UNDER REVIEW

RESCIENCE C

Replication / ML Reproducibility Challenge 2022

Reproducibility Study of 'Latent Space Smoothing for Individually Fair Representations'

Anonymous[1, ID]

[1] Anonymous Institution

Edited by
(Editor)

Reviewed by
(Reviewer 1)
(Reviewer 2)

Received
03 February 2023

Published
–

DOI
–

Reproducibility Summary

Scope of Reproducibility — The aim of this work is to study the reproducibility of the paper *'Latent Space Smoothing for Individually Fair Representations'* by Peychev et al., in which a novel representation learning method called LASSI is proposed. We aim to verify the three main claims made in the original paper: (1) LASSI increases certified individual fairness, while keeping prediction accuracies high, (2) LASSI can handle various sensitive attributes and attribute vectors and (3) LASSI representations can achieve high certified individual fairness even when downstream tasks are not known. In addition, we aim to test the robustness of their claims by conducting additional experiments.

Methodology — To reproduce the experiments, we use the step-by-step guidelines supplied by the original authors on their github repository. The experimental setup and datasets used are identical to the work reported in the original paper. We write additional code to run experiments beyond the scope of the work done by Peychev et al. In order to comply with resource limitations, we reproduce only the experiments relevant to the main claims. In total a budget of 45 hours on an NVIDIA Titan RTX GPU is used.

Results — We are able to reproduce and verify the three main claims of the original paper, by reproducing the results within 5% of the reported values. The additional experiments were succesful and strengthen the claims that LASSI increases certified individual fairness compared to the baseline models. Outliers of the experiments are studied and found to be caused by biased and inaccurate input data.

What was easy — Reproducing the original experiments was made possible by the extensive documentation and guidelines created by the authors in their code and public GitHub repository. The theoretical background provided in their paper was clear and detailed, allowing a deeper understanding about the inner workings of their models and metrics.

What was difficult — The main difficulty was found within the complex structure of the original code files and the related functions across these files. The code needed to perform our additional experiments was therefore also complex and required us to alter many different functions in the original code.

Communication with original authors — To keep the reproducibility report a fair assessment, this work has been sent to the original authors to ask for their feedback and comments.

Copyright © 2023 Anonymous, released under a Creative Commons Attribution 4.0 International license.
The authors have declared that no competing interests exists.
Code is available at https://mametchiii.github.io/lassi-reproducibility/.

1 Introduction

In critical domains such as loan applications [1], crime risk assessments [2] and human resources [3], decisions are increasingly being made by deep learning models. The decisions made by these data-driven models can have wide-ranging impacts and consequences on individuals and society as a whole. Recent studies, however, found that these models and datasets can be biased [4, 5], resulting in discrimination based on sensitive attributes such as race or gender [6, 7, 8]. The field of fairness in artificial intelligence attempts to reduce the biases in decision-making algorithms to ensure a fair treatment of groups and individuals.

In order to ensure that similar individuals are treated similarly Peychev et al. [9] propose LASSI: a novel representation learning method that is able to certify individual fairness on high-dimensional data. This is done by using recent advances in generative models [10] and the scalable certification of deep models [11]. On multiple image classification tasks, the authors claim that LASSI increases certified individual fairness compared to the baselines, while keeping prediction accuracies high. In addition the authors claim that through transfer learning, the representations obtained by LASSI can be used to solve tasks that were unseen during the training of the model.

Our contributions — In this paper we aim to reproduce the results and verify the claims presented in the original paper by Peychev et al. [9] In addition, we aim to extend their research by performing additional experiments to validate the robustness of their claims and investigate the encountered outliers.

2 Scope of reproducibility

Adapting individual fairness and providing similar decisions for similar individuals in machine learning algorithms has proven to be difficult [12]. This is mainly due to the subjectivity and high domain dependence of such a similarity metric [13]. In their paper Peychev et al. [9] present a novel input similarity metric, together with LASSI: a representation learning method with certified individual fairness.

