Master in Artificial Intelligence 2022-23

Explainable and Trustworthy AI

I4+I5+I6 Answer Report

Pedro Guijas Bravo

March 18, 2023

Contents

1 Introduction

2 Week 4: Notebook 4

2.1 Exercise 1

The function fatf_dam.sampling_bias_grid_check uses a default threshold of 0.8 to compute the pairwise sampling bias. Let us now modify this threshold to use a value of 0.5, and check the resulting output.

- **Do results now change? How?**

  Indeed the result has changed. This is since the *sampling_bias_grid_check* function checks the bidirectional ratio, so that if it is below a threshold the pair is considered to suffer from a sampling bias.

  By lowering the threshold value, the bias case identified above is not detected.

- **What does the given threshold account for? (you might want to check the corresponding API documentation)**

  As mentioned above, the threshold will determine whether sampling bias occurs for a bidirectional proportion. That is, if the systematic bias is violated.

- **Given the defined group counts, what would be the minimal threshold that would cause the function to inform of existence of sampling bias?**

  The function computes the bidirectional ratio as the logical operation of applying the threshold on the absolute value of the directional ratios - 1. In fact, the threshold that is passed is also transformed as threshold - 1.

  Thus, since the bidirectional ratios are 0.222 periodic and 0.18, the minimum condition for a bias to be detected is that the threshold is higher (remember that we are dealing with the inverse) than the ratio with a higher value.

  Therefore, the minimum threshold will be 0.777 periodic $(1 - (55/45 - 1))$.

2.2 Exercise 2

Similar to the previous fatf_dam.sampling_bias_grid_check, fatf_accountability_models.systematic_performance_bias_grid, you can use certain thresholds for this evaluation and answer the following questions:

- **What is the default threshold applied in this case?**

  The default threshlohld value applied is again 0.8.

- **Is the criterion for bias evaluation somehow different to the case of sampling bias seen in the previous section?**

  No, the operation is identical. This can be seen by checking the source code of the function, where the same operations mentioned in the previous question are computed.

- **Experiment with the code below and find a threshold that would result in the model to be qualified as containing systematic performance bias. Which groups would be involved?**

  Applying the same reasoning as in the previous exercise we will look for the groups with accuracy values with a higher differentiation, in this case 1 and 0.9090 periodic (group 0 and 2). Calculating the maximum inverse ratio, we conclude that it is the following : $1/0.9090 - 1$. Thus, the minimum threshold will be: $1 - (1/0.9090 - 1)$, which corresponds to 0.899.

2.3 Exercise 3

- **Which conclusions do we obtain now?**

  Given a model (KNN) and 2 groups of instances, formed from the values of a feature, i.e., values greater than and less than or equal to 3 of feature 1 (sepal width). The confusion matrices for each group are obtained using the model, as well as their associated accuracy.

  Looking first at the confusion matrices, a bias can be observed, and that is that for the first partition most of the instances are classified as class 1, although 33 of them are misclassified; note that 8 instances of class 0 are well classified. As for the second matrix, again all instances of class 1 and 2 are erroneously classified as class 1. Furthermore, for this second matrix most samples will belong to class 0, which is linearly separable and are well classified. The differentiations between the matrices show a correlation between attributes as well as a bias between matrices.

  Next, given the accuracies for each group (0.60 and 0.75) the systematic performance bias is calculated with a default threshold of 0.8. The existence of a bias is confirmed.

- **Execute the previous code again, but now select "Positive Predictive Value (PPV)" as performance metric. Do you observe any difference?**

  The only thing that will change will be the values of the metrics, abandoning the use of accuracy in favor of Positive Predictive Value. In this case a peculiarity has to be taken into account, and that is that since we are not dealing with a binary problem, it is necessary to select a class index by means of the property "label_index" on which to compute the metric (explained in detail in the next question).

  Thus, by computing the metric on the different classes, different values will be obtained:

  - **Class 0**: The PPV values for each group are perfect (1). Recall that this class is the linearly separable one. There is no bias.
  - **Class 1**: The PPV values obtained are 0.56 and 0.32 respectively for each group. That is, for the instances with a sepal width greater than 3 the model has more misclassifications when the model result is that it belongs to class 1.
  - **Class 2**: The PPV values for each group are null (0). No instances are classified as class 2, so there is no bias.

- **What does "PPV" account for?**

  The positive predictive value is calculated by the following formula $PPV = \frac{TP}{TP+FP}$, i.e., it will allow to determine how many of the samples that are classified as positive, are actually positive.

  As previously mentioned this is posed for a binary problem, being in this case focused with the parameter *label_index* to a one-versus-all approach (classified as X which really are not).

