A Human-in-the-Loop Fairness-Aware Model Selection Framework for Complex Fairness Objective Landscapes Supplementary Material

Jake Robertson[1, 2], **Thorsten Schmidt**[1], **Frank Hutter**[1, 3], **Noor Awad**[1]
[1]University of Freiburg,
[2]Zuse School ELIZA,
[3]ELLIS Institute Tübingen
robertsj@cs.uni-freiburg.de, thorsten.schmidt@stochastik.uni-freiburg.de,
fh@cs.uni-freiburg.de, awad@cs.uni-freiburg.de

A Supplementary Results

A.1 Criminal Recidivism: Negative Conflicts

The Compas data set is a seminal instance of algorithmic bias, where a sentencing algorithm used in the Florida judicial system to predict criminal recidivism (likelihood of re-offending) to aid parole decisions was found to be severely biased against Black defendants (Angwin et al. 2016). In this section, we take a closer look into the so-called *inverse conflicts* we observed in the RF-Compas experiment, where optimizing for different fairness metrics found stronger solutions than optimizing for a fairness metric directly. This result suggests a possible explanation for the strong performance of the MaO problem formulation in this scenario (Main Figure 3), where interaction between fairness metrics enables MaO to discover strong overall solutions.

In Main Figure 4 we observe negative contrast values between DDSP with respect to DEOD ($C = -0.078$), DEOP ($C = -0.171$), and INVD ($C = -0.111$), indicating that a fairer solutions in terms of DEOD/P and INVD were discovered when optimizing for F_1-Score and DDSP. We also observe MaO experiments in Main Figure 3 with negative regret (-5% to -10%), indicating that a higher $\mathcal{H}_{DEOD/P}$ and $\mathcal{H}_{INVD}$ was achieved by the MaO experiment than by their corresponding BiO experiments. These results are attributed to a single model discovered on $\mathcal{P}(\mathcal{Y}_{DDSP})$ which achieves reasonable accuracy ($1 - F_1 = 0.3$) and the lowest unfairness in terms of all fairness metrics (Appendix Figure 1).

In order to better understand how strong overall fairness was achieved by this model, we compare its behavior with a slightly more accurate ($1 - F_1 = 0.27$) but less fair model in terms of all fairness metrics (Figure 1). In comparison, the all-fair model has a lower sentencing rate for Black defendants that re-offended ($P(Sentenced|Black, Guilty) = 0.09$) than the overall-unfair model ($P(Sentenced|Black, Guilty) = 0.13$). However, because both models have a high parole rate for White defendants that re-offend ($P(Parole|White, Guilty) \geq 0.20$), the decreased sentencing rate from the overall-fair model has the effect of improving overall fairness. First of all, the between-group parole rate $P(Parole|White) - P(Parole|Black)$ is improved from DDSP = 0.07 in the overall-unfair model to DDSP = 0.04 in the overall-fair model. In addition, the between-group parole rate for non-re-offending defendants $P(Parole|White, Innocent) - P(Parole|Black, Innocent)$ is improved from DEOP = 0.02 in the overall-unfair model to DEOP = 0.01 in the overall-fair model. Finally, similarity-based individual fairness INVD is also improved in the overall-fair model, as 4% more similarly re-offending defendants receive similar parole outcomes. This result suggests that optimizing for multiple notions of fairness can have the effect of unlocking regions of the objective space that are otherwise inaccessible using the BiO problem formulation.

A.2 Asymmetry of Fairness Metric Conflicts

In this section, we outline a scenario where a substantial difference in base rates leads to an asymmetric fairness metric conflict. An asymmetric fairness metric conflict occurs when the impact of satisfying fairness metric f_i on violating fairness metric f_j is different (larger or smaller) from the impact of satisfying f_j on violating f_i.

In Main Figure 4, we observe an interesting phenomenon on the RF-Lawschool experiment, where the conflict between INVD with respect to group fairness metrics DDSP and DEOP/D ($C = 0.2$) is significantly larger than the conflict between group fairness metrics DDSP and DEOP/D with respect to INVD ($C = 0.1$). In plainer terms, if INVD is satisfied it leads to a strong violation of DDSP, while the satisfaction of group fairness metrics leads to a relatively weaker violation of INVD.

In the following argument, we explain why asymmetry occurs on the Lawschool dataset by drawing a connection to the significant imbalance (92% White and 8% Black) in privileged and unprivileged applicants (Appendix Table 3). Consider a perfect classifier that satisfies INVD by accepting all qualified applicants and rejecting all unqualified ones. Referring to the distribution in Main Figure 6 (right), the classifier thus accepts all 1% of applicants who are qualified and Black as well as all 50% who are qualified and White. Similarly, the classifier rejects all 7% of applicants who are unqualified and Black as well as all 42% who are

Figure 1: *Inverse Conflicts (RF-Compas)*: Overall-fair model discovered in the RF-Compas experiment, which minimizes all fairness metrics at a reasonable accuracy level. The All-Fair model increases the parole rate for Black defendants that did in fact re-offend to the same rate as White applicants, resulting in low group and individual unfairness.

unqualified and White. Although INVD is satisfied (all individuals receive the outcome they deserve, regardless of demographic group), such an admissions strategy strongly violates DDSP (precisely resulting in $P(Accept|White) - P(Accept|Black) = \frac{42}{92} - \frac{1}{8} = 0.34$ or a difference in between-class acceptance rate of 34%).