The main goal of this reproducibility studies is to reproduce and verify the following three main claims made by Peychev et al. [9]:

- **Claim 1:** LASSI significantly increases certified individual fairness compared to the naive baseline model, while keeping prediction accuracies high.
- **Claim 2:** LASSI can handle various sensitive attributes and attribute vectors and increase certified individual fairness compared to the naive baseline model.
- **Claim 3:** LASSI representations transfer to unseen tasks and can still achieve high certified individual fairness when the downstream tasks are not known.

We extend the verification of these claims by executing additional experiments, testing the robustness of the claims, and taking a deeper dive into possible outliers of the model.

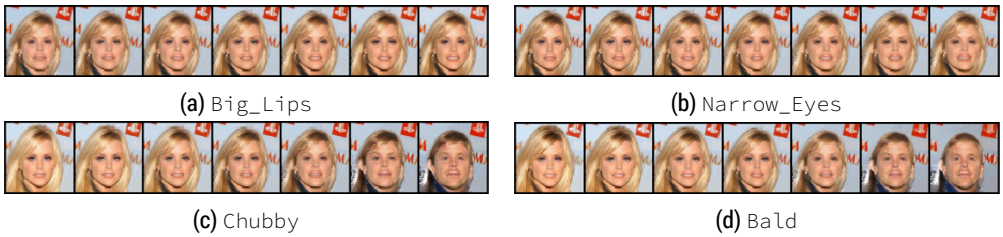
In Section 3, a short theoretical background on LASSI is given, combined with a detailed methodology of the reproducibility studies. In Section 4.1 and 4.2 we present the results of the reproduced experiments and our own contributions respectively. To conclude, in Section 5 we discuss the results and workflow of this research.

3 Methodology

In this section we will describe the two models that are used in the fair representation learning method proposed by Peychev et al. [9]: GLOW and LASSI. The datasets used for training and evaluating will be explained and a definition of fairness will be covered. To conclude the methodology, there will be a description of our experimental set up, together with the computational requirements.

3.1 Model descriptions

In order to ensure fairness, similar individuals that only differ in one or more sensitive attributes such as race or age, need to be treated similar by the LASSI model. To do this, we want to ignore these sensitive attributes in the classification process. This is achieved by using the generative model GLOW [10], which allows us to alter the input data in the latent space, along a specific attribute vector. The images generated by GLOW, shown in figure 1, contain faces that only differ in one or more sensitive attributes and are treated similarly during the training process of the LASSI model.



(a) `Big_Lips` (b) `Narrow_Eyes` (c) `Chubby` (d) `Bald`

Figure 1. Visualizations of the input generated by the GLOW model for a face in the CelebA dataset. The attributes in these figures are the sensitive attributes we use in our additional experiments.

The training process involves balancing fairness, accuracy, and transferability to unknown tasks by finding the optimal value of different loss functions. Once the fair representation of the data is learned, it can be used to train a classifier for any downstream task. The method is compared to a fairness-unaware (naive) baseline model for evaluation. For a more detailed explanation about the models, see Appendix Section A.

3.2 Datasets

This reproducibility research focusses on two datasets used in the original paper. The first is the **CelebA** [14] dataset, which contains 202,599 images of faces of real-world celebrities and is annotated with 40 features. The other dataset used is the **FairFace** [15] dataset, which contains 97,698 images of faces annotated with their race, age and gender. As opposed to the CelebA dataset, the FairFace dataset is balanced, meaning that every race is equally represented. More information about the two datasets is given in appendix Section B.

3.3 Metrics

The LASSI model is evaluated using two metrics: accuracy and fairness. Accuracy is calculated by dividing the amount of correct predictions by the total amount of predictions that has been made.

Fairness — The fairness metric for high dimensional data is a key contribution of the original paper, as described in Section 2. To calculate this metric, the following definition

is given: A model $M : \mathbb{R}^n \to Y$ is individually fair at $\mathbf{x} \subset \mathbb{R}^n$ if it classifies all individuals similar to $\mathbf{x}$ the same [16].

Metric of similarity — As described in Section 3.1 images should be treated similarly, when they only differ in the direction of a certain attribute vector within the latent space.