- **Which is the role of the parameter "label index" in the performance_per_subgroup function? Does it make a difference changing it ("0", "1", or "2") if using "accuracy"? Does it make a difference changing it if using "PPV"? Why?**

  This has already been explained in the previous questions. Note that for accuracy it is not necessary to translate the problem to a binary problem, since the accuracy will be computed as the number of well-classified instances of the total, without distinguishing between classes.

2.4 Exercise 4

- **Check the API documentation and explain what the "density score" actually represents.** The density score is a measure of the density of a particular data point relative to its neighboring data points. It is calculated by looking at the n-th neighbor distance defined by the parameter *neighbours*. If the n-th neighbor distance is relatively large compared to all other distances between data points in the data set, it means that the data point is in a low density region. The density score can be normalized to an interval of [0, 1] by setting the parameter *normalise_scores* to True.

In other words, the density score is a value that represents the degree to which a data point is an outlier or an anomalous point in its local neighborhood. A higher density score means that the data point is more likely to be part of a dense cluster of points, while a lower density score means that the data point is more likely to be an outlier or part of a sparsely populated region of the dataset.

- **Given the previous results. Should we trust the model's output classifications? Is there any difference in this regard between the dense and spare point classifications?**

  For both classifications we have different drawbacks. First, for the sparse point, although the probability of membership points only to class 1 (computed given the N closest instances, 3 by default), it should be noted that no truly similar or representative samples are available. Moreover, in the plot it appears that indeed the 3 closest ones are the blue ones, but by a very little difference, being in a point between the middle of class 0 and 1.

  As for the dense point, although it has similar and representative examples given the high density, it is in a nonlinear zone bordering with class 2 and 3. It may in this case that the bad fit of the model, i.e. a very low parameter K so close to the border induces a failure in the classification. Moreover the entropy in the decision probability is really high, it does not have a high certainty to correctly classify the instance.

3 Week 5: Notebook 5 and IBM Research Trusted AI tool

3.1 Exercise 1

Modify the previous code with the aim of visualizing the fraction of people who earn over $50k regarding race. Then, plot correlations regarding age and race. Finally, remove correlations with alpha=0.7 for race_Black.

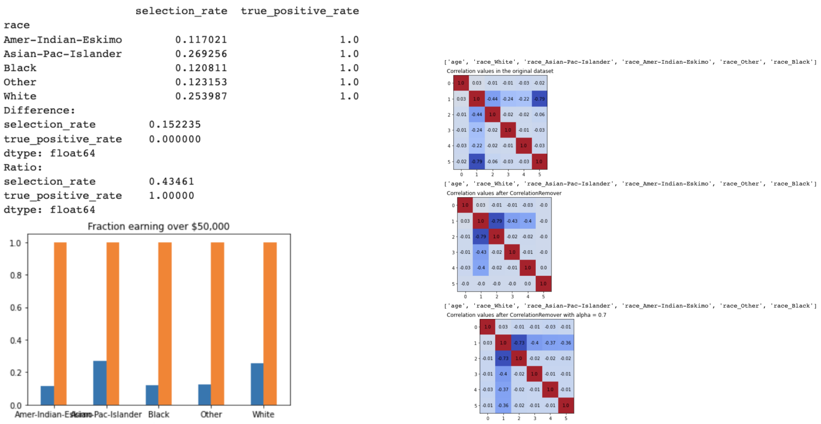
Just modify the given code, essentially changing where *sex_Male* appears by *race_Black* and where *sex_Female* appears by the three attributes. Modifying the alpha is trivial. The resulting code is shown in Listing 1. The output is shown in Figure 1.

