International Core Journal of Social Sciences
ISSN: 3080-8774 (Print), ISSN: 3080-8782 (Online) | Vol. 2, No. 2, 2026

FairPath: Fair and Explainable Talent Promotion Prediction Using Counterfactual Explanations and Fairness Constraints

Elisa Sharma, Avery Hall, Hiro Price, Miles Moore *

Graduate School of Management, University of California, Davis, USA

*** Corresponding author:** Miles Moore (Email: miles1976@ucdavis.edu)

Abstract: Machine-learning models that support promotion decisions are trained on historical HR records and can inherit the unequal treatment those records contain. Recent work such as FairPromote combines adversarial debiasing with SHAP attributions and evaluates outcome-level group fairness, but it does not tell a non-promoted employee what would change the decision, nor whether the effort required to obtain a promotion differs across groups. We present FairPath, which couples fairness constraints with counterfactual explanations. FairPath builds on the FairPromote adversarial encoder–predictor, replaces its loss-maximizing adversary with a confusion objective, and adds a counterfactual logit-pairing constraint that makes promotion scores invariant to flipping the sensitive attribute. For explanation we introduce TCAR, a model-agnostic generator of temporally consistent actionable recourse that searches a lattice of effort improvements and elapsed years exactly, advancing age, tenure and time in role jointly. TCAR also supports a counterfactual audit based on the attribute-flip rate (AFR) and the recourse-burden gap (RBG) between groups. On the IBM HR dataset and a 54,808-employee promotion dataset, under a controlled historical-label-bias protocol with clean test labels, FairPath reduced AFR from 10.2% (MLP) and 14.8% (FairPromote) to 0.1% on IBM-HR and to 0.0% on WNS, and lowered the IBM-HR recourse-burden gap of the MLP from 0.78 to 0.15 cost units without a significant loss of AUC. TCAR returned valid and feasible counterfactuals for every query, whereas none of the counterfactuals produced by DiCE or a gradient-based generator was both valid and feasible.

Keywords: Algorithmic Fairness; Algorithmic Recourse; Counterfactual Explanations; Explainable AI; Human Resource Analytics; Talent Promotion Prediction.

1. Introduction

Promotion decisions shape careers, pay trajectories and the composition of leadership. Organizations increasingly support these decisions with data-driven models that score employees on performance, tenure and engagement [1]. The same models raise well-documented concerns: historical HR records reflect past managerial judgments, and when those judgments treated groups differently, a model trained on them can reproduce the disparity at scale [2]–[4]. Sociological evidence shows that equal performance evaluations do not always translate into equal career rewards for women and minorities [3], which makes historical promotion labels a plausible carrier of bias.

FairPromote [5] is a recent framework that addresses this problem for promotion prediction. It learns employee representations against an adversarial discriminator that tries to recover the sensitive attribute, combines a neural head with gradient boosting, and audits the result with SHAP attributions; its evaluation reports demographic parity and equalized odds. The framework shows that accuracy, outcome fairness and transparency can be pursued in one pipeline. Two questions nevertheless remain open, and both concern the employees who are not promoted.

The first question is what an employee could do to be promoted. Attribution methods explain why a score is low, but they do not specify a feasible change. Counterfactual explanations and algorithmic recourse answer this question by searching for a minimal change of the input that flips the decision [6]–[8]. In a career setting, recourse has a temporal structure that generic counterfactual generators ignore: waiting t years increases age, total experience, tenure and years in the current role together, and nobody becomes younger. Treating these quantities as independent features yields recommendations such as “gain three more years of tenure” while age stays fixed, which no employee can follow.

The second question is whether the path to promotion is equally demanding for different groups. Outcome metrics such as demographic parity compare decision rates; they do not reveal that a model may require a woman to wait several more years, or to improve more performance indicators, than a comparable man. Counterfactual fairness [9] asks whether a decision would change had the sensitive attribute been different, and work on the fairness of recourse compares the cost of recourse across groups [10], [11]. These notions have not yet been brought to promotion prediction, where time-dependent recourse makes burden disparities particularly consequential.

This paper presents FairPath, a fair and explainable promotion predictor that uses counterfactual reasoning both as a fairness constraint and as an explanation. Our contributions are as follows.

- We extend the adversarial encoder–predictor of FairPromote with a confusion-based adversary, which stabilizes training, and with a counterfactual logit-pairing constraint that removes the direct dependence of the promotion score on the sensitive attribute.
- We propose TCAR, a model-agnostic generator of temporally consistent actionable recourse. TCAR searches an action lattice of effort improvements and elapsed years exactly, advances all time-indexed features jointly, and respects immutability, monotonicity and a

retirement bound. For each employee it returns an effort path, a waiting path and their cost-optimal combination.

- We define a counterfactual audit for promotion models, consisting of the attribute-flip rate and the recourse-burden gap, and show that it exposes disparities that outcome metrics miss.
- We evaluate FairPath on two public HR datasets against nine baselines, including a re-implementation of FairPromote, under a controlled historical-label-bias protocol with clean test labels, and report ablations, an analysis of an explicit recourse regularizer, and a comparison of explanation quality against DiCE and a gradient-based generator.

