UNDER REVIEW

RESCIENCE C

Replication / ML Reproducibility Challenge 2023

A Reproducibility Case Study of "Fairness Guarantees under Demographic Shift"

Anonymous Authors

Edited by
(Editor)

Reviewed by
(Reviewer 1)
(Reviewer 2)

Received
03 February 2023

Published
–

DOI
–

Reproducibility Summary

Scope of Reproducibility — This work studies the reproducibility of the paper *Fairness guarantees under demographic shift* (2022) by Giguere et al. Specifically, the authors discuss `Shifty`, an algorithm that provides high-confidence guarantees that a user-specified fairness constraint will hold in the case of a demographic shift between training and deployment data. The authors claim that *Shifty* achieves this without any significant loss of accuracy when compared to a number of other baseline algorithms.

Methodology — Using the open-source code provided by the authors, experiments were conducted to collect the results of `Shifty` and a number of other baseline algorithms when deployed on three different datasets. Results were collected in the form of accuracy, failure rate, and the probability of not finding a fair solution. The experiments in this reproducibility study were conducted on a total of 115 CPU hours.

Results — The claim that `Shifty` guarantees fairness with high confidence is strongly confirmed by the reproduction results of this study. It was also found in this reproducibility study that `Shifty` achieves accuracy scores comparable to those of other fairness algorithms.

What was easy — The open-source code was structured in a way that allowed us to make alterations to the experimental setup or the implementations of the models. The original datasets were also provided in a structured manner and were already standardized.

What was difficult — Modifications to the code were necessary in order to run this code efficiently and without errors; in the original code, there were packages missing, redundant functions and files, and mistakes in the handling of the user-specified fairness constraints.

Communication with original authors — The authors did not respond to our inquiries, resulting in no communication with the original authors.

Copyright © 2023 Anonymous, released under a Creative Commons Attribution 4.0 International license.
Correspondence should be addressed to ()
The authors have declared that no competing interests exists.
Code is available at this link.

1 Introduction

Recent work in the field of AI concerned with bias in machine learning has been focused on creating fair algorithms [2, 3]. However, if demographic shifts occur within the data the model was trained on, such a model can often not maintain its fairness [1]. A demographic shift occurs when certain subgroups in a population are found more or less frequently in deployment. This can result in the model being biased towards certain groups, even if the model was originally trained to be equally fair to all groups [4]. It is therefore crucial to not only ensure the fairness of the model, but to also take into account possible shifts occurring after deployment. This encourages that the models are still making fair and accurate predictions in a possibly *shifted* population.
To train models that are not only fair during training, but also have high-confidence fairness guarantees after deployment, Giguere et al. introduce a new type of learning algorithm called `Shifty`. This algorithm ensures fairness after known and unknown demographic shifts in the data [1]. According to the authors, `Shifty` provides a high probability of the fairness constraints being met when deployed on data that is distributed differently, with respect to a demographic group, than it was during training. In the case of a known shift, the exact demographic shift is known, while with an unknown shift this is approximated by user-specified intervals.
The main contributions of this work are:

- Determining the degree of reproducibility, by recreating the main experiments conducted by Giguere et al., and concluding whether claims made in their research are factual, while considering what resources, such as computational power, are needed to come to this conclusion.
- Improving the efficiency and workings of the code by resolving errors, cleaning up the structure, and removing redundancies.
- Examining the validity of the claims by introducing an unseen dataset with an unknown demographic shift.
- As the authors mention that `Shifty` in practice works with any learning algorithm, the original research is expanded by using a different training algorithm than the one examined in the original paper.

2 Scope of reproducibility

The central claim introduced in this paper is that `Shifty` returns a machine learning algorithm that has a high probability of performing fair classification, before and after deployment, given a user-specified fairness constraint. This constraint is defined to maintain fairness concerning a sensitive attribute such as race or sex. Importantly, `Shifty` does not require access to the deployment data to ensure these guarantees. More specifically, the authors claim the following:

- **High-confidence fairness guarantees:** In the case of a known or unknown demographic shift, `Shifty` provides high-confidence guarantees that a certain user-specified fairness constraint will hold even after the shift.
- **Minor loss of accuracy:** In the case of a known or unknown demographic shift, if enough data is provided, `Shifty` is able to train models, whose resulting accuracy is then comparable to that of models which do not account for demographic shift, such as `Fairness Constraints` [5] and `Seldonian` [2].
- **Finding a solution:** As the amount of training data increases, the probability of `Shifty` returning `No Solution Found` (NSF) decreases.

The remainder of this paper is structured as follows: In section 3, we give a detailed description of `Shifty` and the fairness algorithms it was compared to. In section 4, the methods that were used to verify the claims of the authors are set out. This includes an explanation of the datasets, the hyperparameters, the experimental setup and code, and lastly the computational requirements needed to come to the results. Moreover, we elaborate on the steps taken to conduct additional research. Then, the results of this study are discussed and compared to the original results in section 5. Finally, we conclude this study by discussing and evaluating the used approach.

3 Model descriptions

3.1 `Shifty`

The `Shifty` algorithm is made up of three main parts of which a short overview is given below and which are later discussed more elaborately:

1. The first step consists of partitioning the dataset $D = \{(X_i, Y_i, S_i, T_i)\}_{i=1}^{n}$ into two parts. Each sample in D is uniformly distributed and consists of a feature, label, fairness attribute, and demographic attribute respectively. One part of D is used for candidate model selection, called D_c and the other is used to perform a fairness test, called D_f.

2. Secondly, the candidate model, denoted as θ_c, is trained using D_c, which can be any classification model, while a user-specified fairness constraint is taken into account, also denoted by a function of g. During training, the features, labels, and fairness attributes are used. The candidate model is deemed fair if $g(\theta_c) \leq 0$. This is calculated by inverting Student's t-test [6] and is given by Equation 1, where $\mathbf{E}[H|\xi] - \tau = g(\theta_c)$.

