

Fairness-Aware and Interpretable Policy Learning

Nora Bearth, Michael Lechner, Jana Mareckova, Fabian Muny

ESIF Economics and AI + ML Meeting 2026
Ithaca, NY, United States

June 16, 2026

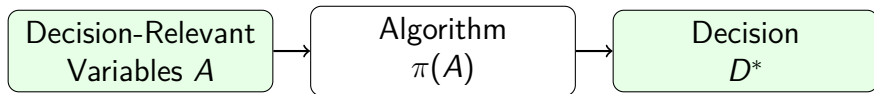
Motivation

- Algorithmic decision-making needs to consider fairness / non-discrimination (EU AI Act, AI Bill of Rights in the US)
- Successful deployment of algorithms often requires transparency (Bansal et al., 2021; Panigutti et al., 2022)
- **Goal:** Promote fairness and interpretability in algorithmic decision-making with respect to protected attributes S
- **Application:** Apply an interpretable decision algorithm promoting fair access to active labor market programs (ALMPs) in Switzerland

Protected Attributes

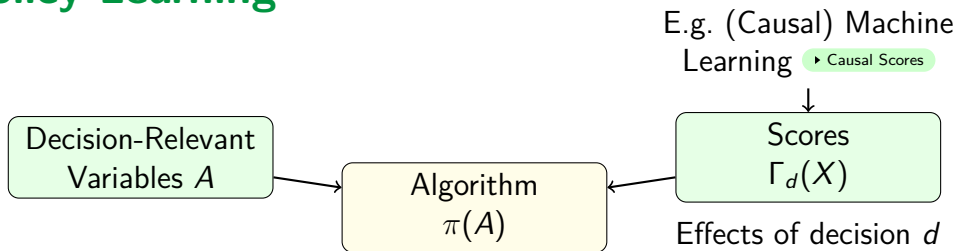
- Protected attributes S are a policy choice
- **Legal context can guide this choice.** The Swiss Federal Constitution prohibits discrimination on grounds including *origin, race, gender, age, language, ...*
- **Application:** For fair access to ALMPs in Switzerland, the focus is on *gender and origin*.
 - Swiss men, Swiss women, Non-swiss men, Non-swiss women

Decision-making Algorithm



- **Decision-Relevant Variables (A):** Labour market history, Education, Age, ...
- **Recommended Decision (D^*):** Assign an unemployed person to none or some ALMP such that policy value (welfare) is maximized
- **Evaluation of D^* :** E.g. fairness criteria assessing link between decisions and sensitive attributes S

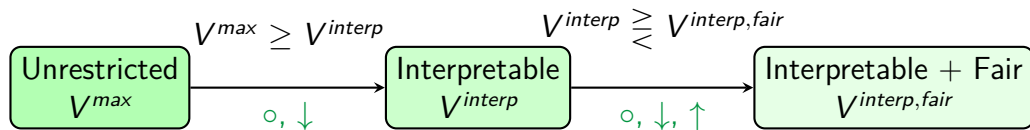
Policy Learning



- **Policy learning:** learns the policy function $\pi(\cdot)$ such that policy value V (welfare) is maximized
- **Policy function** $\pi(\cdot)$ can come from an unrestricted or restricted class of functions Π – modeling choice, e.g. a class of interpretable functions

Three Objectives

- **Policy value:** assign people to programs with the highest estimated employment gains, V^{max} .
- **Interpretability:** use transparent rules, such as shallow policy trees, V^{interp} .
- **Fairness:** reduce dependence between recommended treatments and protected attributes, V^{fair} .