Listing 1: Notebook 5, Exercise 1 - Code

```
metrics_dict = {"selection_rate": selection_rate, "true_positive_rate": true_positive_rate}
# true_positive_rate, true_negative_rate, true_negative_rate, false_positive_rate
selection_rates = MetricFrame(
    metrics=metrics_dict, y_true=y_true, y_pred=y_true, sensitive_features=race
)
print(selection_rates.by_group)
print("Difference:")
print(selection_rates.difference())
print("Ratio:")
print(selection_rates.ratio())
fig = selection_rates.by_group.plot.bar(
    legend=False, rot=0, title="Fraction earning over $50,000"
)

X_raw = adult_data.data[["age", "race"]]
X_raw = pd.get_dummies(X_raw)
plot_heatmap(X_raw, "Correlation values in the original dataset")
cr = CorrelationRemover(sensitive_feature_ids=["race_Black"])
X_cr = cr.fit_transform(X_raw)
X_cr = pd.DataFrame(X_cr, columns=[’age’, ’race_White’, ’race_Asian-Pac-Islander’,
     ’race_Amer-Indian-Eskimo’, ’race_Other’])
X_cr["race_Black"] = X_raw["race_Black"]
plot_heatmap(X_cr, "Correlation values after CorrelationRemover")
```

```
cr_alpha = CorrelationRemover(sensitive_feature_ids=['race_Black'], alpha=0.7)
X_cr_alpha = cr_alpha.fit_transform(X_raw)
X_cr_alpha = pd.DataFrame(X_cr_alpha, columns=['age', 'race_White',
    'race_Asian-Pac-Islander', 'race_Amer-Indian-Eskimo', 'race_Other'])
X_cr_alpha["race_Black"] = X_raw["race_Black"]
plot_heatmap(X_cr_alpha, "Correlation values after CorrelationRemover with alpha = 0.7")
```



```
                    selection_rate  true_positive_rate
race
Amer-Indian-Eskimo        0.117021                 1.0
Asian-Pac-Islander        0.269256                 1.0
Black                     0.120811                 1.0
Other                     0.123153                 1.0
White                     0.253987                 1.0
Difference:
selection_rate        0.152235
true_positive_rate    0.000000
dtype: float64
Ratio:
selection_rate        0.43461
true_positive_rate    1.00000
dtype: float64
```

```
['age', 'race_White', 'race_Asian-Pac-Islander', 'race_Amer-Indian-Eskimo', 'race_Other', 'race_Black']
```

```
['age', 'race_White', 'race_Asian-Pac-Islander', 'race_Amer-Indian-Eskimo', 'race_Other', 'race_Black']
```

```
['age', 'race_White', 'race_Asian-Pac-Islander', 'race_Amer-Indian-Eskimo', 'race_Other', 'race_Black']
```

Figure 1: (Notebook 5, Exercise 1) - Results

3.2 Exercise 2

Repeat the previous analysis but considering now education as sensitive feature. Pay special attention to HS-grad versus Masters and comment the observed results.

As with the example relating to sex, it can be seen how the predictor that does not take fairness into account leads to low fairness decisions. The results related to education as a sensitive feature are shown in Figure 2. In these, a clear differentiation can be seen for certain metrics depending on the type of education.

Accuracy is the only metric that, although there is a differentiation between different types of education, there is hardly any difference in the results of the metrics. However, for the rest of the metrics, there are very abrupt variations between metrics for each type of education. It can be seen that these variations are not only due to the availability of a very differentiated number of samples, since education with more samples such as *HS-grad*, *bachelors* or *home-college* show very different values for various metrics, among which *selection_rate* or *true_negative_rate* stand out. Thus, in addition to the fact that the dataset is clearly unbalanced in relation to education, it can be stated that considering the differences in metrics obtained for each education class, the model is not being entirely fair.

In particular, and as the statement suggests, paying special attention between *HS-grad* vs. *Masters*, a ridiculous differentiation in multiple metrics can be appreciated. As mentioned before, taking out the accuracy, for the rest of metrics the difference is really considerable, starting with the number of samples and ending with the rest of metrics.

