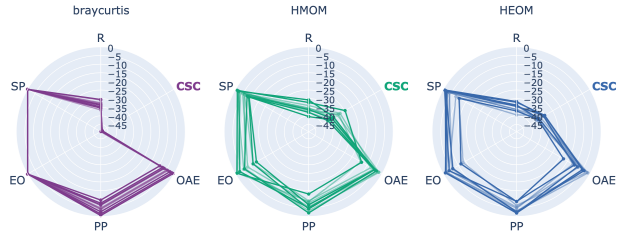
E.1 Individual fairness under different distance metrics

As individual fairness notions are impacted by the chosen distance metric, we explore their impact on the learned policies. Figure 1 shows a representative set of policies when optimising for an individual fairness notion under different distance metrics. Note how for the Bray-Curtis distance metric, both individual fairness notions are difficult to optimise to as high value as the heterogeneous distance metrics HEOM and HMOM [7]. Consequently, choosing the appropriate distance metric is context dependent, and must be decided a priori by stakeholders [1]. For our experiments, we have chosen to exclude the Bray-Curtis distance metric, as no policies could be found with any variation in the

trade-offs for both individual fairness notions. Therefore, this distance metric is not informative on the fairness in our experiments. Instead, we opted for HMOM, but note that HEOM could have been an equally suitable choice.



(a) Individual Fairness (IF)



(b) Consistency Score Complement (CSC)

Figure 1: Representative set of learned hiring policies when optimising for an individual fairness notion.

E.2 Job hiring with reward bias

Figure 2 shows the additional combinations of optimising the reward, a group fairness notion as well as an individual fairness notion under different reward configurations. In general, group fairness is easier to maximise. Note how the only exceptions are policies which prioritise the reward. The default configuration of Figure 2c in particular has found a policy which improves the reward, at the cost of both group and individual fairness.

F ADDITIONAL FRAUD DETECTION RESULTS

F.1 Individual fairness under different distance metrics

We explored the impact of different distance metrics in the fraud detection setting as well. We observe a similar effect to the job hiring setting, where the Bray-Curtis distance metric is very difficult to increase. For the same reasons, we have opted for the HMOM distance metric instead for the rest of the experiments.

F.2 Multi-objective policies under different history sizes

Figure 4 shows the learned policies when optimising the reward and an individual fairness notion, under different history sizes. We note a similar effect on the learned policies when optimising for the reward, a group fairness notion and an individual fairness notion, where the largest differences are across the reward R, EO and OAE. Note that the window size also impacts how much CSC can be maximised.

REFERENCES

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

[2] 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.

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

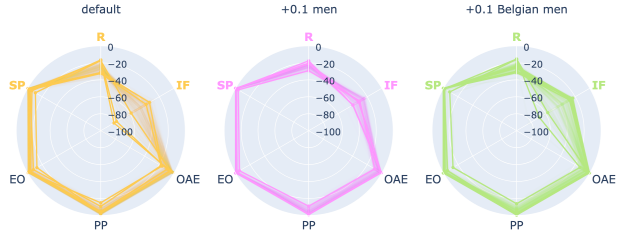
[4] STATBEL. 2023. Transitions on the labour market. https://statbel.fgov.be/en/themes/work-training/labour-market/transitions-labour-market#figures

[5] STATBEL. 2023. Volwasseneneducatie. https://statbel.fgov.be/nl/themas/werk-opleiding/opleidingen-en-onderwijs/volwasseneneducatie#figures

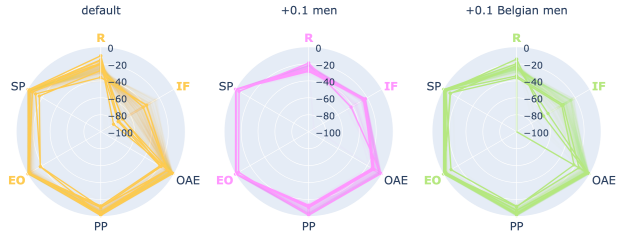
[6] Richard S. Sutton, Andrew G. Barto, and et al. 2018. *Reinforcement Learning : An Introduction.* MIT Press. 526 pages.

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

[8] 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.



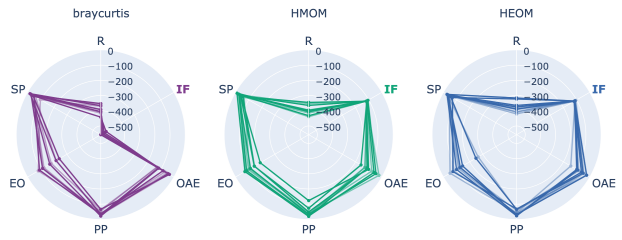
(a) Reward (R), Statistical Parity (SP) and Individual Fairness (IF)



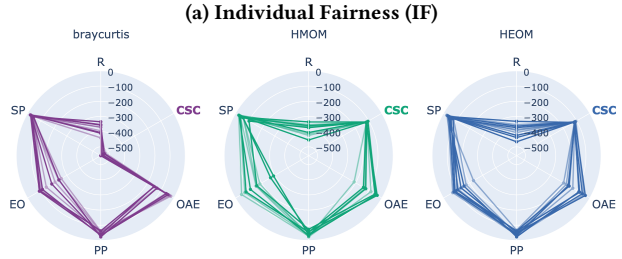
(b) Reward (R), Equal Opportunity (EO) and Individual Fairness (IF)

(c) Reward (R), Statistical Parity (SP) and Consistency Score Complement (CSC)

Figure 2: Representative set of learned hiring policies when optimising the reward, a group fairness notion and an individual fairness notion. Showing results for different reward configurations.

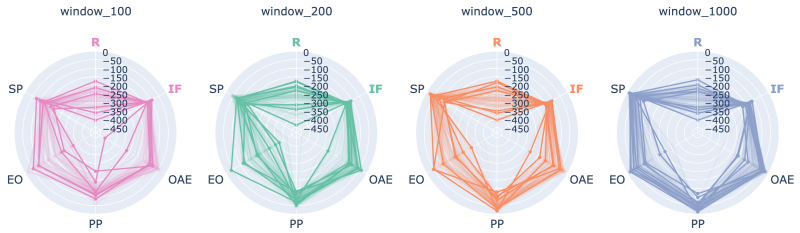


(a) Individual Fairness (IF)

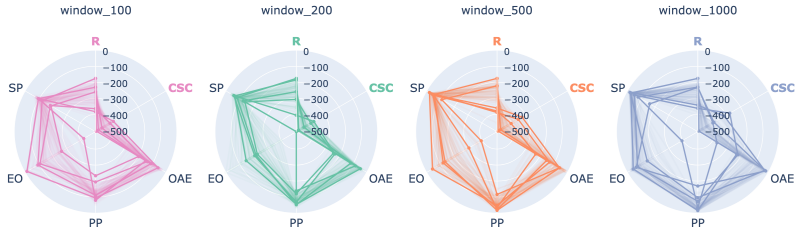


(b) Consistency Score Complement (CSC)

Figure 3: Representative set of learned fraud detection policies when optimising for an individual fairness notion.



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



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

Figure 4: Representative set of learned fraud detection policies when optimising the reward and an individual fairness notion. Showing results for histories with different sliding window sizes.