Fairness Criterion - Statistical Parity

- Requires an equal share of individuals assigned to a program, e.g.:

$$P(D^* = PC \text{ course} | S = men) = P(D^* = PC \text{ course} | S = women)$$

- More generally, statistical parity requires:

$$D^* \perp S$$

- Promotes **equal opportunity**: same chance of getting treatment across S

Fairness Implementation via Pre-processing

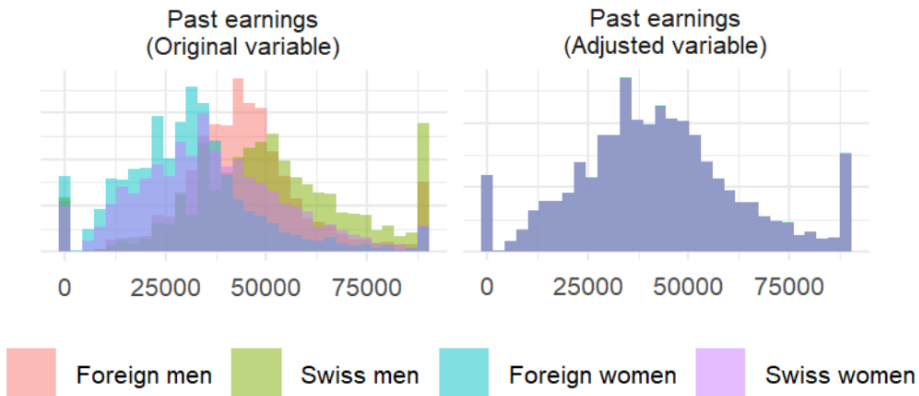
- Modifying the input **data** to promote statistical parity in decisions made by the algorithm
- Let $\tilde{\cdot}$ denote a transformation that makes some variable A (pairwise) independent of S while preserving maximal information
- If $\tilde{A} \perp S$, then $D^* = \pi(\tilde{A}) \perp S$
- Algorithm is trained using scores (capturing effects of decisions) and $\tilde{A}$ such that policy value is maximized, yielding $\hat{\pi}^{fair}(\tilde{A})$
- **Interpretability of $\tilde{A}$:** Even when $\pi(\cdot)$ is interpretable, transformations of A to $\tilde{A}$ can hinder interpretability

Pre-Processing: MQ-Adjustment (1)

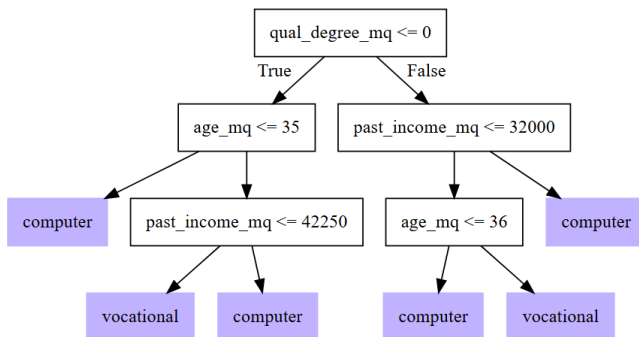
- Yielding the same distribution of $\tilde{A}$ across groups S , while preserving the rank within groups $\rightarrow \tilde{A} \perp S$ and maximum information from A is preserved (Feldman et al., 2015; Johndrow and Lum, 2019; Strack and Yang, 2024)
- Formally, $\tilde{A} = F_A^{-1} \left(F_{A|S}(A|S) \right)$, where $F_{A|S}(A|S)$ is the conditional cdf of the original A and $F_A^{-1}(\cdot)$ is its marginal quantile (MQ) function.
- For discrete A , additional randomization of tied cdf values.

$\rightarrow$ **MQ-adjustment**

Pre-Processing: MQ-Adjustment (2)



Policy Tree $\pi(\tilde{A})$



Source: Muny (2025)

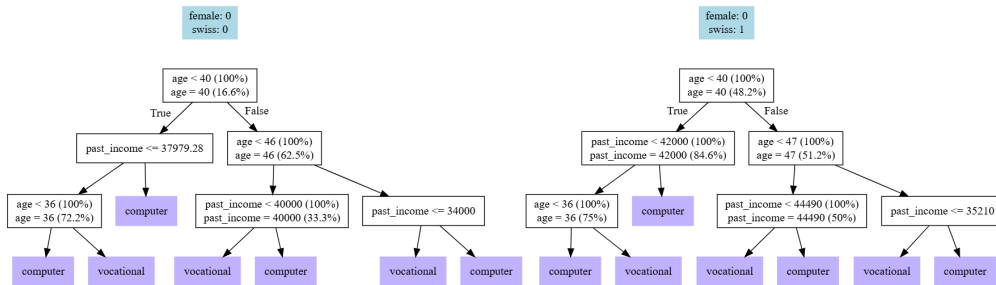
Recovering Interpretability in A : CQ-Adjustment (1)