To develop this similarity metric, center smoothing is applied to the representation of each input image and its similarity set generated by GLOW, in order to bound the distance between these representations by a radius, d_{cs}. The classifier is also randomly smoothed to obtain its l_2 radius, d_{rs}. If d_{cs} is less than d_{rs}, the model provably classifies similar images in the same manner, which is considered as certified individual fairness for an image. The overall fairness of the model is then calculated as the percentage of images that have been certified as 'fair' predictions.

3.4 Experimental setup and hyperparameters

To reproduce the original results, the guidelines explained by the original authors in their GitHub repository [9] are followed. Due to the usage of a Windows machine, the shell-files are executed manually up until the training of the LASSI model, which is done on a Linux machine.

Because of the limited GPU capacity and budget, we do not reproduce all results from the original paper and discard the data-augmentation model, which serves as a model between the naive baseline model and LASSI. The reproduced experiments that are discarded are those deemed least relevant to contribute to a final conclusion. For the CelebA dataset, the model was trained on the tasks `Smiling` and `Earrings` using the sensitive attributes `Pale_Skin`, `Young`, `Blond_Hair` and their combinations. For the FairFace dataset, the model was trained on the tasks `Age-2`, which aims to predict if an individual is younger or older than 30 and `Age-3`, which has three target age ranges: [0-19], [20-39] and [40+].

Additional attributes — To test the robustness of the model, the performance of LASSI was evaluated on additional sensitive attributes and tasks not included in the original work. These included `Bald`, `Big_Lips`, `Chubby`, `Narrow_Eyes` for the CelebA dataset and `Race=Indian` for the FairFace dataset. The visualizations of these can be seen in figure 1 and in a larger size in the Appendix, Section D.

Additional Tasks — The additional tasks we trained the model on are Wearing_Hat, Attractive and Wearing_Necklace for the CelebA dataset.

Hyperparameters — To decrease the run time of all experiments, we run the experiments using two different random seeds, as opposed to the five random seeds used by Peychev et al. [9]. The other hyperparameters were identical to the parameters used by the original authors. The full details and the code of our reproducibility research can be found on our dedicated GitHub page (https://mametchiii.github.io/lassi-reproducibility/).

3.5 Computational requirements

To reduce the training time, we cache the image representations in the latent space of the generative models. This is done with a build-in GPU in series NVIDIA Quadro P1000 which combines a 640 CUDA core Pascal GPU and a 4 GB GDDR5 on-board memory. The total run time of caching the data takes approximately 15 hours. The experiments of training the LASSI model are executed on a LISA cluster with an NVIDIA Titan RTX with a total budget of 20 hours per job and 3233 SBUs compute units.

4 Results

In this section we report the results of the reproducibility studies. These results are two-fold: in Section 4.1 we present minimal differences between the reproduced experiments and the original research by Peychev et al. [10], supporting the three main claims. In Section 4.2 experiments beyond the original research demonstrate that the claims made about the LASSI model are generally robust and possible flaws are explained.

4.1 Results reproducing original paper

Using the code from Peychev et al. [9] we are able to reproduce the experiments exploring the three main claims of the original paper. We present the results of these reproduced experiments claim by claim in the following subsections.

Definition — *All reproduced results within a 5% range of the original results are considered to be 'similar' and displayed in green; values outside this range are 'dissimilar' and shown in red.*

Claim 1 — The first main claim of the original paper states that LASSI significantly increases certified individual fairness, while keeping prediction accuracies high. To verify this claim, we reproduce the experiments by evaluating the performance of the baseline naive model and the LASSI model on two different datasets and multiple tasks.

In table 1 the reproduced performance of both the naive and LASSI model on two datasets is shown, together with the corresponding values found by Peychev et al. [10] in *italics*. As explained in Section 3.4, our results are averaged over two runs with random seeds, as opposed to the five runs with random seeds used in the original research. We measure a similar performance compared to the original paper.

These results indicate that the LASSI model significantly improves certified fairness compared to the naive model, with only a minor loss in accuracy on the `Smiling` task. It even acts as a regularizer on the imbalanced `Earrings` task, where an improved accuracy is measured.