2. Related Work

2.1. Fairness in Machine Learning and HR Analytics

Fairness interventions are commonly grouped into pre-processing, in-processing and post-processing methods [12]. Reweighing assigns instance weights that make the sensitive attribute and the label independent in the training data [13]. Post-processing selects group-specific thresholds that equalize error rates [14]. The reductions approach turns fairness-constrained classification into a sequence of cost-sensitive problems solved by exponentiated gradient [15]. Adversarial debiasing trains a predictor jointly with an adversary that tries to infer the sensitive attribute from the model's internal representation [16]. In HR, systematic reviews document discrimination risks in recruitment and development decisions [2], and FairPromote combined adversarial representation learning with SHAP-based auditing for promotion prediction.

Individual-level notions complement group metrics. Counterfactual fairness [9] requires a decision to be unchanged in a counterfactual world in which the sensitive attribute differs. Garg *et al.* [17] operationalized a simple form of this idea for text classifiers through counterfactual logit pairing, which penalizes differences between the logits of an input and its attribute-swapped twin. FairPath adapts logit pairing to tabular promotion data and combines it with representation-level debiasing.

2.2. Counterfactual Explanations and Algorithmic Recourse

Wachter *et al.* [6] defined counterfactual explanations as the closest input that obtains a desired prediction. Ustun *et al.* [7] introduced actionable recourse for linear classifiers and distinguished mutable from immutable features. DiCE [18] generates diverse counterfactuals for arbitrary models, FACE [19] favors counterfactuals connected to dense regions of the data, and Mahajan *et al.* [20] preserved known causal constraints between features. Karimi *et al.* [21] argued that recourse should be framed as interventions whose consequences propagate through causal relations and surveyed the field in [8]. Ross *et al.* [22] trained models that guarantee the existence of recourse. Our temporal coupling is a domain-specific causal constraint: elapsed time is a single intervention that advances several features at once. SHAP [24] provides additive attributions grounded in Shapley values and is widely used to audit HR models; attributions describe why a score is low but not which change would raise it, which motivates pairing them with recourse. Sun et al. offer an applied example of combining nonlinear predictive modeling with SHAP-based feature interpretation for transparent risk-sensitive decisions [31].

2.3. Fairness of Recourse

Gupta *et al.* [10] observed that classifiers with similar outcome rates can impose different recourse costs on different groups and regularized the gap in distance to the decision boundary for linear models. Von Kügelgen *et al.* [11] analyzed the fairness of causal recourse at group and individual levels, and CERTIFAI [23] measured fairness through the burden of counterfactuals. We use recourse burden as an audit metric for promotion models and compute it with temporally consistent counterfactuals.

3. Problem Formulation

We observe employees (x_i, s_i, y_i), $i = 1, \ldots, N$, where x is the feature vector, $s \in \{0, 1\}$ is a binary sensitive attribute ($s = 1$ for women in our experiments) that is included among the inputs, and $y \in \{0, 1\}$ is the promotion outcome. A model produces a logit $g(x)$ and a promotion probability $p(x) = \sigma(g(x))$; the decision is $\hat{y} = 1[p(x) \geq \tau]$, with τ tuned on validation data. Features play one of three roles: *effort* features E are employee-actionable, can only increase and are bounded by their observed maximum; *time* features T advance together when time elapses; all other features, including s, are *immutable*.

Outcome fairness is measured with the demographic parity difference and the equalized odds difference,

$$\mathrm{DPD} = |P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1)|, \quad (1)$$

$$\mathrm{EOD} = \tfrac{1}{2}\,(|\mathrm{TPR}_0 - \mathrm{TPR}_1| + |\mathrm{FPR}_0 - \mathrm{FPR}_1|), \quad (2)$$

where TPR_s and FPR_s are the true and false positive rates of group s. Let $x^{s^* \leftarrow 1 - {}^* s}$ denote x with the sensitive input flipped and every other feature unchanged. The attribute-flip rate

$$\mathrm{AFR} = (1/n)\ \Sigma_i\ 1[\hat{y}(x_i) \neq \hat{y}(x_i^{s^* \leftarrow 1 - {}^* s})] \quad (3)$$

is the share of decisions that would change with the sensitive attribute alone; for policies with group-specific thresholds, the flipped employee is also judged by the other group's threshold.

An action $a = (k, t)$ applies k_j improvement steps of size δ_j to each effort feature j and lets t years elapse. The transition $x \oplus a$ adds $k_j \delta_j$ to effort feature j and t to every time feature, subject to the observed bounds and the retirement cap age + $t \leq A_{\max}$. The cost of an action is

$$c(a) = \Sigma_{j^* \in^* E}\ k_j \delta_j / R_j + c_t\, t, \quad (4)$$

where R_j is the observed range of feature j and $c_t = 0.1$, so that ten years of waiting cost as much as moving one feature across its whole range. The recourse burden of a non-promoted employee and the recourse-burden gap are

$$B(x) = \min\ \{\ c(a) : a \in A(x), \hat{y}(x \oplus a) = 1\ \}, \quad (5)$$

$$\mathrm{RBG} = |\mathrm{E}[B \mid \hat{y} = 0, s = 0] - \mathrm{E}[B \mid \hat{y} = 0, s = 1]|, \quad (6)$$

where $A(x)$ is the set of feasible actions; when no feasible action reaches a positive decision, $B(x)$ is set to the cost of the largest action in the lattice, so infeasibility counts as burden.