- Let $\tilde{A}$ be constructed by MQ-adjustment, i.e. $\tilde{A} = F_A^{-1} \left(F_{A|S}(A, S) \right)$.
- For a given s , there exists a unique reverse mapping back to a for any threshold $\tilde{a}$: $g(p, s)$, where $p = F_A(\tilde{a})$.
- $g(\cdot, \cdot)$ is linked to the conditional quantile $F_{A|S}^{-1}(\cdot, \cdot) \rightarrow$ **CQ-adjustment**
- Let $\pi(\tilde{A})$ be a policy tree. CQ-adjustment yields a separate policy tree $\pi_s(A)$ for each sensitive group.

Recovering Interpretability in A : CQ-Adjustment (2)

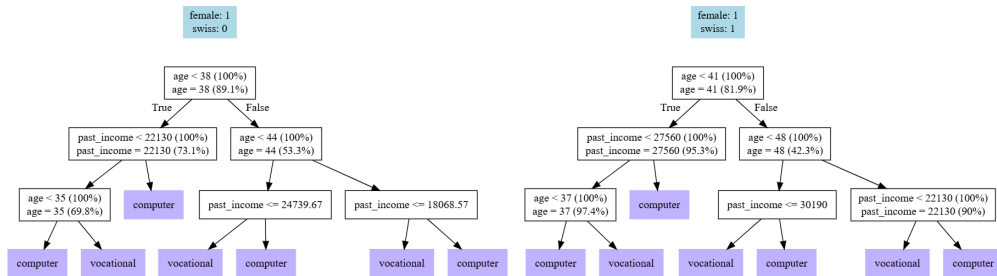
- Let A be continuous and $\tilde{A}$ be constructed by MQ-adjustment. Then CQ-adjustment yields deterministic splits in $\pi_s(A)$.
- Let A be discrete or having mass points and $\tilde{A}$ be constructed by MQ-adjustment. Then CQ-adjustment yields probabilistic splits in $\pi_s(A)$ where $\tilde{p}_s$ indicates what share of units with value a goes to the left daughter leaf.

Policy Trees $\pi_s(A)$



Source: Muny (2025)

Policy Trees $\pi_s(A)$



Source: Munny (2025)

Empirical Application

- Algorithmic assignments of unemployed (registered in 2003, aged 24-55) to active labor market policies in Switzerland
- **6 Treatments** D : No program, job search assistance, vocational training, computer course, language course, public employment
- **4 Sensitive groups** S : Gender (female/male) and nationality (swiss/foreign)
- **Decision variables** A : Past income, age, qualification
- **Outcome** Y : Months employed in 31 months after program start

Main Results

Policy	$\hat{V}$	Cramer's V	$\log(\text{BF})$
Black box	18.16	0.11	123
Policy tree	17.88	0.27	422
Fair black box	18.14	0.03	-25
Fair policy tree	17.88	0.09	36
Observed assignment	14.53	0.07	18

- Fairness can be improved with little loss in estimated average policy value.

► Policy

► Fairness Measures

Policy Tree → Fair Policy Tree: Winners & Losers

Variable	Strong loss	Moderate loss	"No change"	Moderate gain	Strong gain
Difference in policy value	-2.07	-0.90	-0.00	0.84	1.82
Number of observations in cluster	135	268	11022	276	159
Job seeker is female	0.41	0.50	0.44	0.41	0.46
Swiss citizen	0.73	0.47	0.64	0.54	0.62
Age of job seeker	41.47	41.02	36.28	41.24	41.28
Earnings in CHF before unemployment	31609	31990	43070	32394	33958
Qualification: with degree	0.38	0.19	0.59	0.23	0.21
Months employed in last 2 years	0.75	0.75	0.81	0.79	0.81
Employment spells in last 5 years	1.53	1.16	1.18	1.31	1.12
Job seeker is married	0.50	0.64	0.46	0.63	0.60
Unofficial mother tongue	0.25	0.53	0.33	0.51	0.40
Unemployment spells in last 2 years	0.94	0.79	0.55	0.81	0.59
Lives in no city	0.56	0.62	0.68	0.73	0.72
Lives in medium city	0.16	0.13	0.12	0.12	0.15
Lives in big city	0.28	0.25	0.20	0.16	0.13