			Naive model		LASSI model	
Dataset	Task	Sensitive attrib.	Acc	Fair	Acc	Fair
CelebA	Smiling	Pale_Skin	**85.4** | ***86.3***	0.3 | *0.6*	84.9 | *85.9*	**97.3** | ***98.0***
		Young	**85.3** | ***86.3***	54.8 | *38.2*	85.1 | ***86.3***	**98.6** | ***98.8***
		Blond_Hair	85.6 | *86.3*	5.6 | *3.4*	**86.4** | ***86.4***	**97.0** | ***94.7***
		Pale+Young	85.1 | ***86.0***	0.3 | *0.4*	**85.3** | *85.8*	**97.4** | ***97.3***
		P + Y + B	85.3 | ***86.2***	0.0 | *0.0*	**85.7** | *85.5*	**91.8** | ***86.5***
	Earrings	Pale_Skin	83.3 | *81.3*	10.6 | *24.3*	**85.7** | ***85.5***	**97.0** | ***98.5***
		Young	83.3 | *81.4*	30.3 | *59.2*	**86.7** | ***84.5***	**99.0** | ***98.0***
		Blond_Hair	83.3 | *81.4*	4.3 | *9.2*	**86.4** | ***84.8***	**97.4** | ***96.2***
FairFace	Age-2	Race=Black	66.1 | *69.0*	5.5 | *5.7*	**70.8** | ***72.0***	**95.3** | ***95.0***
	Age-3	Race=Black	64.1 | ***67.0***	0.0 | *0.0*	**64.6** | *65.1*	**93.4** | ***90.8***

Table 1. Evaluation of the Naive and LASSI models on the CelebA and FairFace datasets. The results are reported as 'our results | *original results [9]*'. Highlighted in bold are the highest accuracy and fairness between the naive and LASSI model. The reproduced values that are similar to the original values ($\Delta \leq 5\%$) are marked in green, the dissimilar values in red.

Claim 2 — The second claim made by Peychev et al. [9] states that LASSI can correctly handle various sensitive attributes and attribute vectors.

The first part of this claim is supported by the results in table 1. These results indicate that LASSI increases the certified individual fairness using multiple different sensitive attributes: `Pale_Skin`, `Young`, `Blond_Hair` and combinations of two or more of these sensitive attributes. In addition LASSI keeps the prediction accuracies high, and even increases them for unbalanced tasks.

To examine whether LASSI is independent of the computation of the attribute vector a, we evaluate the performance of the LASSI model by using two different attribute vector types. In table 2 we show that the results of the reproduced experiments are similar to the values found in the original paper, supporting the claim that LASSI correctly handles various different attribute vector types.

		Naive model		LASSI model	
a-vector type	Sensitive attrib.	Acc	Fair	Acc	Fair
`orthogonal`	`Pale_Skin`	**85.3** | *86.4*	57.5 | *34.0*	**85.3** | *86.5*	**98.4** | *98.8*
	`Young`	**85.3** | *86.3*	74.5 | *73.1*	84.9 | *86.8*	**98.6** | *97.9*
	`Blond_Hair`	**85.3** | *86.2*	76.9 | *71.4*	84.6 | *86.7*	**97.4** | *98.8*
`sample avg`	`Blond_Hair`	**85.3** | *86.2*	87.8 | *90.8*	85.1 | *86.8*	**98.1** | *98.8*

Table 2. Evaluation of the Naive and LASSI models on the CelebA dataset using two different attribute vectors. The results are reported as 'our results | *original results [9]*'. Highlighted in bold are the highest accuracy and fairness between the naive and LASSI model. The reproduced values that are similar to the original values ($\Delta \leq 5\%$) are marked in green.

Claim 3 — The third main claim made in the original paper is that LASSI can learn transferable representations and still achieve high certified individual fairness, also when the downstream tasks are not known. To examine this, consistent with prior work [55] and similar to the original research, we turn off the classification loss and enable the reconstruction loss.

In table 3 we report the performance of the LASSI model on the downstream tasks `Smiling` and `High_Cheeks`, using `Pale_Skin` and `Young` as the sensitive attributes. Our reproduced results are similar to the results from Peychev et al. [10], indicating that the models perform slightly worse than when the tasks are known, but still maintaining high individual fairness. The claim that LASSI can achieve high certified individual fairness even when the downstream tasks are not known is supported by these results.