4. The FairPath Framework

4.1. Overview

Fig. 1 summarizes FairPath. During training, fairness is enforced at two levels: an adversary removes sensitive information from the learned representation, and a

counterfactual logit-pairing term removes the direct effect of the sensitive input on the score. After training, TCAR generates temporally consistent recourse for non-promoted employees and computes the counterfactual audit, alongside SHAP attributions.

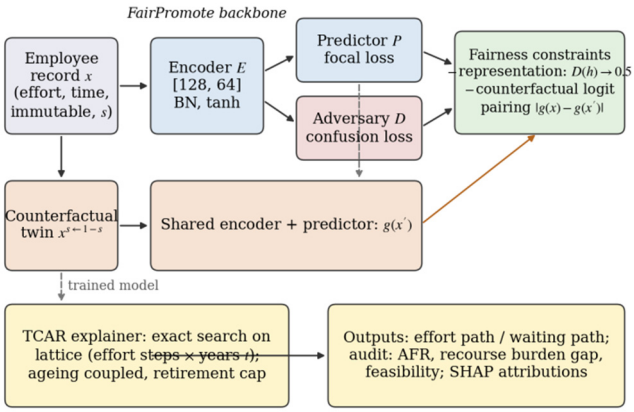


Fig 1. Overview of FairPath. The encoder–predictor–adversary backbone follows FairPromote; FairPath replaces the adversarial objective with a confusion loss and adds counterfactual logit pairing. TCAR explains and audits any trained model through black-box queries.

4.2. Adversarial Representation Learning

We build on the FairPromote architecture [5] and keep its components and hyperparameters: an encoder

$$h = \tanh(W_2\,\mathrm{Dropout}(\mathrm{BN}(\mathrm{ReLU}(W_1 x + b_1))) + b_2) \quad (7)$$

with hidden sizes [128, 64] and dropout 0.3, a predictor head with layers [64, 32], and an adversary D with one hidden layer of 32 units that predicts s from h. The predictor is trained with the focal loss [25]

$$L_{\mathrm{foc}} = -(1/N)\,\Sigma_i\,[(1-p_i)^{\gamma} y_i \log p_i + p_i^{\gamma}(1-y_i)\log(1-p_i)] \quad (8)$$

with $\gamma = 2$. FairPromote trains the encoder and predictor to minimize $L_{\mathrm{foc}} - \alpha L_{\mathrm{adv}}$, where L_{adv} is the adversary's cross-entropy. Maximizing the adversary's loss can be satisfied by representations from which s is predicted confidently but wrongly, which still encode the attribute [26]. FairPath therefore uses the confusion objective of [26], which drives the frozen adversary toward maximal uncertainty,

$$L_{\mathrm{conf}} = -(1/N)\,\Sigma_i\,\tfrac{1}{2}\,[\log D(h_i) + \log(1 - D(h_i))] - \log 2. \quad (9)$$

The adversary itself is updated to minimize L_{adv} on detached representations, alternating with the encoder–predictor step.

4.3. Counterfactual Logit Pairing

Representation-level debiasing removes information that can be decoded from h, but it does not stop the predictor from responding to the sensitive input directly. Following counterfactual logit pairing [17], we penalize the difference between the logits of each employee and of the attribute-flipped twin,

$$L_{\mathrm{cf}} = (1/N)\,\Sigma_i\,|g(x_i) - g(x_i^{s^*\leftarrow 1-s^*})|. \quad (10)$$

We pair logits rather than probabilities because probabilities saturate for clear non-promotions, which are precisely the employees who need recourse. The training problem is

$$\min_{\theta^{*E},\theta^{*P}} L_{\mathrm{foc}} + \alpha L_{\mathrm{conf}} + \lambda_{\mathrm{cf}} L_{\mathrm{cf}}, \quad \min_{\theta^{*D}} L_{\mathrm{adv}}, \quad (11)$$

solved by alternating Adam updates with $\alpha = 1.5$, the value used by FairPromote, and $\lambda_{\mathrm{cf}} = 1$.

4.4. Temporally Consistent Actionable Recourse (TCAR)

Given a trained model and an employee with $\hat{y}(x) = 0$, TCAR enumerates the action lattice of all step vectors k with $0 \le k_j \le K_j(x)$ and all horizons $0 \le t \le \min(T_{\max}, A_{\max} - \mathrm{age})$, where $K_j(x)$ keeps feature j inside its observed range and $T_{\max} = 10$ years. Each lattice point is mapped to a feature vector in which every time feature advances by t; all points are scored by the model in one batch, and the valid point of minimum cost is returned. The search is exact on the lattice and needs only query access to the model, so it applies equally to tree ensembles, neural networks and their mixtures. The lattice is small for HR data: at most 3,080 points per employee for IBM-HR and 25,740 for WNS. Besides the cost-optimal combined action, TCAR reports two restricted counterfactuals that suit career conversations: an *effort path* ($t = 0$) and a *waiting path* (no effort, smallest t). The same search yields $B(x)$, the RBG and the share of employees with feasible recourse.

4.5. Attribution and Audit

We complement TCAR with SHAP attributions computed on the original features by permutation sampling [24]. For hybrid predictors such as FairPromote, the attributions of the boosting and neural components are combined with the ensemble weights, mirroring FairPromote's interpretation module. Attributions explain why a score is low; TCAR states what would change it and at what cost.

