A. Generalization to other fairness metrics

In this section, we present the performances of `LaundryML` with other fairness metrics are used. To control the rest of the parameters, the experiments are done only for the model rationalization algorithm `LaundryML-global` for a black-box model trained on the Adult Income dataset. We use the same black-box model trained with Random Forest, that we use in our previous experiments. In addition to the *demographic parity* metric, we use three different fairness metrics, namely the *overall accuracy equality* metric, the *statistical parity* metric and the *conditional procedure accuracy* metric. The definitions of all these additional fairness metrics can be found in (Berk et al., 2018). For each of these scenarios, we enumerate 50 models and use the regularization parameters $\lambda = 0.005$ and $\beta = \{0, 0.1, 0.2, 0.5, 0.7, 0.9\}$.

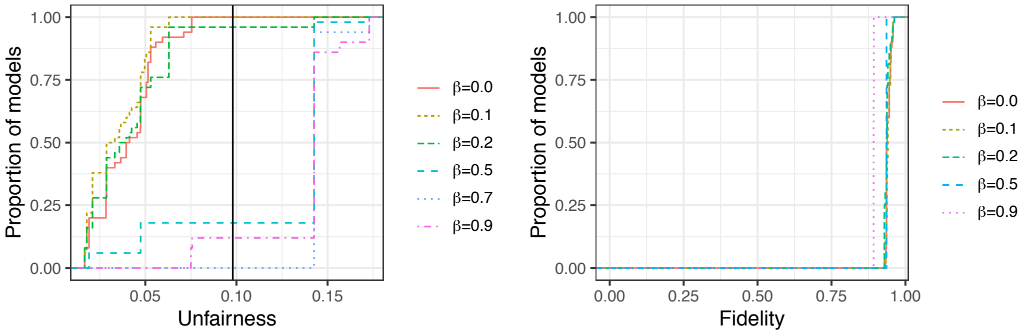

Figure 4. CDFs of the unfairness (left) and the fidelity (right) of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income. Results are for the *Overall Accuracy Equality* metric and the *Random Forest* black-box model. The vertical line on the left figure represents the unfairness of the black-box model. The CDFs on the right figure are the CDFs of the fidelity of explanation models whose unfairness are less than that of the black box model.

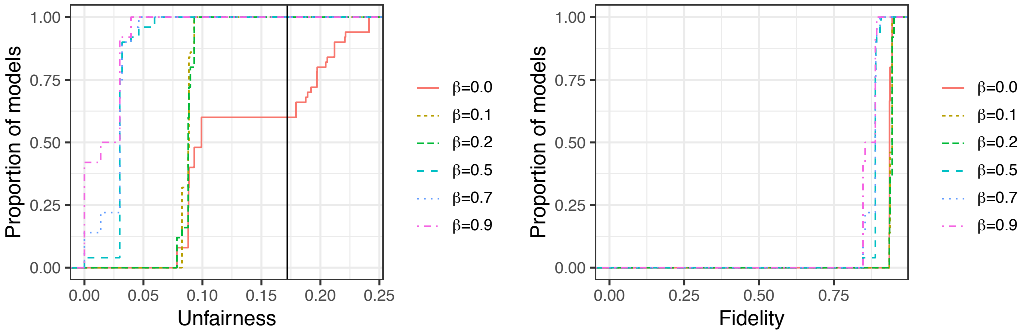

Figure 5. CDFs of the unfairness (left) and the fidelity (right) of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income. Results are for the *Statistical Parity* metric and the *Random Forest* black-box model. The vertical line on the left figure represents the unfairness of the black-box model. The CDFs on the right figure are the CDFs of the fidelity of explanation models whose unfairness are less than that of the black box model.

Figures 4, 5, 6 and 7 show the performances of `LaundryML-global` when respectively *overall accuracy equality*, *statistical parity*, *conditional procedure accuracy* and *demographic parity* are used as fairness metrics, and the black-box model is a *Random Forest* classifier. Overall, for all these fairness metrics, `LaundryML-global` can find explanation models to use for fairwashing. In particular, when the unfairness of the black-box model is high (*i.e.,* unfairness ≥ 0.1), a higher unfairness regularization (*i.e.,* $\beta \geq 0.5$) allows to find more potential candidate models for fairwashing. In contrast, when the unfairness of the black-box model is low (*i.e.,* unfairness < 0.1), a lower unfairness regularization (*i.e.,* $\beta < 0.5$) enables to find more potential candidate models for fairwashing. In general, the smaller the regularization, the better the fidelity. Overall, these results confirm that the possibility of performing fairwashing is agnostic to the fairness metric considered.