Sensitive attrib.:	`Pale_Skin`		`Young`	
Transfer task	Acc	Fair	Acc	Fair
`Smiling`	82.1 | *86.2*	96.3 | *93.1*	85.7 | *86.0*	96.2 | *95.4*
`High_Cheeks`	79.6 | *81.7*	96.2 | *92.6*	81.2 | *82.3*	97.4 | *96.0*

Table 3. Evaluation of the accuracy and fairness of LASSI when the downstream tasks are not known, using transfer learning. The results are reported as 'our results | *original results [9]*'. The reproduced values that are similar to the original values ($\Delta \leq 5\%$) are marked in green.

4.2 Results beyond original paper

Robustness of LASSI — In order to assess the robustness of LASSI, we conducted additional experiments using the CelebA and FairFace datasets. The scope of the experiments was expanded to include a wider range of sensitive attributes and tasks. These experiments serve to further complete the experimental setup presented in the original paper by incorporating nearly all relevant options for sensitive attributes and tasks.

The results of the experiments are reported in Table 4 and 5. Table 4 shows that LASSI increases fairness scores on all examined sensitive attributes, while maintaining high prediction accuracies. The results in this table are similar to those presented earlier in table 1, supporting the robustness of claim 1 and 2. Table 5 shows that LASSI achieves high individual fairness on two additional transfer tasks, further strengthening claim 3.

Two surprising values to note are the low individual fairness scores for the LASSI model on the `Attractive` task with the sensitive attributes `Bald` and `Chubby`, highlighted in red in table 4. This decrease, however, does not necessarily compromise the robustness of the model. Further investigation of these outliers will follow in Section 4.2.2 and we will discuss the consequent results in Section 5 and Appendix Section C.

			Naive		LASSI	
Dataset	Task	Sensitive attribute	Acc	Fair	Acc	Fair
CelebA	Smiling	Bald	85.3	42.6	**85.4**	**96.0**
		Big_Lips	**85.3**	77.7	**85.3**	**99.5**
		Chubby	85.3	38.1	**85.7**	**98.4**
		Narrow_Eyes	85.3	9.0	**87.2**	**98.4**
	Wearing_Hat	Bald	96.0	31.9	**97.9**	**99.8**
		Big_Lips	96.2	98.1	**97.4**	**100.0**
		Chubby	96.0	65.2	**98.1**	**99.7**
		Narrow_Eyes	96.2	98.8	**97.6**	**99.8**
	Attractive	Bald	**74.2**	0.0	72.9	**10.9**
		Big_Lips	**77.4**	6.9	76.9	**89.4**
		Chubby	**74.7**	0.0	71.3	**5.6**
		Narrow_Eyes	77.7	17.1	**78.4**	**99.0**
	Necklace	Bald	**84.8**	0.8	84.3	**98.7**
		Big_Lips	**84.8**	95.5	84.0	**99.8**
		Chubby	**84.8**	56.9	84.0	**99.0**
		Narrow_Eyes	**84.8**	97.8	83.3	**99.0**
FairFace	Age-2	Race=Indian	68.1	8.5	**69.7**	**97.5**
	Age-3	Race=Indian	**64.3**	0.0	64.1	**94.6**

Table 4. Evaluation of the naive and LASSI models using a wider range of sensitive attributes and tasks. Highlighted in bold are the highest accuracy and fairness between the naive and LASSI model. Highlighted in red are two unexpected values, further discussed in Section 4.2.2.

Sensitive attrib.:	Pale Skin		Young		Brown hair		Bags	
Transfer task	Acc	Fair	Acc	Fair	Acc	Fair	Acc	Fair
Oval_Face	68.9	97.0	67.6	97.3	67.9	97.8	68.8	98.7
Wearing_Hat	94.7	99.4	95.5	99.2	94.7	99.8	95.4	99.4

Table 5. Evaluation of the accuracy and fairness of LASSI using transfer learning, on two new unseen downstream tasks. These results strengthen the third claim made by Peychev et al. [9].

Is LASSI flawed? — To understand the significant decline in certified individual fairness under specific settings, as documented in table 4, we explore the impact of the input data (generated by the GLOW model) on LASSI. In order to do this, we select a random sample of faces and visualize the input generated by the GLOW model. In addition, we calculate the certified individual fairness scores for each face.