4.6. Variant: Explicit Recourse Regularization

A natural alternative is to regularize recourse disparities directly during training. We implemented a differentiable first-order proxy of the recourse cost,

$$r(x) = [t - g(x)]_+ / \max_k\,[\nabla g(x)^{\mathrm{T}} a_k]_+, \quad (12)$$

where a_k are unit-cost action directions (one per effort feature plus the coupled time direction) and t is the logit of the marginal promoted employee at the training promotion rate, which makes r invariant to affine rescaling of g. Adding $\lambda_{\mathrm{rec}}|\bar{R}_0 - \bar{R}_1|$, where $\bar{R}_g$ is the mean proxy over non-promoted employees of group g in group-balanced mini-batches, yields the variant analyzed in Section VI-E; the max is smoothed with a log-sum-exp and the gradient of g is differentiated by double back-propagation.

5. Experimental Setup

5.1. Datasets

IBM-HR. The IBM HR Analytics Employee Attrition & Performance dataset [27] contains 1,470 fictional employee records created by IBM data scientists. It has no explicit promotion event, so, following the JobLevel-based labeling used by FairPromote, we define promotion to a senior grade as $y = 1[\mathrm{JobLevel} \ge 3]$ (26.7% positives). MonthlyIncome and JobRole are removed because they determine the grade. Effort features are JobInvolvement, PerformanceRating, TrainingTimesLastYear and Education; time features are Age, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager and YearsSinceLastPromotion.

WNS. The HR Analytics Employee Promotion data were released for the WNS Analytics Wizard 2018 hackathon and describe 54,808 anonymized employees of a large multinational company with an explicit promotion label (8.5% positives). We use the publicly distributed training file and

verified its size and class balance against the competition description. Effort features are the previous-year rating, KPIs met (above 80%), awards won, average training score (steps of 5 points), number of trainings and education level; time features are age and length of service.

Both datasets show small recorded gender gaps that slightly favor women (Table 1). Gender is not predictable from the other IBM-HR features (proxy AUC 0.47), but it is moderately predictable in WNS (0.73), mainly through department.

Table 1. Dataset summary

	IBM-HR	WNS
Employees	1,470	54,808
Raw / encoded features	27 / 38	12 / 55
Promotion label	JobLevel ≥ 3	is_promoted
Positive rate (all)	26.7%	8.5%
Positive rate (women / men)	28.7% / 25.4%	9.0% / 8.3%
Share of women	40.0%	29.8%
Gender proxy AUC	0.47	0.73
Effort / time features	4 / 6	6 / 2
Max. lattice size per employee	3,080	25,740

5.2. Historical Label-Bias Protocol

Because the recorded labels exhibit little gender disparity, we evaluate bias mitigation with a controlled label-bias protocol [28]: in the training and validation splits, each promoted woman is relabeled as not promoted with probability β, simulating deserved promotions withheld by past decision makers. Models are trained and tuned on the biased labels, and all metrics are computed against the clean test labels. We report $\beta = 0$ (original labels) and $\beta = 0.4$, and sweep $\beta \in \{0, 0.2, 0.4, 0.6\}$.

5.3. Baselines

We compare nine methods: logistic regression (LR); XGBoost [29] with 100 trees of depth 6; an MLP with the FairPromote architecture trained with cross-entropy; Unaware-XGB, which drops the gender input; Reweighing [13] followed by XGBoost; EO-Post, which selects group-specific thresholds on XGBoost scores that equalize odds on validation data [14]; ExpGrad-EO, the reductions approach with equalized-odds constraints [15] as implemented in Fairlearn [30]; EqRec-LR, a linear model with the equalized-recourse penalty of Gupta *et al.* [10] in our action geometry; and FairPromote [5], re-implemented with its published configuration (focal loss with $\gamma = 2$, $\alpha = 1.5$, Adam with learning rates 10^{-3} and $5 \cdot 10^{-4}$, early stopping with patience 15, XGBoost on the learned representation, and a validation-selected ensemble weight). Our implementation freezes the adversary's parameters during the encoder–predictor step so that the reversed adversarial gradient reaches the encoder.

5.4. Implementation and Metrics

Each run uses a stratified 70/15/15 split by label and gender. We repeat every configuration with 10 seeds on IBM-HR and 3 seeds on WNS; secondary studies use fewer seeds, as stated with each result. Numeric features are standardized and categorical features one-hot encoded. Neural models are implemented in JAX with batch sizes of 32 (IBM-HR) and 256 (WNS) and early stopping on the validation objective. The decision threshold of every model is chosen on validation data to maximize F1. Recourse is audited for all predicted non-promoted test employees on IBM-HR and for a gender-balanced pool of 300 test employees per seed on WNS. We report accuracy-related metrics (F1, AUC) on clean labels; DPD and EOD as defined above, following FairPromote's evaluation protocol; AFR; the mean recourse burden per group; RBG; and the share of non-promoted employees with feasible recourse. All experiments ran on a single CPU core.