Conclusion

- Construction of interpretable (approximate) fair decision algorithms
 - Fairness is implemented through pre-processing of decision variables.
 - Interpretability is recovered by mapping adjusted tree splits back to the original variables.
- Application: Allocation of ALMPs in Switzerland
 - In this particular case, fairness has a low cost in terms of policy value
 - Reallocation of ALMPs within unemployed with higher age, lower prior earnings and lower probability to hold a formal qualification

References I

- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.
- Johndrow, J. E. and Lum, K. (2019). An algorithm for removing sensitive information. *The Annals of Applied Statistics*, 13(1):189–220.

References II

- Lechner, M. (2018). Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487*.
- Muny, F. (2025). Replication notebook for fairness-aware and interpretable policy learning. https://fmuny.github.io/fairpolicytree/replication_paper/replication_notebook.html.
- Panigutti, C., Beretta, A., Giannotti, F., and Pedreschi, D. (2022). Understanding the impact of explanations on advice-taking: A user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9.

References III

- Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81.
- Strack, P. and Yang, K. H. (2024). Privacy-preserving signals. *Econometrica*, 92(6):1907–1938.
- Zhou, Z., Athey, S., and Wager, S. (2023). Offline multi-action policy learning: Generalization and optimization. *Oper. Res.*, 71(1):148–183.

Scores

To rewrite $V(\pi)$ causally, i.e., $V(\pi) = E[\Gamma_{\pi(A)}(X)] = E[\sum_{d \in \mathcal{D}} \mathbf{1}\{\pi(A) = d\} Y^d]$

- suitable score $\Gamma_d(X)$
 - Individual Average Potential Outcome (IAPO):
 $\Gamma_d^{\text{IAPO}}(X) = \mu_d(X)$ with $\mu_d(X) = E[Y|D = d, X]$
 - Augmented Inverse Probability Weighted (AIPW) score:
 $\Gamma_d^{\text{AIPW}}(X, Y, D) = \mu_d(X) + \frac{\mathbf{1}\{D=d\}(Y - \mu_d(X))}{e_d(X)}, \quad e_d(x) = P(D = d|X = x)$
- identifying assumptions
 - randomized control trial
 - unconfoundedness

► Back

Estimation

- Policy scores $\Gamma_d(X)$ estimated by Modified Causal Forest (Lechner, 2018)
 - Black box $\hat{\pi}(\Gamma) \rightarrow \max V$
 - Fair black box $\hat{\pi}(\tilde{\Gamma}) \rightarrow \text{fair}$
- Policy trees (Zhou et al., 2023) trained with $\Gamma_d(X)$ and A or $\tilde{A}$
 - Policy Tree $\hat{\pi}(A) \rightarrow \text{interpretable in } A$
 - Fair Policy Tree $\hat{\pi}(\tilde{A}) \rightarrow \text{fair and "interpretable" in } \tilde{A}$
 - $\hookrightarrow$ Fair Probabilistic Split Policy Tree(s) $\hat{\pi}_s(A) \rightarrow \text{fair and interpretable in } A$

► Back

Measures

Estimated Policy Value: $\hat{V} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(\hat{\pi}(a_i) = d) \hat{\Gamma}_d(x_i)$

Fairness Indicators:

- **Cramer's V:** Measure of association between two categorical variables
 - 0: no association
 - 1: perfect association
- **Log(Bayes Factor):** Log of ratio of two marginal likelihoods
 - $<< 0$: more evidence for independence
 - $>> 0$: more evidence for dependence
- Analyze interplay between efficiency, fairness, and interpretability

► Back