3. Lastly, a high-confidence upper bound (HCUB) on the candidate model is calculated after it is deployed in an environment affected by a demographic shift, which is simulated using the demographic attribute. $g'(\theta_c)$ denotes the fairness of θ_c after deployment and should also be less or equal to zero with high confidence, which then again is calculated using the inverted t-test.

For the third step, the description of the possible demographic shifts is defined by $\mathcal{Q} := \{(a_t, b_t)\}_{t \in \mathcal{T}}$. It is a user-defined set of upper and lower boundaries on the marginal probability of each demographic attribute value after deployment. If the demographic shift is known then $\forall_{t \in \mathcal{T}}[a_t = b_t = Pr(T' = t)]$, where $Pr(T' = t)$ is the probability of demographic attribute t occurring after deployment.

$$Pr(\mathbf{E}[H|\xi] \leq U_{ttest}(g, D, \theta, \delta)) \geq 1 - \delta, \text{ where } \delta \in [0, 1] \tag{1}$$

3.2 Baselines

To assess `Shifty`'s effectiveness, it was compared to other algorithms, namely `Fairness Constraints` [5], `Seldonian`, `Quasi-Seldonian` [2], `Fairlearn` [3], and `RFLearn` [7]. The features of these algorithms are summarized in Table 1. Of these algorithms, `Shifty` is thus the first algorithm to provide fairness guarantees under demographic shift.

3.3 Fairness Constraints

The fairness constraints that `Shifty` uses are user-specified. In this study and in the original paper, two fairness constraints are utilized, namely Demographic Parity (DP)

Algorithm	Classifier	Difference with Shifty
FC	Decision boundary	Does not account for demographic shift
Seldonian	Linear decision boundary	Does not account for demographic shift
Q-Seldonian	Linear decision boundary	Does not account for demographic shift
Fairlearn	Linear SVC	Does not account for demographic shift
RFLearn	Logistic regression	Does not provide fairness guarantees

Table 1. Algorithms overview. For each algorithm, its classifier and its difference with `Shifty` are specified. 'FC' stands for `Fairness Constraints` and 'Q-Seldionian' is the `Quasi-Seldonian` algorithm.

and Disparate Impact (DI). Equations 2 and 3 show the definitions of DP and DI, respectively. In these equations, g represents the function to define unfair behaviour, $\theta(X)$ the model, ϵ the fairness constraint tolerance hyperparameter, and S the sensitive fairness attribute. In the case of S being the attribute `sex`, s_0 and s_1 would represent `male` and `female`.

$$g_{DP} := |\mathbb{E}[\theta(X)|S = s_0] - \mathbb{E}[\theta(X)|S = s_1]| - \epsilon_{DP} \tag{2}$$

$$g_{DI} := -\min(\frac{\mathbb{E}[\theta(X)|S = s_0]}{\mathbb{E}[\theta(X)|S = s_1]}, \frac{\mathbb{E}[\theta(X)|S = s_1]}{\mathbb{E}[\theta(X)|S = s_0]}) + \epsilon_{DI} \tag{3}$$

4 Methodology

We reproduced the original results from the paper using the open-source implementation of the code as provided by the authors on GitHub [8]. This code was analyzed to understand how the results were achieved. The provided code is partly based on the papers that cover the different fairness algorithms that `Shifty` is compared to (section 3.2). A code coverage analysis revealed that a large fraction of the code was not used to obtain the results corresponding to the experiments as discussed in the paper.

4.1 Datasets

To verify the results, this reproducibility study used the same data to conduct the experiments. The authors provided the pre-processed datasets and the code to pre-process the original data. This pre-processing resulted in datasets with zero mean and unit variance. For the non-Seldonian algorithms, both datasets were split up into a 6:4 ratio of train and test data. For the Seldonian algorithms, the ratio was a 6:4 of data for the candidate selection (D_c) and fairness testing (D_f) subsets.
The original paper conducted experiments on the *UCI Adult Census* dataset [9] and the *UFRGS Entrance Exam and GPA* dataset [10]. To further validate the claims made by the authors, an additional dataset was acquired containing approximately 50k diabetic patient encounters collected over a period of 10 years from 130 US hospitals [11]. This dataset will subsequently be referred to as *Diabetes*. Each encounter recorded several statistics to help determine the relationship between the probability of readmission and hbA1c measurement depending on primary diagnosis. For all datasets, we present the relevant statistics and their main purpose in Table 2.

Datasets	Task	Samples	Fairness attr.	Demographic attr.
UCI Adult Census [9]	Predict income above $50k	43k	Race	Sex
UFRGS [10]	Predict GPA above 3.0	43k	Sex	Race
Diabetes [11]	Predict readmission	47k	Race	Sex
		49k	Sex	Race

Table 2. Datasets overview. All three datasets were used for binary classification tasks. In the *UCI Adult Census* and *Diabetes* datasets, the samples were filtered down to only include white and black individuals, when race is considered the fairness attribute.

4.2 Hyperparameters

The original authors' code included a batch file to run the experiments with specified hyperparameters (shown in Table 3 for the *UCI Adult Census* dataset; for *UFRGS* results, see Appendix 7). The provided code contained pre-determined values for the fairness constraint tolerance hyperparameter ϵ (used in equations 2 and 3). It also provided the split ratio for training data and test data, and the split ratio for candidate data and fairness data for the Seldonian algorithms. The batch file also included the number of iterations considered for the training algorithm a, and the confidence bound δ (Eq. 1). Lastly, for the unknown distributional shift the width of the intervals around true marginals representing the valid demographic shifts is given by an α-value.

Constraint	ϵ	train / test	Dc / Df	n-iters	δ	α*
DI	-0.8	0.4	0.4	2000	0.05	0.5
DP	0.1	0.4	0.4	2000	0.05	0.5

Table 3. Hyperparameter values for the experiments run with the *UCI Adult Census* dataset specified for DI or DP for both an unknown demographic shift and a known demographic shift. The α is only used in the case of an unknown shift.