6. Results

6.1. Prediction and Fairness under Historical Bias

Tables 2 and 3 report results under historical label bias ($\beta = 0.4$). Every model that receives the gender input inherits part of the bias: flipping gender alone changes 13.8% of LR, 10.8% of XGBoost and 10.2% of MLP decisions on IBM-HR, and 2.7% of XGBoost decisions on WNS. FairPath reduces AFR to 0.1% and 0.0%, below every baseline that uses the gender input on both datasets (paired t-tests over seeds on IBM-HR: all $p < 0.001$). The constraints cost little predictive performance: FairPath reaches an AUC of 0.869 on IBM-HR and 0.902 on WNS, against 0.869 and 0.898 for the MLP with the same architecture (IBM-HR: difference not significant, $p = 0.891$; WNS: significantly higher, $p = 0.014$). Gradient-boosted trees remain the most accurate family on IBM-HR (best AUC 0.896, Unaware-XGB).

Table 2. Results on IBM-HR under historical label bias ($\beta = 0.4$)

Method	AUC ↑	F1 ↑	DPD ↓	EOD ↓	AFR ↓	RBG ↓
LR	86.3±2.3	65.5±3.9	12.3±6.8	15.4±7.4	13.8±4.8	1.14±0.26
XGBoost	88.8±2.1	59.8±6.0	9.6±6.2	16.2±7.4	10.8±3.8	0.44±0.31
MLP	86.9±2.2	65.2±4.4	7.8±3.7	11.4±4.8	10.2±3.4	0.78±0.69
Unaware-XGB	**89.6±1.5**	64.5±4.8	**3.2±2.1**	**5.7±3.5**	**0.0±0.0**	0.31±0.19
Reweighing	89.1±2.0	60.3±5.2	7.2±4.0	12.7±6.6	8.4±3.5	0.29±0.24
EO-Post	88.8±2.1	61.2±3.6	7.0±5.6	9.5±7.1	9.0±4.0	0.36±0.41
ExpGrad-EO	89.4±1.8	62.5±8.8	7.1±4.2	11.8±6.0	8.3±2.1	0.41±0.36
EqRec-LR	87.5±2.5	**67.7±3.6**	5.0±4.0	7.2±6.2	5.2±2.0	**0.09±0.04**
FairPromote	85.7±3.0	63.9±4.8	9.7±5.9	13.2±7.7	14.8±4.5	0.28±0.22
FairPath (ours)	86.9±2.5	65.9±2.9	4.3±2.0	**5.7±4.0**	0.1±0.3	0.15±0.18

Mean ± std over 10 seeds on clean test labels. AUC, F1, DPD, EOD and AFR ×100; RBG in cost units. Best value in bold. Unaware-XGB has zero AFR by construction

Outcome metrics alone give an incomplete picture. On IBM-HR, non-promoted women face a mean recourse burden of 1.60 cost units under the MLP, against 0.82 for men; with ten years of waiting priced at one unit, the difference corresponds to about 7.8 additional years. LR shows the same asymmetry (2.22 versus 1.08). FairPath lowers the RBG from 0.78 (MLP) to 0.15 (significant, $p = 0.029$), and its group burdens are close (1.13 for men, 1.04 for women).

The fairness baselines expose complementary trade-offs. EqRec-LR, which penalizes recourse disparities explicitly, attains the lowest RBG on IBM-HR (0.09; the difference to FairPath is not significant, $p = 0.408$), but keeps an AFR of

5.2% because its linear score still uses gender. Unaware-XGB performs well on IBM-HR, where gender has no proxies (DPD 3.2%). On WNS, where department carries gender information, its DPD and EOD are 0.1% and 1.3%, against 0.4% and 1.5% for FairPath. Reweighing, EO-Post and ExpGrad-EO reduce some outcome gaps on IBM-HR but leave AFR between 8.3% and 9.0%, since their decision rules still depend on gender. The FairPromote re-implementation has the highest AFR on IBM-HR (14.8%), above the MLP without an adversary; Section VI-D relates this to its adversarial objective.

Table 3. Results on WNS under historical label bias (β = 0.4)

Method	AUC ↑	F1 ↑	DPD ↓	EOD ↓	AFR ↓	RBG ↓
LR	86.7±1.4	44.5±1.8	4.6±1.3	10.8±3.5	5.5±1.2	0.03±0.01
XGBoost	90.5±0.9	50.6±1.8	2.4±0.9	6.6±1.5	2.7±0.9	0.06±0.01
MLP	89.8±0.7	50.2±2.0	1.7±0.7	4.7±1.6	2.4±0.8	0.01±0.00
Unaware-XGB	**90.9±0.8**	**51.7±1.7**	**0.1±0.0**	**1.3±1.1**	**0.0±0.0**	0.02±0.03
Reweighing	**90.9±0.9**	51.4±1.6	1.0±1.0	3.7±2.2	1.6±0.9	0.03±0.02
EO-Post	90.5±0.9	48.7±5.1	1.5±0.4	3.7±2.4	1.9±0.5	0.08±0.07
ExpGrad-EO	90.6±1.0	51.0±2.8	0.6±0.4	1.9±0.5	0.8±0.5	0.03±0.03
EqRec-LR	87.3±1.4	46.0±1.0	0.8±0.6	3.0±2.8	1.8±0.6	**0.01±0.00**
FairPromote	89.9±0.5	50.8±1.9	1.1±0.5	3.8±2.3	1.5±0.0	0.02±0.01
FairPath (ours)	90.2±0.7	50.9±3.1	0.4±0.4	1.5±0.9	**0.0±0.0**	0.01±0.00