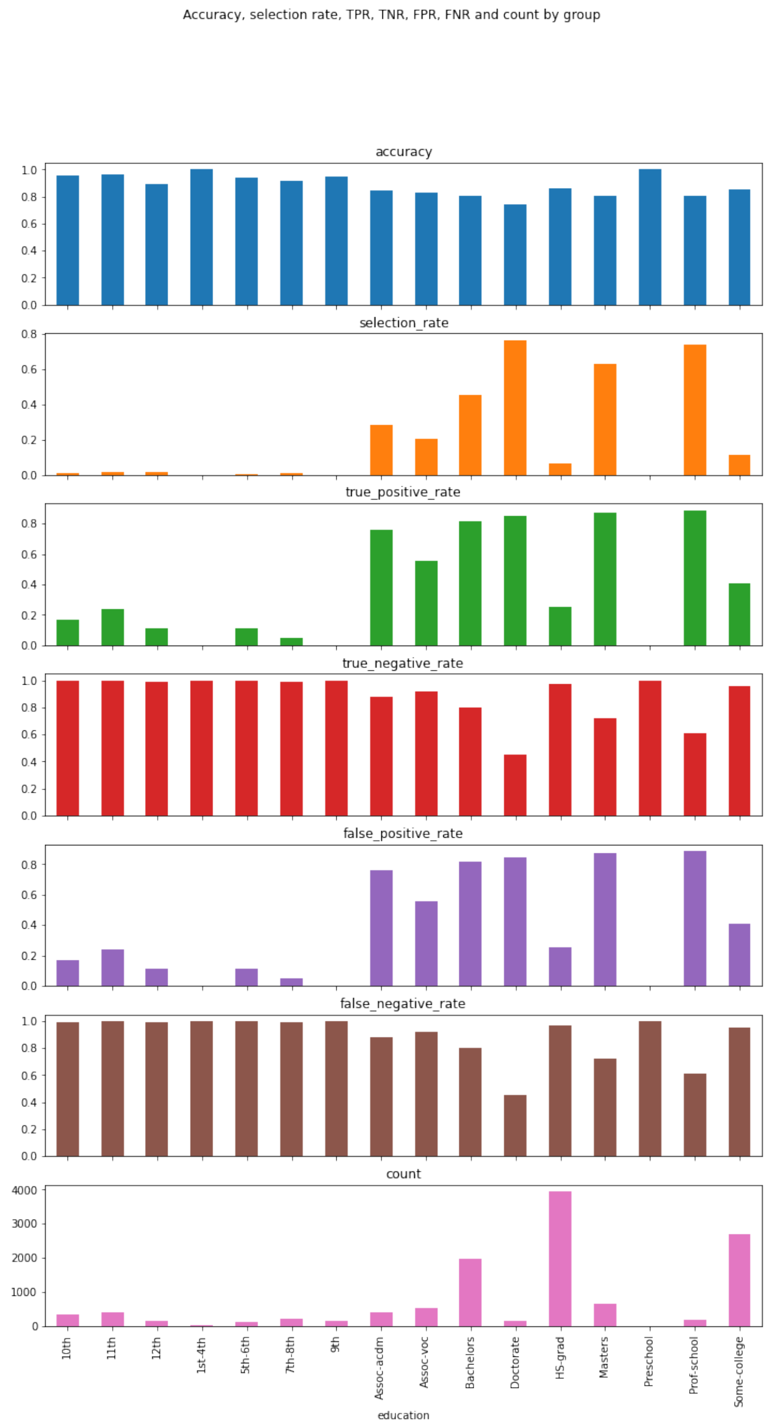


Figure 2: Accuracy, selection rate, TPR, TNR, FPR, FNR and count by educational groups

3.3 Exercise 3

First, compare the quality of the models with and without mitigating demographic bias regarding sex. Discuss the observed difference. Second, run the code below and use the fairness dashboard to try with other sensitive features and metrics. For example:

- **sensitive feature = race; performance metric = Accuracy; fairness metric = Demo-**

graphic parity ratio

- **sensitive feature = race; performance metric = F1-score; fairness metric = Demographic parity difference**

The results for the different models that attempt to mitigate bias are shown in Figures 1, 2, 3 and 4, while Figure 5 shows the results resulting from not mitigating any bias.

Considering the quality as such, first of all pay attention to how the number of samples remains stable, however, the accuracy and selection rate are affected. Since the bias is evident for the model that does not mitigate it, we will analyze those that in principle try to mitigate it. For these models the accuracy is affected, being lower or equal (only for non_dominated 3) to the model without mitigations. In addition, a relationship can be seen between the models that most closely match the selection rate and those that obtain worse results in relation to accuracy. It should be noted that in no way similar accuracies are obtained for both genders. In a more visual way, the differences can be seen in Figure 3.

	accuracy	selection_rate	count
Female	0.898868	0.149852	4064.0
Male	0.798822	0.153062	8147.0

Table 1: Results for non_dominated 0

	accuracy	selection_rate	count
Female	0.912648	0.125738	4064.0
Male	0.808641	0.183749	8147.0

Table 2: Results for non_dominated 1

	accuracy	selection_rate	count
Female	0.923720	0.099409	4064.0
Male	0.814287	0.217872	8147.0

Table 3: Results for non_dominated 2

	accuracy	selection_rate	count
Female	0.928150	0.077264	4064.0
Male	0.817233	0.253713	8147.0

Table 4: Results for non_dominated 3

	accuracy	selection_rate	count
Female	0.928150	0.077264	4064.0
Male	0.817233	0.253713	8147.0

Table 5: Results for unmitigated_predictor

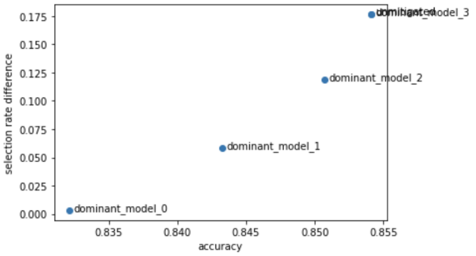


Figure 3: Trade-off between selection rate difference and accuracy

In relation to the second part of the exercise, the results are analyzed from the fairness dashboard according to race, since gender has already been discussed previously. In the following, therefore, the mentioned examples will be discussed, and the results can be visualized in Figures 4 and 5 respectively.

First of all, it should be taken into account that for the fairness demographic parity ratio metric, the optimum value will be 1, while values close to 0 are undesirable. For demographic parity difference the opposite will happen, and the values closest to 0 will be the best (there is no difference between demographic parities).