Now consider modifying this classifier (e.g. with a post-processing technique) such that DDSP is satisfied by increasing the acceptance likelihood for unqualified Black students by 3% and decreasing the acceptance likelihood for qualified White students by 4% (positive discrimination), resulting in $P(Accept|White) - P(Accept|Black) = \frac{46}{92} - \frac{4}{8} = 0$. Such a modified classifier results in only a 24% violation of INVD, as 3% and 4% of similarly qualified (or unqualified) applicants from different demographic groups receive different admissions/rejection outcomes.

Note that the impact of DDSP on INVD depends on the base rate of privileged/unprivileged applicants, and asymmetry would increase in this scenario if the *overall* proportion of Black applicants increased while the ratio of qualified and unqualified Black applicants stayed the same. For example, if 2% of applicants were qualified and Black, while 14% of applicants were unqualified and Black, satisfying DDSP, would require a 6% (as opposed to the previous 3%) increase in acceptance likelihood for unqualified Black applicants, leading to a larger increase in INVD than in the previous example. We thus exemplify how fairness metric conflicts can be asymmetric, while also identifying the impact that dataset characteristics (e.g. difference in base rates) can have on their occurrence, significance, and symmetry. This identification suggests that fairness metric conflicts can potentially be anticipated during domain-knowledge-driven deliberations, adding a technical and concrete angle to these discussions.

B Experimental Details

B.1 Multi-Criteria Objective Function

Our objective function takes as input a hyperparameter configuration $\lambda \in \Lambda$, a FairML data set $\mathcal{D} = (X, Y, A)$, and a subset of the fairness metrics $\{f_0, f_1, f_2, ..., f_d\}$. The objective function applies Nested Stratified k-Fold Cross-Validation to iteratively partition the data set into training, testing, and validation folds $\mathcal{D}_{train}, \mathcal{D}_{val}$ and, $\mathcal{D}_{test}$. Each fold is stratified by both the target Y and protected attribute A in order to maintain a realistic distribution of these variables.

Given a candidate hyperparameter configuration $\lambda \in \Lambda$, a model $\mathcal{M}$ is defined and fit to the training fold $\mathcal{D}_{train}$, generating predictions $\hat{Y}$ on the validation set $\mathcal{D}_{val}$. The predictive performance of the hyperparameter configuration $f_0(Y, \hat{Y})$ is calculated using the F_1-Score, an appealing performance metric in the face of significant class imbalance. Because a higher F_1-Score is better with respect to predictive performance and defined in the range $(0, 1)$, we minimize $f_0(Y, \hat{Y}) := 1 - F_1$ during optimization. The *unfairness* of the hyperparameter configuration $f_{1:d}(Y, \hat{Y}, A)$ is calculated using the measures of fairness defined in Table **??**. The objective values of each evaluated hyperparameter configuration are added to an archive of all observations $\mathcal{Y}$.