4.3 Experimental setup and code

In the original experiments, each algorithm was trained with different-sized subsets of the data to determine how much data was needed to maintain a fairness guarantee under demographic shift. For a known demographic shift, we used subset sizes ranging from 10k to 60k points, in intervals of 5k datapoints. For an unknown demographic shift, we increased the intervals to 10k while the range remained the same. In a single trial for an user-specified constraint, the results per subset size were collected for each algorithm for both cases of a known and unknown demographic shift.

The original paper specified that 25 trials were executed for each fairness constraint in both cases of a known or unknown demographic shift and for each dataset mentioned in section 4.1. In our experiments, due to a lack of computational resources, we only executed 10 trials for each case.

The original classifier that we used for the three Seldonian algorithms was a linear decision boundary. Additionally, we implemented a multi-layer perceptron (MLP) as a classification model that works with the Seldonian algorithms. The preferred classifier and the sizes of the hidden layers can be specified within the batch file and we used an MLP with 2 hidden layers of sizes 16 and 8. Furthermore, we instantiated the weight parameters according to the normal distribution and optimized these in the same way as the original `Shifty` implementation. This experiment was run for 2 trials for each constraint mentioned previously, for a known and unknown demographic shift.

After every single trial, we saved the results for each algorithm and subset size. These results contained the probability of a `NSF`, the accuracy of the original model and the accuracy of the deployed model, as well as the failure rate of the original model and the failure rate of the deployed model. After completing all the trials, the mean and standard error were determined for each of these measurements. The code of this reproducibility study can be found here.

4.4 Computational requirements

The code provided by the original authors contains an elaborate launcher that allows the experiments to be run with CPU multiprocessing. The CPU used to obtain the results was an *AMD Ryzen 7 3800X* processor, utilizing 8 cores simultaneously.

During the process, the time required per trial was recorded. This was done for each dataset, as well as for each constraint for both the known and unknown demographic

shift. These times were then averaged. For the *UFRGS* dataset this resulted in 15.5 minutes per trial, while the *UCI Adult Census* dataset resulted in 24 minutes per trial. The total (CPU) time required, resulted in 26 hours for the reproduction of the original experiments. The additional experiments entail 28 additional hours for the experiments run with the *Diabetes* dataset, and 60 hours for the experiments run with MLP classifier, concluding to roughly 115 hours total needed for the considered experiments.

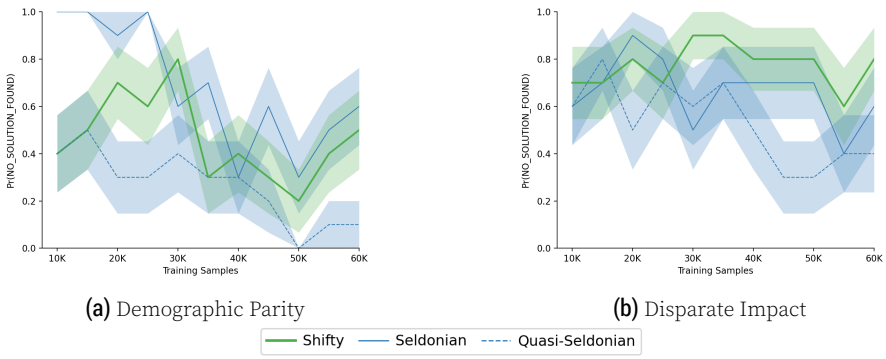
5 Results

In this section, we discuss the reproduction results from the original paper and the results from the additional experiments. To verify the claims set out in section 2, the experiments were executed as described in section 4.3. The results of these experiments include the probabilities that the Seldonian algorithms return `NSF`, and the accuracies and failure rates of the algorithms before and after deployment. The graphs of these results show the means and standard errors. The resulting accuracies and failure rates of the models can be found in Appendix 7.2 & 7.3.
The two fairness constraints, DP and DI, were examined for both datasets. As in the original paper, only the reproduced results of the *UCI Adult Census* dataset are shown. The results of the *UFRGS* dataset can be found in Appendix 7.1.

5.1 Results reproducing original paper

The following section showcases the results of the experiments as discussed in section 4.3. In the sections 5.1.1, and 5.1.2 the specific results for the known and unknown demographic shift are considered, respectively.

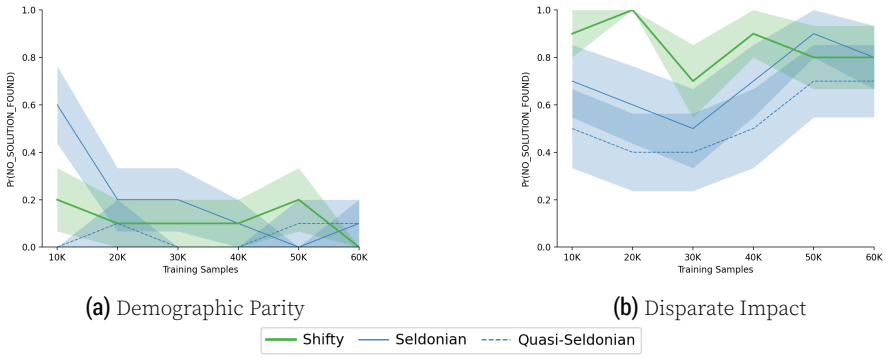
Result 1: Known Demographic Shift — Figure 1 plots the probability that the `Seldonian`, the `Quasi-Seldonian`, and `Shifty` algorithms return `NSF` for a known demographic shift for the fairness constraints DP and DI, per data subset size.



(a) Demographic Parity **(b)** Disparate Impact

Figure 1. Probabilities of returning `NSF` per number of training samples when enforcing fairness constraints DP and DI using the *UCI Adult Census* dataset under known demographic shift.