Thus, as can be seen and following the line of the explanations given, the attempts to mitigate discrimination are closely related to the decrease in the overall performance of the model, reflected in the obtaining of worse results in the metrics than the original model. However, it should be noted that the models extracted by the grid do not practically improve the fairness of the original model. For the demographic parity ratio the original model is the closest to 1 (although far away), while for the demographic parity difference, although other models manage to have a lower value, the difference with the unmitigated model is negligible (besides the notorious decrease in F1 score).

The conclusions drawn for these two examples of model performance and fairness metrics as they relate to race can be carried over to all other metric configurations. However, these will not be discussed individually for the sake of brevity, since there are many combinations of metrics. It is worth mentioning, however, that with the use of metrics that account for imbalance (balanced precision, for example), there will be models with better results on performance metrics than the original. However, this is not considered representative of the overall model performance, as they will actually be ”punishing” the correct classification of most samples. It is worth noting, by way of reflection, that learning models based on data without biases would be impossible, since the models exploit them in order to perform classifications. Furthermore, if we wish to obtain a simple model, without overtraining to mitigate biases, we will have to try to increase the complexity of the data or of the model itself.

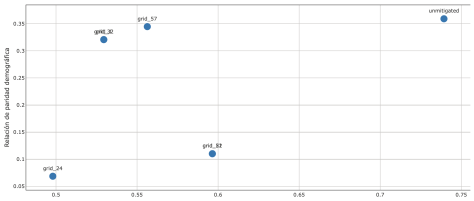


Figure 4: Trade-off between demographic parity ratio and accuracy

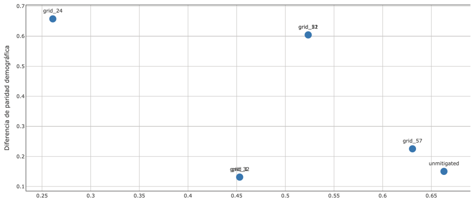


Figure 5: Trade-off between demographic parity difference and F1 score

3.4 IBM Research Trusted AI

Here we will complete the last part of the work for week 5. For this part, the *IBM Research Trusted AI* platform will be used, specifically the *AI Fairness 360 - Demo*. In order not to change the context, it has been decided to use the same dataset as for the notebook questions, that of *Adult census income*.

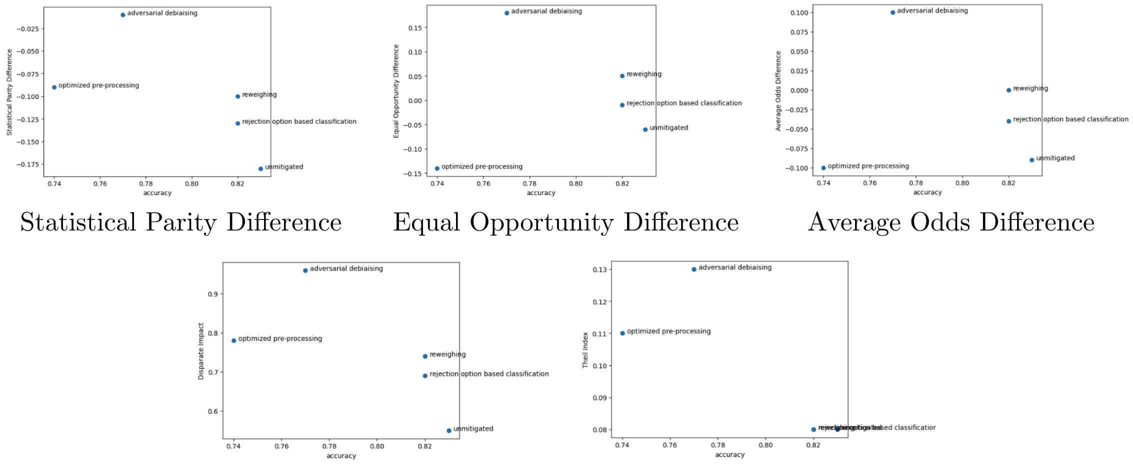
In this dataset, there are two protected attributes, race and sex, so different bias mitigation algorithms will be employed for each attribute. Figures 6 and 7 show the results in the form of a Pareto Front (Accuracy vs Bias metrics) for the protected attributes race and sex respectively. It should be noted that the optimal point for the bias metrics will be 0, with the exception of *Disparate Impact*, where 1 implies perfect fairness.