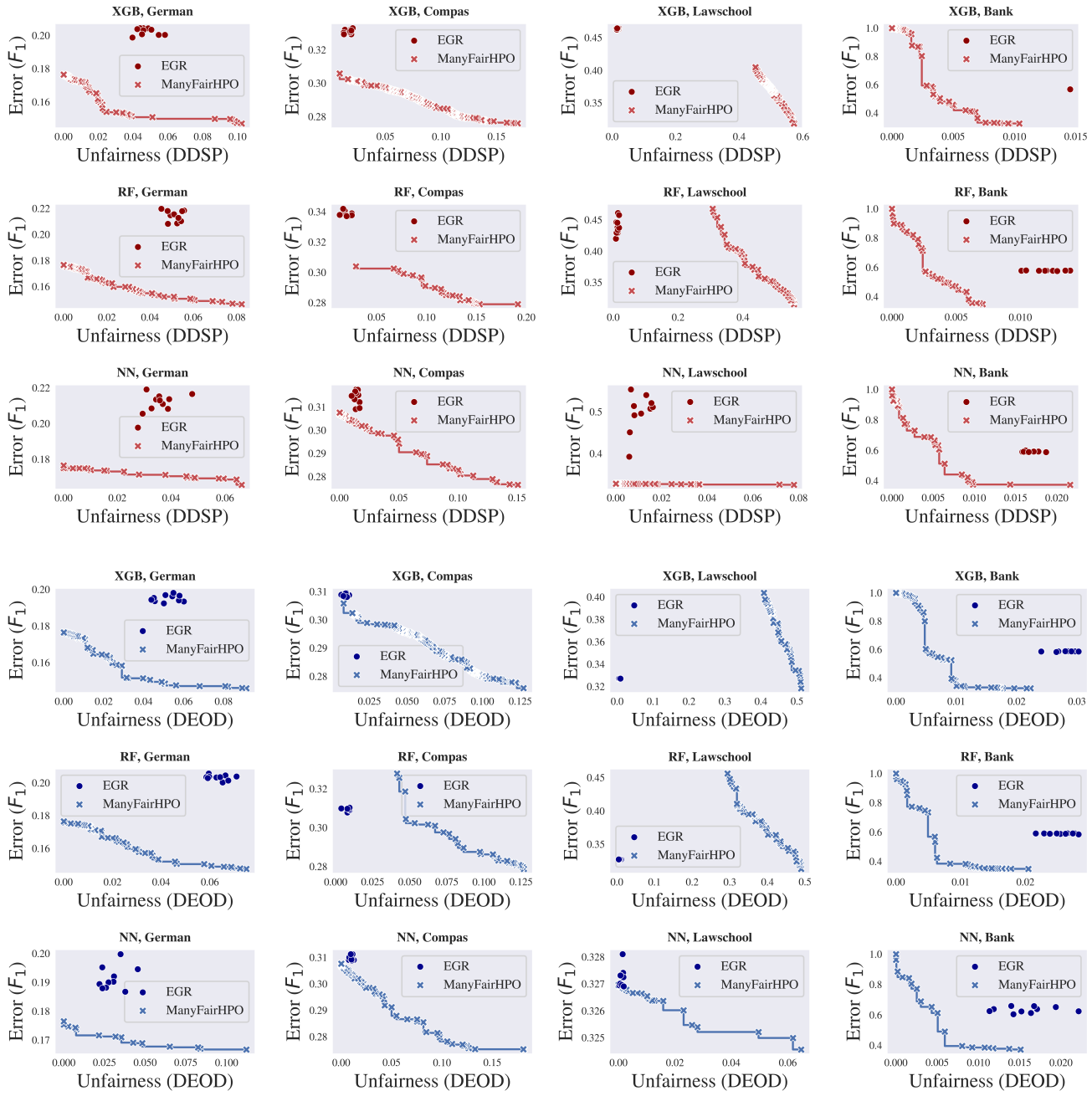

Figure 2: *ManyFairHPO vs. EGR*: Relative fairness-accuracy objective space locations of hyperparameter configurations found by ManyFairHPO and those post-processed with Exponentiated Gradient Reduction (EGR) to minimize DDSP (top) and DEOD (bottom). ManyFairHPO Pareto Fronts dominate EGR models in the majority of cases (9/12 for DDSP and 7/12 for DEOD), suggesting that HPO alone is a competitive approach to bias-mitigation.

Random Forest (NN)

Name	Range	Scale
max_depth	(1, 50)	Log
min_samples_fold	(2, 128)	Log
min_samples_leaf	(1, 20)	Uniform
max_features	(0, 1)	Uniform
n_estimators	(1, 200)	Log

XGBoost (XGB)

Name	Range	Scale
eta	$(2^{-10}, 1.0)$	Log
max_depth	(1, 50)	Log
colsample_bytree	(0.1, 1.0)	Uniform
reg_lambda	$(2^{-10}, 2^{10})$	Log
n_estimators	(1, 200)	Log

Multi-Layer Perceptron (NN)

Name	Range	Scale
depth	(1, 3)	Uniform
width	(16, 1024)	Log
batch_size	(4, 256)	Log
alpha	$(10^{-8}, 1)$	Log
learning_rate_init	$(10^{-5}, 1)$	Log
n_iter_no_change	(1, 20)	Log

Table 1: *HPO Search Spaces*: Summary of hyperparameter search spaces drawn from HPOBench.

Formulation	Name	Optimizer	Objectives	Pop. Size	Func. Evals.	Seeds
BiO	F1-DDSP	NSGA-II	2	20	1000	10
	F1-DEOD	NSGA-II	2	20	1000	10
	F1-DEOP	NSGA-II	2	20	1000	10
	F1-INVD	NSGA-II	2	20	1000	10
MaO	F1-MULTI	NSGA-III	5	42	1000	10

Table 2: *ManyFairHPO Experiments:* Summary of ManyFairHPO experiments, spaning across two problem formulations, four fairness metrics, three HPO search spaces, and five data sets. We run each experiment for 10 seeds with a maximum wall-clock time of 1 CPU day.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. *ProPublica, May*, (201barocas6): 139–159.

Name	Prot. Attr.	Samples	Features	Pos./Neg.	Priv./Unpriv.
German Credit	sex	1,000	59	70/30	69/31
Criminal Recidvism	race	5278	7	53/47	40/60
Bank Marketing	age	764	31	23/77	64/36
Census Income	sex	15,315	44	25/75	85/14
Lawschool Admissions	race	22,342	3	25/75	92/8

Table 3: *FairML Datasets:* Summary of data sets drawn from the `aif360` library.