Result 2: Unknown Demographic Shift — Figure 2 plots the probability that the `Seldonian`, the `Quasi-Seldonian`, and `Shifty` algorithms return `NSF` for an unknown demographic shift for the fairness constraints DP and DI, per data subset size.



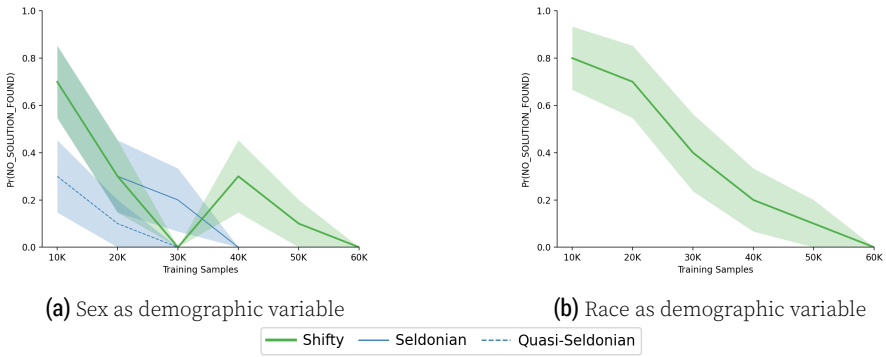
(a) Demographic Parity (b) Disparate Impact

Figure 2. Probabilities of returning NSF per number of training samples when enforcing fairness constraints DP and DI using the *UCI Adult Census* dataset under unknown demographic shift.

5.2 Results beyond original paper

Additional experiments were conducted using a different dataset to substantiate the claims made by the authors. Further experiments were carried out to validate whether `Shifty` is successful independent of the classifier used. In the following section additional results will be discussed, using the *Diabetes* dataset as specified in section 4.1 and the MLP classifier as specified in section 4.3.

Additional Results 1 — Using the *Diabetes* dataset, Figure 3 shows the probability that the `Seldonian`, the `Quasi-Seldonian`, and `Shifty` algorithms return NSF for an unknown demographic shift under the fairness constraint DI, per data subset size. In the left plot *race* is considered the demographic attribute, and in the right plot *sex* is considered the demographic attribute.



(a) Sex as demographic variable (b) Race as demographic variable

Figure 3. Results when enforcing fairness constraint DI using the *Diabetes* dataset under unknown demographic shift, with either the *sex* or *race* as demographic attribute.

Additional Result 2 — For the experiments run with an MLP classifier, the accuracies of the `Seldonian`, the `Quasi-Seldonian`, and `Shifty` algorithms are shown before and after deployment in Table 4, using the *UCI Adult Census* dataset. In case the algorithm was not able to return a fair model in any trial, the results show a NaN-value.

Accuracy - Known Demographic Shift - Demographic Parity											
Samples	10k	15k	20k	25k	30k	35k	40k	45k	50k	55k	60k
Seldonian Classifier											
Original	nan	nan	nan	nan	nan	nan	nan	nan	63.40	62.07	nan
Deployed	nan	nan	nan	nan	nan	nan	nan	nan	63.72	63.29	nan
Difference	nan	nan	nan	nan	nan	nan	nan	nan	0.32	1.22	nan
Quasi-Seldonian-Robust Classifier (`Shifty`)											
Original	nan	nan	nan	nan	53.99	nan	nan	72.87	50.04	73.24	54.11
Deployed	nan	nan	nan	nan	54.86	nan	nan	70.25	47.76	70.77	53.04
Difference	nan	nan	nan	nan	0.87	nan	nan	-2.62	-2.28	-2.47	-1.07
Quasi-Seldonian Classifier											
Original	nan	nan	40.99	62.80	nan	55.11	nan	62.98	68.80	71.24	65.09
Deployed	nan	nan	40.66	60.54	nan	55.71	nan	59.87	68.60	70.81	65.67
Difference	nan	nan	-0.33	-2.26	nan	0.60	nan	-3.11	-0.20	-0.43	0.58

Table 4. Results table showcasing the numerical mean accuracy (in percentages) of each Seldonian algorithm using the MLP-classifier and the *UCI Adult Census* dataset, for both the original and deployment distribution when trained on a known demographic shift with the fairness constraint DP. The decrease or increase in accuracy is shown in the rows named 'difference'.

6 Discussion

6.1 Claim 1: High-confidence fairness guarantees

The results found in this reproducibility study validate the first claim made by the original authors, which asserts the high-confidence fairness guarantee of `Shifty`. The `Shifty` algorithm never returns an unfair model after deployment for both a known and an unknown demographic shift, while other baseline algorithms do. The figures portraying the results of the reproduction experiments supporting this claim can be found in section 7.2 of the appendix.

6.2 Claim 2: Minor loss of accuracy

The results found in this reproducibility study show strong support for the second claim made by the original authors, namely that there is only a minor loss of accuracy with `Shifty` when compared to the other baseline algorithms. This is the case under both a known and an unknown demographic shift, which can be seen in tables 6, 7, 8, and 9 in the appendix section 7.3.
What does stand out is that in the case of DI as the fairness constraint and under a known demographic shift, `Shifty` achieves an accuracy that is approximately 10% lower than that of `RFLearn` and `Fairlearn`.

6.3 Claim 3: Finding a solution

The results found in this reproducibility study do not show strong support for the third claim, namely that `Shifty` avoids returning `NSF` when there is a reasonable amount of training data available. Under a known demographic shift, the probability Pr(`NSF`) shows great fluctuations when altering the number of training samples, and thus showing no support for the third claim. This can be seen in Figures 1a and 1b.
The results under an unknown demographic shift, which can be seen in Figures 2a and 2b, are more stable compared to those under a known demographic shift. There are fewer fluctuations in the probability Pr(`NSF`) when the number of training samples is increased, thus showing more support for the third claim. This is especially the case with the fairness constraint being DP, which even results in `Shifty` having a probability of 0% for 60K training samples. The results with DI as the fairness constraint show more fluctuations, where the probability also increases when more training samples are used.