In figure 2 we compare this analysis for a setting of the LASSI model that results in a high individual fairness score of 98% (trained on the `Smiling` task, using `Pale_Skin` as sensitive attribute, see table 1); to a setting of the LASSI model that results in a low individual fairness score of 5.6% (trained on the `Attractive` task, using `Chubby` as sensitive attribute, see table 4).

A visual inspection of figure 2 reveals that the faces resulting in a 0% fairness score are not only altered in their chubbiness, but also in various other facial features, such as gender and age. These distinct variations of faces, serving as a collection of representative input examples, suggest that the input data generated by the GLOW model can be inaccurate under certain settings. The resulting unwanted alterations of facial features may impact the classification task at hand. As an example, in figure 2 the faces are varied in a manner that likely affects their level of attractiveness. In comparison, the faces on the left are varied only in skin tone, serving as a more accurate sample of input data.

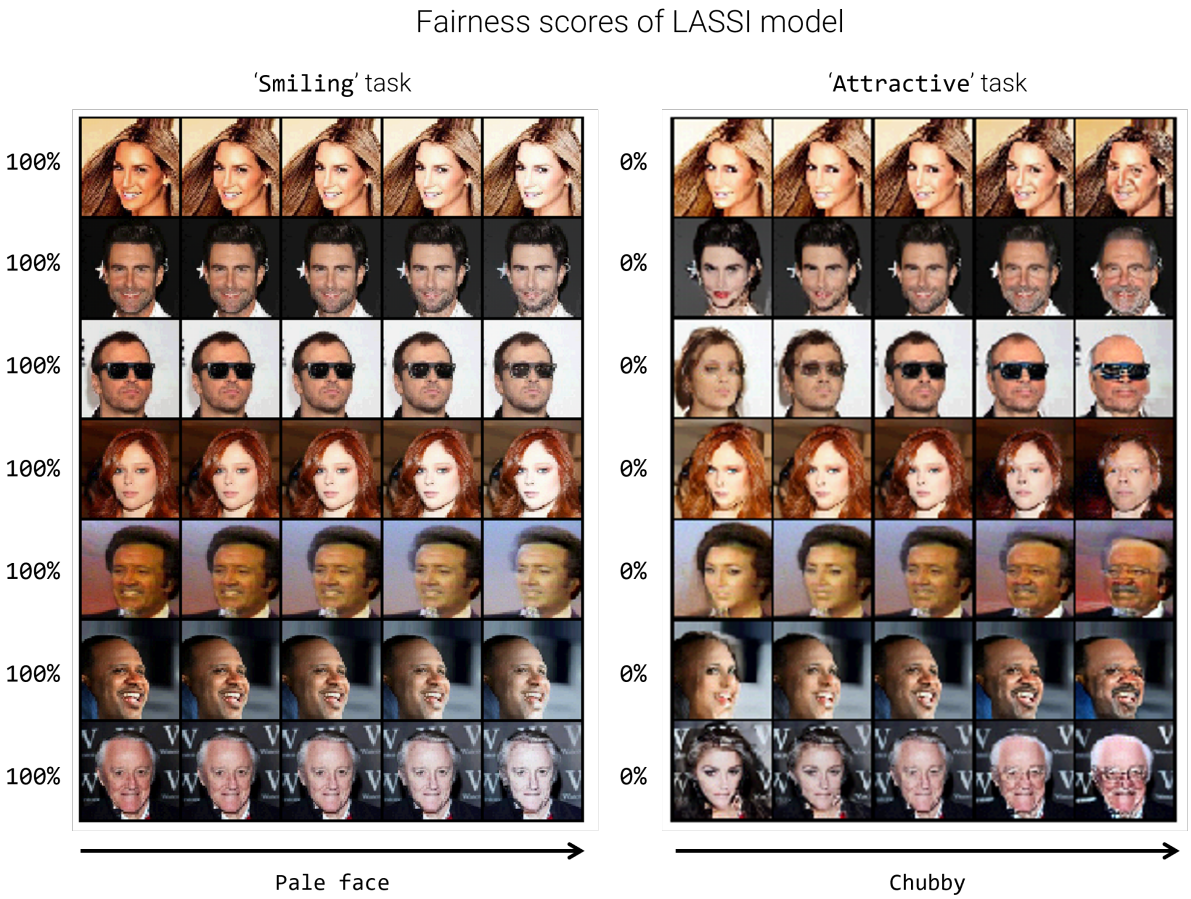


Figure 2. Visualization comparing face variations generated by GLOW and the resulting certified individual fairness scores of the LASSI model. On the left the model is trained on the smiling task, using pale_face as sensitive attribute, resulting in high fairness scores; on the right the model is trained on the attractive task, using chubby as sensitive attribute, resulting in low fairness scores.

5 Discussion

In this reproducibility studies we conducted multiple experiments in an attempt to reproduce the main findings of the work done by Peychev et al. [9]. As detailed in Section 4.1, the three main claims made in the original paper were found to be reproducible and supported by our own results. We showed that LASSI increases certified fairness on various sensitive attributes and attribute vectors, while keeping prediction accuracies high. In addition, the results indicate that LASSI achieves high certified individual fairness even when the downstream tasks are not known.

Outliers of LASSI — The additional experiments executed beyond the original paper investigated the robustness of the three main claims, by experimenting on a wider-range of sensitive attributes and tasks. Interestingly, the results from this analysis in Section 4.2 indicate a significant drop in LASSI's individual fairness scores under certain settings.