Mean ± std over 3 seeds on clean test labels. AUC, F1, DPD, EOD and AFR ×100; RBG in cost units. Best value in bold. Unaware-XGB has zero AFR by construction

6.2. Recourse Burden by Group

Fig. 2 shows the group-wise burdens behind the RBG. On IBM-HR, senior grades are reached mainly through experience, so recourse is dominated by waiting time and a score penalty against women translates into years. Tree models also leave many non-promoted employees without any feasible recourse inside the lattice (65.9% feasible for XGBoost versus 89.0% for FairPath), because their piecewise-constant scores do not respond to small changes; such employees carry the maximal burden. On WNS, effort actions such as raising the average training score are cheap and effective, so mean burdens lie between 0.20 and 0.26 cost units for all models and bias produces small absolute gaps (RBG at most 0.076). The same score bias therefore has very different consequences for employees depending on the action geometry, which outcome metrics cannot reveal.

6.3. Original Labels

Table 4 repeats the comparison on the original labels. With little gender disparity in the data, all models have small outcome gaps, and FairPath's AUC (0.879 on IBM-HR, 0.902 on WNS) is close to that of the MLP (0.882 and 0.904). FairPromote is the exception in counterfactual terms: its AFR is 9.6% on IBM-HR although the labels carry no injected bias, against 4.4% for the MLP and 0.1% for FairPath. On unbiased data, FairPath thus stays close to the MLP in accuracy while keeping its decisions independent of the gender input.

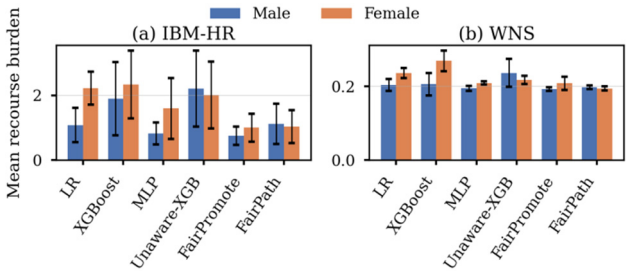


Fig 2. Mean recourse burden of non-promoted men and women under historical label bias (β = 0.4); error bars show the standard deviation over seeds. One cost unit equals ten years of waiting or one full-range improvement of an effort feature.

Table 4. Results on the original labels (β = 0); mean values

	IBM-HR				WNS			
Method	AUC	DPD	AFR	RBG	AUC	DPD	AFR	RBG
LR	89.1	4.2	0.5	0.20	87.3	0.7	0.3	0.01
XGBoost	89.4	4.0	0.5	0.30	91.1	0.4	0.5	0.01
MLP	88.2	4.7	4.4	0.19	90.4	0.7	1.6	0.01
Unaware-XGB	89.4	4.6	0.0	0.23	91.0	0.4	0.0	0.01
Reweighing	89.4	3.1	0.5	0.23	91.0	0.1	0.7	0.01
EO-Post	89.4	7.2	4.0	0.60	91.1	1.4	0.8	0.01
ExpGrad-EO	89.6	3.1	0.3	0.20	90.8	2.9	2.8	0.01
EqRec-LR	88.7	4.0	1.9	0.09	87.1	0.4	0.9	0.01
FairPromote	87.9	4.1	9.6	0.23	90.0	0.6	0.7	0.01
FairPath (ours)	87.9	3.5	0.1	0.10	90.2	0.9	0.0	0.01

AUC, DPD and AFR ×100; RBG in cost units.

6.4. Ablation Study

Table 5 removes or replaces one component at a time (β = 0.4; 5 seeds on IBM-HR and 1 on WNS). Without logit pairing, AFR rises to 9.2% on IBM-HR and 3.1% on WNS: an adversary on the representation does not stop the predictor from reading the gender input directly. The confusion objective alone is nevertheless less harmful than the sign-flipped objective of FairPromote, whose AFR is 15.8% on the same seeds. Removing the adversary leaves DPD and AFR on WNS unchanged (0.2% and 0.0%), so in our settings most of the counterfactual benefit comes from logit pairing. With the sign-flipped objective instead of the confusion loss, accuracy degrades sharply: the AUC on WNS falls from 0.893 to 0.674, and on IBM-HR it drops from 0.866 to 0.772. The complete FairPromote pipeline does not show this drop, plausibly

because its boosting head and validation-selected ensemble weight can down-weight a degraded neural branch. Finally, attaching FairPromote's boosting head to the FairPath encoder leaves AFR at 0.0% on both datasets.

Table 5. Ablation of FairPath (β = 0.4); mean values

	IBM-HR				WNS			
Variant	AUC	DPD	AFR	RBG	AUC	DPD	AFR	RBG
FairPath (full)	86.6	4.5	0.0	0.09	89.3	0.2	0.0	0.01
w/o logit pairing	86.7	4.4	9.2	0.69	88.6	3.3	3.1	0.04
w/o adversary	86.9	3.9	0.0	0.14	89.2	0.2	0.0	0.01
sign-flip adversary	77.2	1.5	0.4	0.26	67.4	0.4	0.0	0.07
with boosting head	86.6	5.3	0.0	0.08	89.5	0.2	0.0	0.01

AUC, DPD and AFR ×100; RBG in cost units.