6.4 Additional Statements

The original paper mentions that the `Shifty` algorithm works with any underlying classification model. However, the additional experiments shown in Table 4 contain `NaN`-values, meaning that `Shifty` accepted candidate models whilst they do not hold the fairness constraints after deployment. From this, we conclude that when a non-linear model is implemented, the guarantee does not necessarily always hold. It is important to note that we only ran 2 trials for each experiment with the MLP classifier.
Additionally, when we ran the experiments on the *Diabetes* dataset, the corresponding results in Figure 3 show that there is no strong evidence for claim 3, which states that larger subsets lead to a lower $Pr(\texttt{NSF})$. While the results in Figure 3b show support for this claim, Figure 3a displays strong fluctuations in the value of $Pr(\texttt{NSF})$ and therefore does not show strong evidence for the claim. However, for the two other claims in both experiments, there are still strong indications that they are kept.

6.5 What was easy

Since the repository containing all the code used to run the experiments was made available by the authors of the original paper, nothing needed to be implemented from scratch. This also provided the tuned hyperparameters, resulting in no extra time needed to search for these. The pre-processed data was also supplied, avoiding any extra time needed to match these to the implementation.

6.6 What was difficult

The code required a thorough analysis to determine its functioning and redundant parts, and while the authors provided a file containing all the necessary requirements, multiple modules were not included. Furthermore, debugging was necessary to be able to run the provided set-up successfully, since the original code contained mistakes in handling the fairness constraint expressions and loading the data.
The original paper did not present its results in numerical values but rather only showed graphs. This made it complicated to fully validate whether the reproduced results approximate these, and thus to fully support the claims. Additionally, discrepancies were found between the number of trials conducted according to the published code and the amount mentioned in the paper. While the paper indicates to have run 25 trials for each algorithm with each constraint per dataset, the code showed a lower number for the experiments with an unknown demographic shift.

6.7 Communication with original authors

The original authors did not respond to our inquiry, so there was no communication. It would have been useful to have received the numerical values of the figures in the original paper, so that a quantitative comparison of the values could be performed.

References

1. S. Giguere, B. Metevier, Y. Brun, P. S. Thomas, S. Niekum, and B. C. da Silva. "Fairness Guarantees under Demographic Shift." In: **International Conference on Learning Representations**. 2022.
2. P. S. Thomas, B. Castro da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. "Preventing undesirable behavior of intelligent machines." In: **Science** 366.6468 (2019), pp. 999–1004.
3. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. "A reductions approach to fair classification." In: **International Conference on Machine Learning**. PMLR. 2018, pp. 60–69.
4. J. Schrouff, N. Harris, O. Koyejo, I. Alabdulmohsin, E. Schnider, K. Opsahl-Ong, A. Brown, S. Roy, D. Mincu, C. Chen, et al. "Maintaining fairness across distribution shift: do we have viable solutions for real-world applications?" In: **arXiv preprint arXiv:2202.01034** (2022).
5. M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. "Fairness constraints: Mechanisms for fair classification." In: **Artificial intelligence and statistics**. PMLR. 2017, pp. 962–970.
6. Student. "The Probable Error of a Mean." In: **Biometrika** 6.1 (1908), pp. 1–25.
7. W. Du and X. Wu. "Fair and robust classification under sample selection bias." In: **Proceedings of the 30th ACM International Conference on Information & Knowledge Management**. 2021, pp. 2999–3003.
8. S. Giguere, B. Metevier, Y. Brun, P. S. Thomas, S. Niekum, and B. C. da Silva. **Fairness Guarantees under Demographic Shift**. https://github.com/sgiguere/Fairness-Guarantees-under-Demographic-Shift. 2022.
9. R. Kohavi and B. Becker. **UCI Machine Learning Repository: Adult Data Set**. 1996.
10. B. C. da Silva. **UFRGS Entrance Exam and GPA Data**. Version V2. 2019.
11. B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore. "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records." In: **BioMed research international** 2014 (2014).