Further experiments into these surprising values show that these values do not necessarily compromise the robustness of LASSI. Two possible explanations we find are the high bias between certain tasks and sensitive attributes and the possibly corrupted input data generated by the GLOW model for certain attributes. A detailed study into these possible limitations of LASSI is given in Appendix C. In general, we conclude that the additional experiments we conducted support and even further strengthen the three main claims made in the original paper.

Conclusion — During this study resource limitations prevented us from reproducing every experiment done by Peychev et al. [9]. In addition the lower amount of random seeds used by us might affect the results found in our studies. Despite these compromises, we find very similar results in all reproduced and additional experiments conducted. In this work, the three main claims made by the original authors are reproducible and found to be robust.

5.1 Reflection

What was easy — In their original paper, the authors give a complete and detailed explanation of the theoretical background of their models and mathematics, giving us a deep understanding about the inner workings of the models and evaluation metrics presented. Together with the clear and well documented code on their GitHub repository [9], it was relatively straightforward to reproduce their experiments as accurately as our resource limits allowed.

What was difficult — The main difficulty was found within the complex structure of the code files and the dependent functions across these files. In our additional experiments we tried to visualize random samples of faces and calculate the corresponding fairness scores of these samples. The code needed to do this correctly was complex and required us to alter many functions in the original code.

5.2 Communication with original authors

Any questions we had could be answered by the extensive documentation or comments made by the original authors, and no reason to contact them was found. However, to keep the reproducibility report a fair assessment, this work has been sent to the original authors to ask for their feedback and comments. In addition, we would like to take this opportunity to thank them for their very interesting and well-documented research!