Firstly, and following the line of comments, it should be noted that for both attributes, the best model in terms of accuracy is the one to which no bias mitigation algorithm (*unmitigated*) has been applied. In general, the models generated after applying *reweighing* and *rejection option based classification* are at very similar levels of accuracy to the *unmitigated* model, but with somewhat lower levels of bias. Specifically, these techniques achieve very good results in metrics such as *Equal Opportunity Difference* or *Average Odds Difference* only by sacrificing a maximum of 2% of accuracy.

Adversarial debiasing is one of the methods that undoubtedly most negatively affects the model accuracy. This method obtains the lowest bias results for both problems according to the metrics *Statistical Parity Difference* and *Disparate Impact*, however, for the rest of the bias metrics it is the algorithm that produces the worst results, surpassing even the *unmitigated* model in *Equal Opportunity Difference*.

The method that reduces model accuracy the most is *optimized pre-processing*, at around 72%, almost 10% less than the 83% of the *unmitigated* model. Furthermore, this method does not correctly reduce the bias, since the only metric that claims to do so is *Statistical Parity Difference*, applying the algorithm on the race attribute. This is undoubtedly the worst algorithm, as it is the one that reduces the accuracy the most and only mitigates the bias according to one metric, in which *adversarial debiasing* achieves a better accuracy and bias value.

Finally, if one had to select a method *reweighing* or *Reject Option Based Classification* would be preferred. Forcing the selection of a single candidate, *reweighing* would be preferred, as it is considered slightly superior looking at the metrics in which neither of the two are outstanding.



Statistical Parity Difference

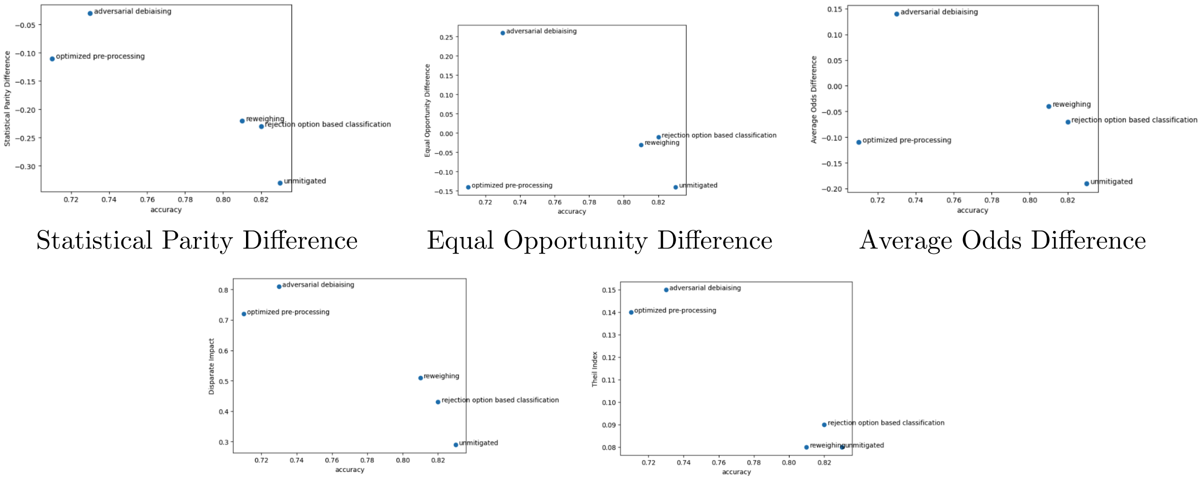
Equal Opportunity Difference

Average Odds Difference

Disparate Impact

Theil Index

Figure 6: Pareto Front (Accuracy vs Bias metrics) after applying different bias mitigation methods on the *race* attribute.



Statistical Parity Difference Equal Opportunity Difference Average Odds Difference

Disparate Impact Theil Index

Figure 7: Pareto Front (Accuracy vs Bias metrics) after applying different bias mitigation methods on the *sex* attribute.

4 Week 6: ALTAI

Summary of the case study Q Industries is a global defense company that develops autonomous vehicles, including bomb-defusing robots and crowd-monitoring drones, which have been widely used by military and law enforcement. Q has faced challenges since individual attackers have started damaging their vehicles. Therefore, Q has experimented with automated active responses, such as facial recognition algorithms and non-lethal responses like tear gas or acoustic weapons. Recently, Q has been approached, in secret, by various governments to expand its response to include lethal capabilities, which led to some of its engineers resigning in protest. Q sued them for violating their confidentiality employment agreement when they planned to speak publicly about their concerns, including the inadequate protection of non-lethal responses from tampering.

Assumptions about the case study Some assumptions have been made in order to answer some of the ALTAI's form. First of all, we have assumed that, although the system is autonomous, there exists a mechanism to take the whole control of the robot/drone, in case it is needed. Also, personal data has been used to train the system such as race, genre, or age. All the classes would be balanced, in principle. Finally, we consider the secret meetings as a kind of stakeholder participation. No other information is assumed but the one provided by the example, so any question that involves specific designs, metrics or any other kind of mechanisms not mentioned in the example, is considered to not be implemented.