7 Appendix

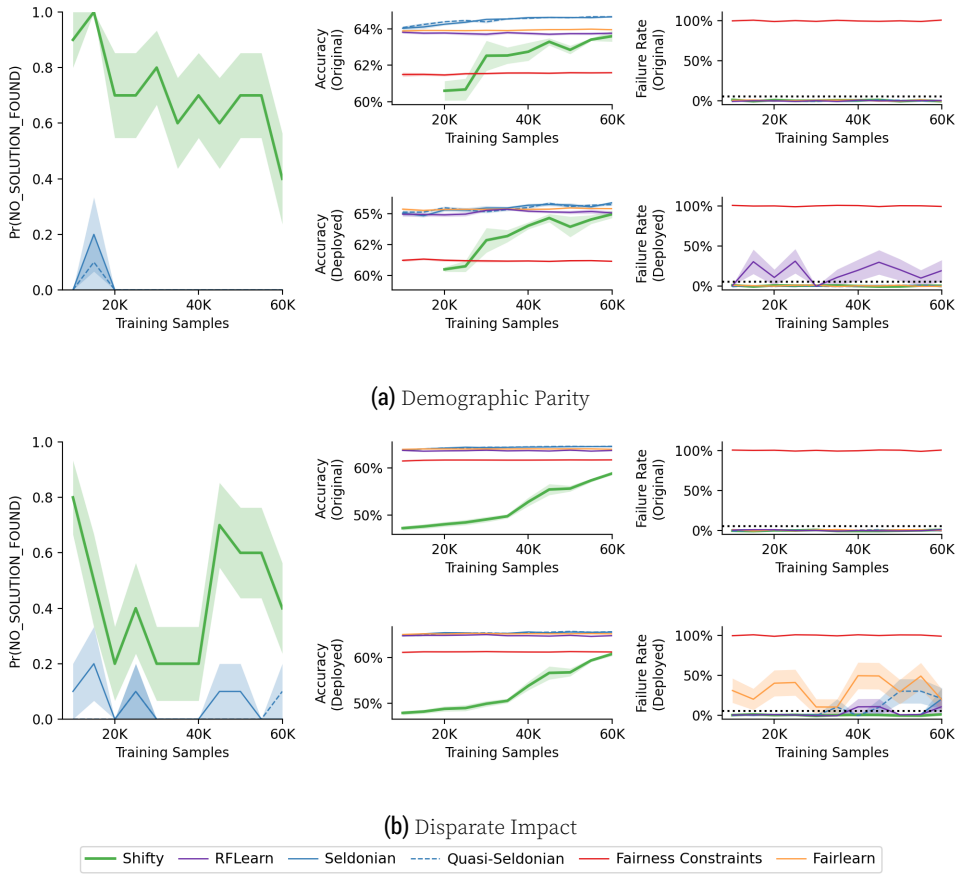
7.1 UFRGS Entrance Exam and GPA dataset

The following section contains the results considering the second dataset used for reproduction of the original experiments as described in section 4.1 and 4.3. In this dataset sex is used as the fairness attribute, and race as the demographic attribute.
The hyperparameters values used for these experiments are summarised in Table 5.

Constraint	ϵ	train / test	Dc / Df	n-iters	δ	α^*
DI	-0.8	0.4	0.4	2000	0.05	0.25
DP	0.1	0.4	0.4	2000	0.05	0.25

Table 5. Hyperparameter values for the experiments run with the *UFRGS GPA* dataset specified for Disparate Impact (DI) or Demographic Parity (DP) for both a known and unknown demographic shift. α is only used in the case of an unknown shift.

Shown in figure 4 are the results under a known demographic shift with DP and DI as the fairness constraints. Results under unknown demographic shift considering DP and DI as the fairness constraints are shown in figure 5.



(a) Demographic Parity

(b) Disparate Impact

Figure 4. Results when enforcing fairness constraints under known demographic shift using the *UFRGS GPA* dataset. For both fairness constraints DP and DI, the leftmost graph shows the probability of `NO_SOLUTION_FOUND`, the middle column shows the accuracies, and the rightmost column shows the failure rates.

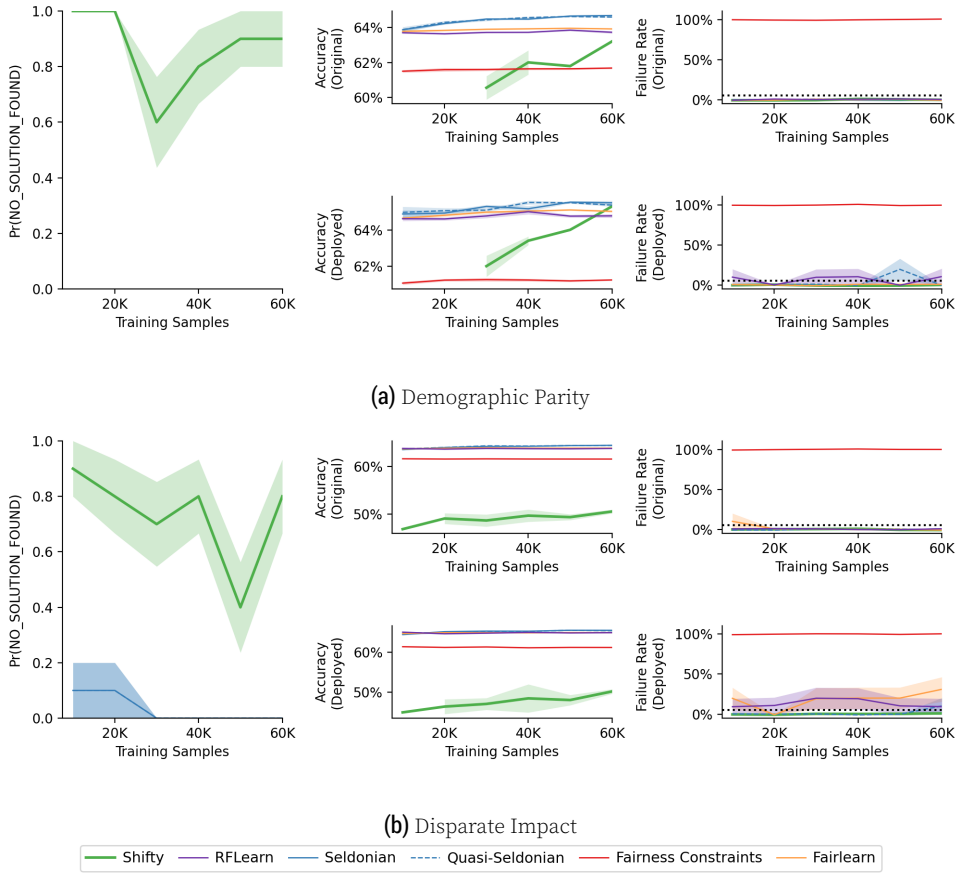


Figure 5. Results when enforcing fairness constraints under unknown demographic shift using the *UFRGS GPA* dataset. For both fairness constraints DP and DI, the leftmost graph shows the probability of `NO_SOLUTION_FOUND`, the middle column shows the accuracies, and the rightmost column shows the failure rates.

7.2 UCI - Adult Census Failure Rates

In this section, the failure rates for each algorithm with the *UCI Adult Census* dataset, as mentioned in section 6.1, are provided in Figures 6 and 7.