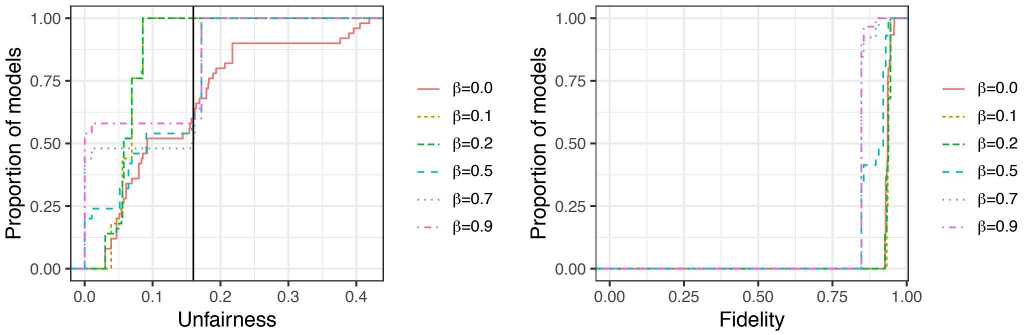

Figure 6. CDFs of the unfairness (left) and the fidelity (right) of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income. Results are for the *Conditional Procedure Accuracy* metric and the *Random Forest* black-box model. The vertical line on the left figure represents the unfairness of the black-box model. The CDFs on the right figure are the CDFs of the fidelity of explanation models whose unfairness are less than that of the black box model.

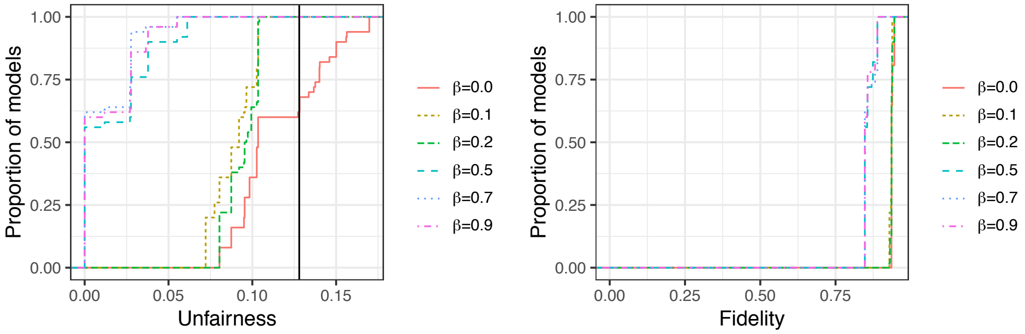

Figure 7. CDFs of the unfairness (left) and the fidelity (right) of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income. Results are for the *Demographic Parity* metric and the *Random Forest* black-box model. The vertical line on the left figure represents the unfairness of the black-box model. The CDFs on the right figure are the CDFs of the fidelity of explanation models whose unfairness are less than that of the black box model.

B. Generalization to other black-Box models

In this section, we present the performances of `LaundryML` when other black-box models are used. To control the rest of the parameters, the experiments are done only for the model rationalization algorithm `LaundryML-global` with *demographic parity* as fairness metrics. In addition to the *Random Forest*, we use three different types of black-box models, namely a *Support Vector Machine* (SVM) classifier, a *Gradient Boosting* (XGBOOST) classifier and a *Multi-Layer Perceptron* (MLP) classifier. For each of these scenarios, we enumerate 50 models and use the regularization parameters $\lambda = 0.005$ and $\beta = \{0, 0.1, 0.2, 0.5, 0.7, 0.9\}$.

Figures 7, 8, 9 and 10 show the performances of `LaundryML-global` when the black-box models are respectively a *Random Forest* classifier, a *SVM* classifier, a *XGBOOST* classifier and a *MLP* classifier, and the fairness metric is *demographic parity*. For each type of black-box model, `LaundryML-global` can find explanation models to use for fairwashing. Overall, these results confirm that the possibility of performing fairwashing is also agnostic to the black-box model.