In summary, the recommendations for the specific use case for each section proposed by ALTAI will be discussed below. See first the spyder-plot in Figure 8, as it reflects at a glance the situation of the use case.

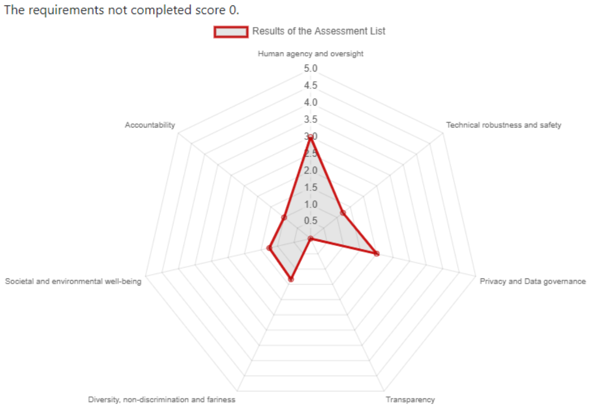


Figure 8: ALTAI results

- **Human Agency and Oversight**: Q Industries's started to behave completely autonomously since the implementation of the active responses. One of the main problems is that this behavior is not controlled, and the developers have not been trained on how to oversight them properly. Therefore, mechanisms must be developed to reflect the autonomous nature of the system, detect undesirable responses and teach humans how to oversight them. One positive aspect regarding this section is that the system can always be interrupted so the control can be derived to a human when necessary, this is why we get 3 points in the spider plot.

- **Technical Robustness and Safety**: About this section, although some possible risks have been identified (e.g. racism), no metric have been extracted, its possible misuse has not been properly assessed, no plan for fault tolerance has been developed, no monitoring or logging is performed, etc. A lot of work has to be performed in this aspect, defining the necessary mechanisms to ensure the technical robustness and safety of the system.

- **Privacy and Data Governance**: The accuracy of the system was preferred over the privacy of the individuals, so a lot of personal information was used, in addition to the use of possible non-legal data provided by different stakeholders. Due to the international scope, specific standards have not been followed, damaging the right to privacy and the protection of individuals' data. As in the previous section, mechanisms to flag issues related to privacy or data protection over the course of the AI system's lifecycle must be implemented.

- **Transparency**: For Q Industries, it is recommended to prioritize transparency in its AI systems. This involves providing explanations of the decisions made by the system. By prioritizing transparency, Q Industries can build trust among its users and ensure that its AI systems are perceived as reliable and trustworthy, especially for such a sensitive topic as warfare and citizen security.

- **Diversity, non-discrimination and fairness**: In terms of ensuring diversity, nondiscrimination and equity in AI systems, a number of measures are recommended. To avoid bias, it suggests establishing strategies and procedures, testing for specific target groups, and using state-of-the-art technical tools to understand data and performance. The section also recommends testing and monitoring for potential bias throughout the AI system lifecycle, educating AI designers and developers about potential bias, and implementing clear measures to flag problems related to bias, discrimination, or poor performance. In addition, it recommends establishing mechanisms to ensure fairness, considering other definitions of fairness, consulting affected communities, and ensuring that the AI system corresponds to society's range of preferences and capabilities. Finally, it recommends assessing the potential disproportionate impact on specific groups and the risk of unfairness on end-user or subject communities.

- **Societal and environmental well-being**: Q Industries' is not taking into account the social and environmental well-being of the systems in its system. Therefore multiple recommendations could be applied such as: taking into account positive and negative impacts on the environment, establishing mechanisms to assess impact, informing and consulting affected workers and stakeholders, analyzing work processes and the socio-technical system, and provide training opportunities for up-skilling and re-skilling.

- **Accountability**: A modular and auditable system that can be tracked and recorded should be included and designed to ensure accountability, as well as involving third parties to report vulnerabilities and risks. It is also recommended to provide appropriate training for developers and deployers, and to include various stakeholders in an ethical review board to monitor and assist the development process. Additionally, risk management processes should be continually revised with new findings, and redress should be provided when incorrect predictions cause adverse impacts to individuals, particularly vulnerable groups.