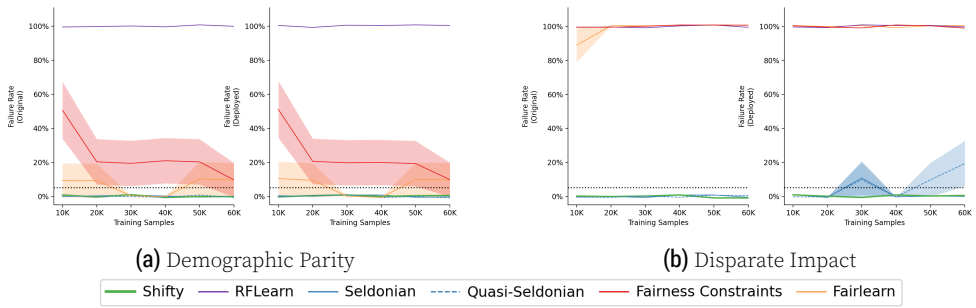


Figure 6. Failure rates for each algorithm under unknown demographic shift for fairness constraints DP and DI with the *UCI Adult Census* dataset. The confidence bound is indicated with the dotted line.

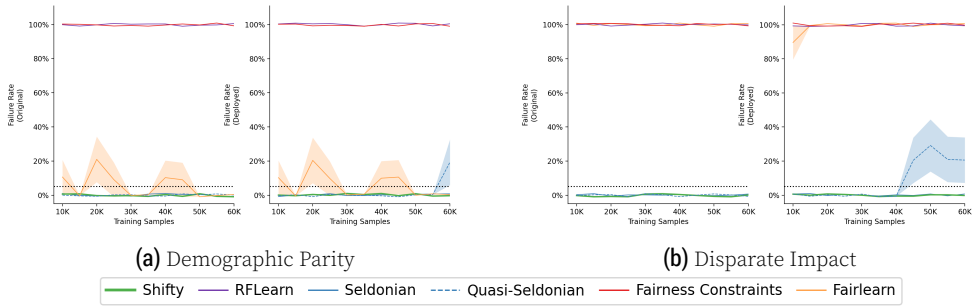


Figure 7. Failure rate (in percentages) for each algorithm under known demographic shift for fairness constraints DP and DI with the *UCI Adult Census* dataset. The confidence threshold is indicated with the dotted line.

7.3 Numerical Results

The tables in this section (tables 6, 7, 8, and 9) provide numerical results of the accuracy scores on the experiments run with the *UCI Adult Census* dataset, as set out in section 6.2.

Accuracy - Known Demographic Shift - Demographic Parity											
Samples	10k	15k	20k	25k	30k	35k	40k	45k	50k	55k	60k
Seldonian Classifier											
Original	nan	nan	76.05	nan	76.31	76.59	76.59	77.75	78.05	78.11	78.19
Deployed	nan	nan	72.80	nan	74.15	75.30	74.72	75.58	75.98	76.06	76.05
Difference	nan	nan	-3.25	nan	-2.16	-1.29	-1.87	-2.17	-2.07	-2.05	-2.14
Quasi-Seldonian-Robust Classifier (`Shifty`)											
Original	76.75	76.25	77.17	77.56	77.34	77.80	77.91	77.99	77.99	77.93	77.95
Deployed	73.60	74.11	74.37	74.99	74.66	75.19	75.62	75.44	75.53	75.72	75.73
Difference	-3.15	-2.14	-2.80	-2.57	-2.68	-2.61	-2.29	-2.55	-2.46	-2.21	-2.22
Quasi-Seldonian Classifier											
Original	76.70	77.03	77.38	78.10	78.31	78.56	78.20	78.56	78.68	78.81	78.86
Deployed	73.47	74.35	74.56	75.70	75.96	76.60	75.66	76.43	76.42	76.67	76.69
Difference	-3.23	-2.68	-2.82	-2.40	-2.35	-1.96	-2.54	-2.13	-2.26	-2.14	-2.17
Fairlearn											
Original	74.03	74.39	74.78	74.11	74.36	74.40	74.13	74.15	74.41	74.40	74.43
Deployed	70.91	71.24	71.76	71.04	71.23	71.25	71.05	71.06	71.25	71.25	71.26
Difference	-3.12	-3.15	-3.02	-3.07	-3.13	-3.15	-3.08	-3.09	-3.16	-3.15	-3.17
RFLearn											
Original	80.99	81.03	81.01	80.92	81.05	81.00	81.03	81.04	81.12	81.02	80.98
Deployed	78.51	78.55	78.56	78.46	78.56	78.52	78.58	78.59	78.68	78.59	78.53
Difference	-2.48	-2.48	-2.45	-2.46	-2.49	-2.48	-2.45	-2.45	-2.44	-2.43	-2.45
Fair Constraints											
Original	80.99	80.77	80.71	80.63	80.60	80.58	80.64	80.49	80.51	80.52	80.52
Deployed	78.67	78.44	78.38	78.28	78.23	78.23	78.29	78.15	78.19	78.17	78.16
Difference	-2.32	-2.33	-2.33	-2.35	-2.37	-2.35	-2.35	-2.34	-2.32	-2.35	-2.36

Table 6. Results table showcasing the numerical mean accuracy percentage of each algorithm, for both the original distribution and the deployed one when trained under a known demographic shift with fairness constraint Demographic Parity. The decrease or increase in accuracy is shown in the rows named 'Difference'.