6.5. Explicit Recourse Regularization

Fig. 3 adds the recourse regularizer of Section IV-F to FairPath (5 seeds on IBM-HR, 1 on WNS). On IBM-HR, the RBG with $\lambda_{rec} > 0$ ranges from 0.034 to 0.170, against 0.090 without the term, while the AUC moves from 0.866 to 0.679 at $\lambda_{rec} = 2$; on WNS the RBG stays between 0.002 and 0.014. The regularizer can push the RBG below that of FairPath (down to 0.034), although not monotonically in its weight and at the price of a lower AUC, so it offers a trade-off that must be tuned by organizations that prioritize recourse parity.

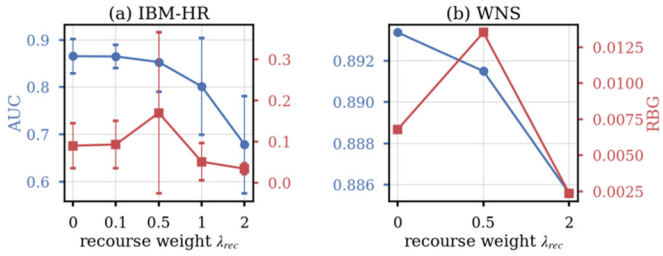


Fig 3. Effect of the explicit recourse-regularization weight λ_{rec} on AUC (left axis) and RBG (right axis) at β = 0.4. $\lambda_{rec} = 0$ is FairPath.

6.6. Strength of Historical Bias

Fig. 4 varies the label-bias strength β (5 seeds on IBM-HR). As β grows from 0 to 0.6, the AFR of XGBoost on IBM-HR rises from 0.2% to 18.0% and that of FairPromote from 10.7% to 20.3%, whereas FairPath stays at 0.0%. The RBG of XGBoost reaches 0.82 at β = 0.6, against 0.18 for FairPath. Unaware-XGB is immune to direct dependence by construction.

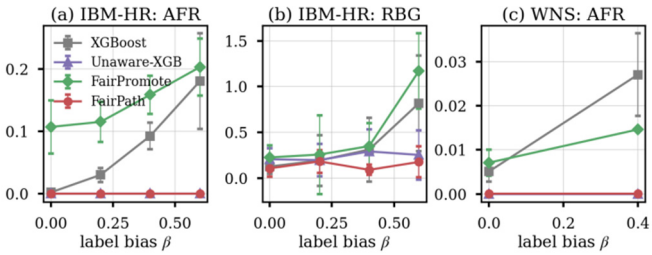


Fig 4. Attribute-flip rate and recourse-burden gap as functions of the label-bias strength β (mean ± std over seeds).

6.7. Explanation Quality

Table 6 compares counterfactual generators on the trained FairPath model. A counterfactual is counted as feasible when all time features advance by the same non-negative integer number of years, no effort or time feature decreases, ordinal features keep integer values inside the observed range, and immutable features are unchanged. TCAR satisfies these conditions by construction and finds valid recourse for every query on both datasets in 5–8 ms per employee. The Wachter-style gradient search always reaches the target but moves time features independently and produces fractional ratings and years, so none of its counterfactuals is feasible. DiCE, configured with the same actionable features, integer data types, permitted ranges and a budget of 5,000 sampled candidates, returned a valid counterfactual for 25% of the IBM-HR and 0% of the WNS queries, and the counterfactuals it did return changed age and tenure independently (0% temporally consistent on IBM-HR). Freezing age while tenure advances, as is common when age is declared immutable, yields recommendations of which 7% become invalid on IBM-HR once the employee is aged consistently; on WNS the cheapest recourse never involves waiting, so the two variants coincide.

Table 6. Counterfactual generators (% of queries; time in ms per employee)

Generator	Valid	Consist.	Monot.	Integr.	Feasible & valid	ms
IBM-HR (FairPath model, 60 non-promoted test employees)						
TCAR (ours)	100	100	100	100	100	5
TCAR, age frozen	97	0	100	100	0	5
Wachter-style	100	0	0	0	0	401
DiCE (random)	25	0	33	100	0	1448
WNS (FairPath model, 40 non-promoted test employees)						
TCAR (ours)	100	100	100	100	100	8
TCAR, age frozen	100	100	100	100	100	9
Wachter-style	100	0	5	0	0	415
DiCE (random)	0	–	–	–	0	1248

Consist., Monot. and Integr. are computed over the counterfactuals actually returned; – indicates that none was returned

6.8. Attributions and a Case Study

Table 7 reports the mean absolute SHAP attribution of the gender input under bias (β = 0.4, 200 test employees). Following FairPromote's interpretation module, the attributions of its boosting and neural components are combined with the ensemble weight. Gender ranks 4th among 27 features for the MLP and 4th for FairPromote on IBM-HR,

but 27th for FairPath, consistent with the counterfactual audit. Attributions and recourse answer different questions: for 45% of non-promoted IBM-HR employees and 60% of WNS employees, at least one of the three most negative attributions belongs to an immutable feature that no action can change, and in 100% of IBM-HR cases a feature changed by TCAR is among the three most negative actionable attributions. Wang et al. evaluate explanation faithfulness in a memory-augmented sequential recommendation agent, providing a useful complementary perspective on assessing the quality of user-facing decision explanations [32].