Recommendations:

- **Human agency and oversight**.
  - Put in place procedures to avoid that end users over-rely on the AI system.
  - Give specific training to humans (human-in-the-loop, human-on-the-loop, human-in-command) on how to exercise oversight.
  - Establish detection and response mechanisms in case the AI system generates undesirable adverse effects for the end-user or subject.
  - Adopt specific oversight and control measures to reflect the self-learning/autonomous nature of the system.
- **Technical robustness and safety**
  - Define risk, risk metrics and risk levels of the AI system in each specific use case.
  - Assess the risk of possible malicious use, misuse or inappropriate use of the AI system.
  - Assess the dependency of critical system's decisions on its stable and reliable behaviour.
  - Assess the dependency of critical system's decisions on its stable and reliable behaviour.
  - Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or "conventional").
  - Develop a mechanism to evaluate when the AI system has been changed enough to merit a new review of its technical robustness and safety.Develop a mechanism to evaluate when the AI system has been changed enough to merit a new review of its technical robustness and safety.
  - Put in place a series of steps to monitor and document the AI system's accuracy.
  - Put in place processes to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated.
  - Put in place a well-defined process to monitor if the AI system is meeting the goals of the intended applications.
  - Test whether specific contexts or conditions need to be taken into account to ensure reproducibility.
  - Put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the system's reliability and reproducibility.
  - Clearly document and operationalize processes for the testing and verification of the reliability and reproducibility of the AI system.
  - Define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them.
  - Put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score.

- **Privacy and Data Governance**
  - Take measures to consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection.
  - Consider establishing mechanisms that allow flagging issues related to privacy or data protection concerning the AI system.
  - When relevant, implement the right to withdraw consent, the right to object and the right to be forgotten in the AI system.
  - Consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's lifecycle.
  - Consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data.
  - Whenever possible and relevant, align the AI-system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for (daily) data management and governance.
- **Transparency**
  - Consider explaining the decision adopted or suggested by the AI system to its end users.
  - Consider continuously surveying the users to ask them whether they understand the decision(s) of the AI system.
  - In case of interactive AI system, consider communicating to users that they are interacting with a machine.
- **Diversity, non-discrimination and fairness**
  - Consider establishing a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design.
  - Test for specific target groups or problematic use cases.
  - Research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance.
  - Assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness).
  - Put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system.
  - Depending on the use case, ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system.
  - You should establish clear steps and ways of communicating on how and to whom such issues can be raised.
  - Your definition of fairness should be commonly used and should be implemented in any phase of the process of setting up the AI system.
  - Consider other definitions of fairness before choosing one.
  - Consult with the impacted communities about the correct definition of fairness, such as representatives of elderly persons or persons with disabilities.
  - Ensure a quantitative analysis or metrics to measure and test the applied definition of fairness.
  - Establish mechanisms to ensure fairness in your AI system.
  - You should ensure that the AI system corresponds to the variety of preferences and abilities in society.
  - You should assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion.

  - You should ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable.
  - You should assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects.
  - You should assess whether there could be groups who might be disproportionately affected by the outcomes of the system.
  - You should assess the risk of the possible unfairness of the system onto the end-user's or subject's communities.

- **Societal and environmental well-being**
  - Consider the potential positive and negative impacts of your AI system on the environment and establish mechanisms to evaluate this impact.
  - Define measures to reduce the environmental impact of your AI system's lifecycle and participate in competitions for the development of AI solutions that tackle this problem.
  - Inform and consult with the impacted workers and their representatives but also involve other stakeholders. Implement communication, education, and training at operational and management level.
  - Take measures to ensure that the work impacts of the AI system are well understood on the basis of an analysis of the work processes and the whole socio-technical system.
  - Provide training opportunities and materials for re- and up-skilling measures.

- **Accountability**
  - Designing a system in a way that can be audited later, results in a more modular and robust system architecture. Thus, it is highly recommended to ensure modularity, traceability of the control and data flow and suitable logging mechanisms.
  - To facilitate 3rd party auditing can contribute to generate trust in the technology and the product itself. Additionally, it is a strong indication of applying due care in the development and adhering to best practices and industrial standards.
  - To foresee 3rd party auditing or guidance can help with both, qualitative and quantitative risk analysis. In addition, it can contribute to generate trust in the technology and the product itself.
  - AI systems should be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. Consequently, developers and deployers should receive appropriate training about the legal framework that applies for the deployed systems.
  - A useful non-technical method to ensure the implementation of trustworthy AI is to include various stakeholders, e.g. assembled in an "ethical review board" to monitor and assist the development process.
  - If AI systems are increasingly used for decision support or for taking decisions themselves, it has to be made sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this. Consequently, all conflicts of values, or trade-offs should be well documented and explained
  - Involving third parties to report on vulnerabilities and risks does help to identify and mitigate potential pitfalls
  - A risk management process should always include new findings since initial assumptions about the likelihood of occurrence for a specific risk might be faulty and thus, the quantitative risk analysis was not correct and should be revised with the new findings.
  - Acknowledging that redress is needed when incorrect predictions can cause adverse impacts to individuals is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.