Accuracy - Known Demographic Shift - Disparate Impact											
Samples	10k	15k	20k	25k	30k	35k	40k	45k	50k	55k	60k
Seldonian Classifier											
Original	65.00	69.93	73.23	73.60	74.40	74.96	75.30	76.72	77.01	77.01	77.65
Deployed	65.65	69.75	71.53	72.22	73.32	73.57	73.95	75.25	75.60	75.52	76.01
Difference	0.65	-0.18	-1.70	-1.38	-1.08	-1.39	-1.35	-1.47	-1.41	-1.49	-1.64
Quasi-Seldonian-Robust Classifier (`Shifty`)											
Original	67.31	68.49	71.50	74.14	74.58	76.98	77.01	76.90	77.03	77.32	77.29
Deployed	67.30	66.93	69.81	72.72	72.79	75.45	75.49	74.69	74.70	75.16	75.33
Difference	-0.01	-1.56	-1.69	-1.42	-1.79	-1.53	-1.52	-2.21	-2.33	-2.16	-1.96
Quasi-Seldonian Classifier											
Original	66.72	71.57	72.95	76.02	75.96	76.90	77.65	77.69	78.20	78.42	78.49
Deployed	66.92	69.32	72.21	74.55	74.05	74.91	75.77	75.80	76.38	76.87	76.70
Difference	0.20	-2.25	-0.74	-1.47	-1.91	-1.99	-1.88	-1.89	-1.82	-1.55	-1.79
Fairlearn											
Original	80.10	79.77	79.94	80.60	80.58	80.68	80.61	80.68	80.68	80.68	80.66
Deployed	77.54	77.20	77.35	78.09	78.12	78.16	78.11	78.15	78.15	78.13	78.12
Difference	-2.56	-2.57	-2.59	-2.51	-2.46	-2.52	-2.50	-2.53	-2.53	-2.55	-2.54
RFLearn											
Original	81.00	81.05	81.07	81.08	81.01	81.00	80.98	81.07	80.97	81.03	80.99
Deployed	78.58	78.64	78.64	78.67	78.58	78.54	78.54	78.64	78.53	78.61	78.54
Difference	-2.42	-2.41	-2.43	-2.41	-2.43	-2.46	-2.44	-2.43	-2.44	-2.42	-2.45
Fair Constraints											
Original	81.04	81.14	80.66	80.61	80.64	80.64	80.69	80.61	80.63	80.54	80.58
Deployed	78.70	78.81	78.33	78.27	78.28	78.27	78.34	78.25	78.27	78.16	78.21
Difference	-2.34	-2.33	-2.33	-2.34	-2.36	-2.37	-2.35	-2.36	-2.36	-2.38	-2.37

Table 7. Results table showcasing the numerical mean accuracy percentage of each algorithm, for both the original distribution and the deployed one when trained under a known demographic shift with fairness constraint Disparate Impact. The decrease or increase in accuracy are shown per number of samples in the rows named 'Difference'

Accuracy - Unknown Demographic Shift - Demographic Parity						
Samples	10k	20k	30k	40k	50k	60k
Seldonian Classifier						
Original	77.23	78.74	79.40	79.54	79.75	80.08
Deployed	76.31	77.06	77.49	77.67	77.93	78.03
Difference	-0.92	-1.68	-1.91	-1.87	-1.82	-2.05
Quasi-Seldonian-Robust Classifier (`Shifty`)						
Original	78.39	79.23	79.32	79.65	79.66	79.83
Deployed	76.95	77.29	77.39	77.61	77.71	77.88
Difference	-1.44	-1.94	-1.93	-2.04	-1.95	-1.95
Quasi-Seldonian Classifier						
Original	78.85	79.70	80.28	80.49	80.61	80.68
Deployed	76.65	77.68	78.16	78.29	78.36	78.46
Difference	-2.20	-2.02	-2.12	-2.20	-2.25	-2.22
Fairlearn						
Original	75.03	75.03	74.38	74.40	75.07	75.05
Deployed	72.17	72.17	71.47	71.47	72.19	72.19
Difference	-2.86	-2.86	-2.91	-2.93	-2.88	-2.86
RFLearn						
Original	80.97	80.92	80.94	80.86	80.86	80.89
Deployed	78.73	78.66	78.70	78.61	78.61	78.62
Difference	-2.24	-2.26	-2.24	-2.25	-2.25	-2.27
Fair Constraints						
Original	81.02	80.67	80.69	80.63	80.60	80.58
Deployed	78.84	78.47	78.51	78.46	78.42	78.41
Difference	-2.18	-2.20	-2.18	-2.17	-2.18	-2.17

Table 8. Results table showcasing the numerical mean accuracy percentage of each algorithm, for both the original distribution and the deployed one when trained under an unknown demographic shift with fairness constraint Demographic Parity. The decrease or increase in accuracy is shown in the rows named 'Difference'.

Accuracy - Unknown Demographic Shift - Disparate Impact						
Samples	10k	20k	30k	40k	50k	60k
Seldonian Classifier						
Original	64.66	71.92	75.22	75.74	77.17	76.74
Deployed	65.43	71.93	75.63	76.54	80.18	79.55
Difference	0.77	0.01	0.41	0.80	3.01	2.81
Quasi-Seldonian-Robust Classifier (`Shifty`)						
Original	52.97	nan	73.30	75.09	75.61	76.39
Deployed	52.36	nan	73.20	74.29	74.40	74.71
Difference	-0.61	nan	-0.10	-0.80	-1.21	-1.68
Quasi-Seldonian Classifier						
Original	67.03	74.15	76.09	77.71	77.41	78.17
Deployed	67.31	75.56	76.82	79.09	79.14	78.73
Difference	0.28	1.41	0.73	1.38	1.73	0.56
Fairlearn						
Original	80.01	80.00	80.68	80.68	80.68	80.69
Deployed	82.78	82.78	84.24	84.34	84.15	84.54
Difference	2.77	2.78	3.56	3.66	3.47	3.85
RFLearn						
Original	81.00	80.95	80.98	80.96	81.03	80.98
Deployed	83.06	83.26	82.14	82.60	82.00	82.69
Difference	2.06	2.31	1.16	1.64	0.97	1.71
Fair Constraints						
Original	80.97	80.65	80.64	80.55	80.65	80.52
Deployed	81.78	80.00	79.97	79.15	79.56	79.37
Difference	0.81	-0.65	-0.67	-1.40	-1.09	-1.15

Table 9. Results table showcasing the numerical mean accuracy percentage of each algorithm, for both the original distribution and the deployed one when trained under an unknown demographic shift with fairness constraint Disparate Impact. The decrease or increase in accuracy is shown in the rows named 'Difference'.