Table 7. Mean |SHAP| of the gender input (×100) and its rank

	IBM-HR		WNS	
Model	|φ|	Rank	|φ|	Rank
MLP	3.32	4/27	2.55	4/12
FairPromote	2.85	4/27	0.75	6/12
FairPath (ours)	0.01	27/27	0.00	12/12

As a case study, consider a non-promoted male employee from the IBM-HR test set with a predicted promotion probability of 0.44. TCAR offers two directly usable options: an effort path that raises job involvement from 3 to 4 and education level from 2 to 4 (cost 0.83), and a waiting path of 2 years during which age, tenure and time in role advance together. The cost-optimal plan combines 2 more years of service (cost 0.20). Because FairPath's score is invariant to the gender input, the same plan would be issued to an otherwise identical employee of the other gender.

7. Discussion and Limitations

Three limitations qualify the findings. First, the bias protocol injects label bias through a known mechanism; real historical bias can also operate through features, for example through biased ratings, and the IBM-HR records are fictional. The protocol is therefore a controlled test of whether a method recovers an unbiased decision rule, not a measurement of bias in any real organization. Second, recourse costs encode policy choices, such as the relative price of a training course and of a year of waiting; organizations would weight them differently, and TCAR accepts any additive cost. Third, temporal coupling is a deliberately simple causal model: it does not represent that effort can change future ratings or that a role change resets time in role. Richer structural models [21] and a validation of the recommendations with HR practitioners are natural next steps. Finally, the attribute-flip analysis treats gender as a direct input, as FairPromote does. Simply removing the input (Unaware-XGB) was competitive on both of our datasets, including WNS where gender is partly predictable from department; when gender is removed, the adversary remains the only safeguard against proxy information, and its benefit should be verified on data with stronger proxies.

8. Conclusion

FairPath shows that counterfactual reasoning serves two complementary roles in promotion prediction. As a training constraint, counterfactual logit pairing on top of a confusion-based adversarial encoder removed the direct dependence of decisions on gender and reduced recourse-burden disparities that outcome metrics did not reveal. As an explanation, temporally consistent recourse turned scores into feasible effort and waiting paths, whereas generic generators produced recommendations that employees could not follow. The audit metrics and the TCAR search are model-agnostic and can be applied to existing promotion models without retraining.

References

[1] Tambe, M., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15–42. https://doi.org/10.1177/0008125619867910.

[2] Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w.

[3] Castilla, E. J. (2008). Gender, race, and meritocracy in organizational careers. American Journal of Sociology, 113(6), 1479–1526. https://doi.org/10.1086/588738.

[4] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[5] Wang, Z., Yang, J. S., Shang, W., & Ding, J. (2026). FairPromote: Explainable and fairness-aware talent promotion prediction via adversarial debiasing and SHAP-based interpretation. IEEE Access, 14, 72891–72904. https://doi.org/10.1109/ACCESS.2026.3692578.

[6] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

[7] Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 10–19). https://doi.org/10.1145/3287560.3287566.

[8] Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2022). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. ACM Computing Surveys, 55(5), Article 95. https://doi.org/10.1145/3527848.

[9] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems, 30, 4066–4076.

[10] Gupta, V., Nokhiz, P., Roy, C. D., & Venkatasubramanian, S. (2019). Equalizing recourse across groups. arXiv:1909.03166.

[11] von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., & Schölkopf, B. (2022). On the fairness of causal algorithmic recourse. Proceedings of the AAAI Conference on Artificial Intelligence, 36(9), 9584–9594. https://doi.org/10.1609/aaai.v36i9.21192.

[12] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607.

[13] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8.

[14] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323.

[15] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In Proceedings of the 35th International Conference on Machine Learning (pp. 60–69).

[16] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 335–340). https://doi.org/10.1145/3278721.3278779.

[17] Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019). Counterfactual fairness in text classification through robustness. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 219–226). https://doi.org/10.1145/3306618.3317950.

[18] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 607–617). https://doi.org/10.1145/3351095.3372850.

[19] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[20] Mahajan, D., Tan, C., & Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv:1912.03277.

[21] Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (pp. 353–362). https://doi.org/10.1145/3442188.3445899.

[22] Ross, A., Lakkaraju, H., & Bastani, O. (2021). Learning models for actionable recourse. Advances in Neural Information Processing Systems, 34.

[23] Sharma, S., Henderson, J., & Ghosh, J. (2020). CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 166–172). https://doi.org/10.1145/3375627.3375812.

[24] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[25] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324.

[26] Teng, D. (2025). TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows. World Journal of Engineering Research, 3(6), 55–66.

[27] IBM. (2026, October 22). IBM HR analytics employee attrition & performance. https://github.com/IBM/employee-attrition-aif360.

[28] Wick, M., Panda, S., & Tristan, J.-B. (2019). Unlocking fairness: A trade-off revisited. Advances in Neural Information Processing Systems, 32.

[29] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

[30] Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023). Fairlearn: Assessing and improving fairness of AI systems. Journal of Machine Learning Research, 24(257), 1–8.

[31] Sun, T. (2024). Enhancing auto insurance risk evaluation with Transformer and SHAP. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3446179.

[32] Wang, G., Li, P., Feng, E., Shen, A., & Rhee, M. (2026). A dual-process memory-augmented large language model agent for explainable sequential recommendation. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3741894.