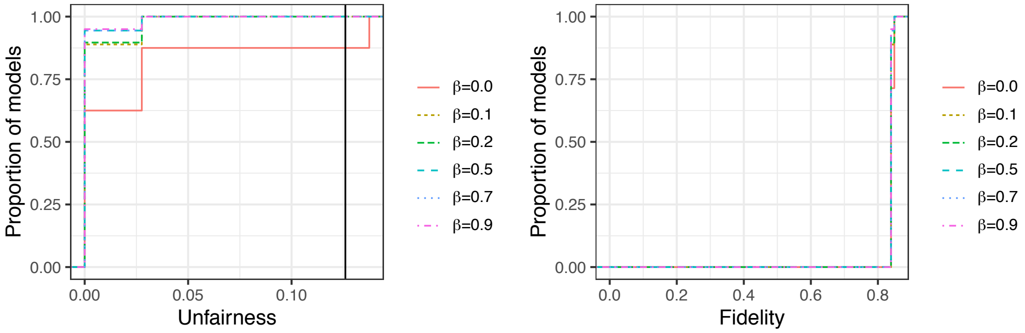

Figure 8. CDFs of the unfairness (left) and the fidelity (right) of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income. Results are for the *Demographic Parity* metric and the *SVM* black-box model. The vertical line on the left figure represents the unfairness of the black-box model. The CDFs on the right figure are the CDFs of the fidelity of explanation models whose unfairness are less than that of the black box model.

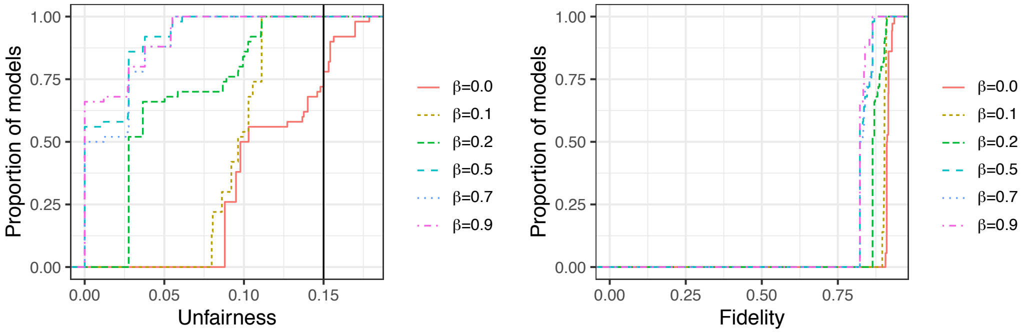

Figure 9. CDFs of the unfairness (left) and the fidelity (right) of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income. Results are for the *Demographic Parity* metric and the *XGBOOST* black-box model. The vertical line on the left figure represents the unfairness of the black-box model. The CDFs on the right figure are the CDFs of the fidelity of explanation models whose unfairness are less than that of the black box model.

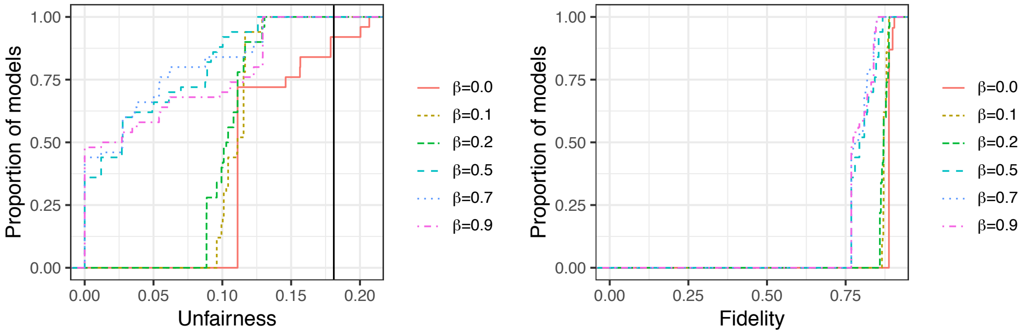

Figure 10. CDFs of the unfairness (left) and the fidelity (right) of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income. Results are for the *Demographic Parity* metric and the *MLP* black-box model. The vertical line on the left figure represents the unfairness of the black-box model. The CDFs on the right figure are the CDFs of the fidelity of explanation models whose unfairness are less than that of the black box model.