References

1. A. E. Khandani, A. J. Kim, and A. Lo. "Consumer credit-risk models via machine-learning algorithms." In: **Journal of Banking Finance** 34.11 (2010), pp. 2767–2787.
2. T. Brennan, W. Dieterich, and B. Ehret. "Evaluating the predictive validity of the COMPAS Risk and Needs Assessment System." In: **Criminal Justice and Behavior - CRIM JUSTICE BEHAV** 36 (Jan. 2009), pp. 21–40.
3. P. Tambe, P. Cappelli, and V. Yakubovich. "Artificial intelligence in human resources management: Challenges and a path forward." In: **California Management Review** 61.4 (2019), pp. 15–42.
4. J. Buolamwini and T. Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In: **Conference on fairness, accountability and transparency**. PMLR. 2018, pp. 77–91.
5. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. "Face recognition performance: Role of demographic information." In: **IEEE Transactions on information forensics and security** 7.6 (2012), pp. 1789–1801.
6. A. Hleg. "Ethics guidelines for trustworthy AI." In: **B-1049 Brussels** (2019).
7. A. I. Act. "Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts." In: **EUR-Lex-52021PC0206** (2021).
8. E. Jillson. "Aiming for truth, fairness, and equity in your company's use of AI." In: **Federal Trade Commission** (2021).
9. M. Peychev, A. Ruoss, M. Balunović, M. Baader, and M. Vechev. "Latent Space Smoothing for Individually Fair Representations." In: **Computer Vision – ECCV 2022**. Springer. 2022, pp. 535–554.
10. D. P. Kingma and P. Dhariwal. "Glow: Generative Flow with Invertible 1x1 Convolutions." In: **Advances in Neural Information Processing Systems**. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.
11. J. Cohen, E. Rosenfeld, and Z. Kolter. "Certified Adversarial Robustness via Randomized Smoothing." In: **Proceedings of the 36th International Conference on Machine Learning**. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 1310–1320.
12. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. "Fairness Through Awareness." In: **CoRR** abs/1104.3913 (2011). arXiv: 1104.3913.
13. M. Yurochkin, A. Bower, and Y. Sun. "Training individually fair ML models with sensitive subspace robustness." In: **arXiv preprint arXiv:1907.00020** (2019).
14. Z. Liu, P. Luo, X. Wang, and X. Tang. "Deep learning face attributes in the wild." In: **Proceedings of the IEEE international conference on computer vision**. 2015, pp. 3730–3738.
15. K. Karkkainen and J. Joo. "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation." In: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**. 2021, pp. 1548–1558.
16. A. Ruoss, M. Balunovic, M. Fischer, and M. Vechev. "Learning certified individually fair representations." In: **Advances in Neural Information Processing Systems** 33 (2020), pp. 7584–7596.

A Model descriptions

As we describe in section 3, the latent representation learning method proposed in the original paper achieves certified individual fairness by training the deep learning model in a way that a specific sensitive attribute is ignored in the classification process. This can be done by obtaining images of individuals who are identical but only differ in the attribute for which individual fairness needs to be certified. If these similar individuals are treated identically by the model during the training process, it is ensured that the classification is not based on the sensitive attribute but only on all the other attributes.

GLOW — To obtain images of faces that only differ in a certain sensitive attribute, Peychev et al. [9] use the generative GLOW model [10] to generate these similar faces. The GLOW model is able to alter the input data in the latent space, along a chosen attribute vector. Examples of the output of the GLOW model for the sensitive attributes used in the original paper can be seen in figure 3. We show the examples of GLOW output generated by us for the additional experiments in figure 1. The GLOW model used in this reproducibility study is pre-trained and could be readily deployed.

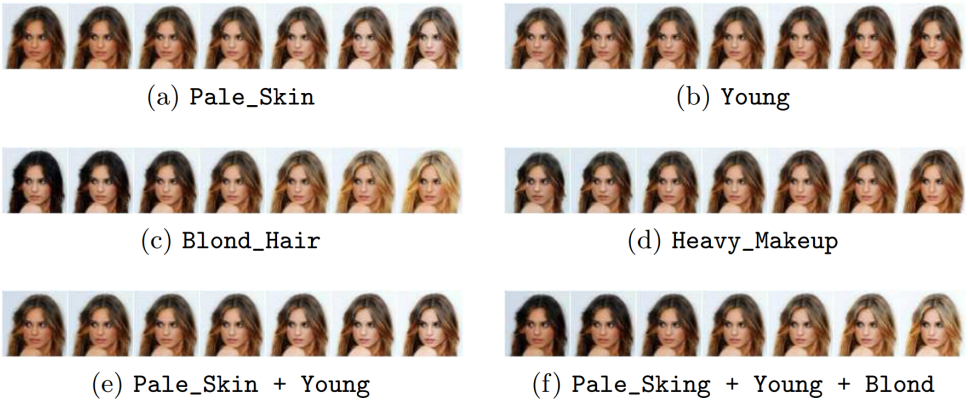


Figure 3. Visualizations of the input generated by the GLOW model for a face in the CelebA dataset. The attributes in these figures are the sensitive attributes and their combinations used by the authors in the original report [9].

LASSI — By treating all the variations of an image generated by the GLOW model in the same manner, the LASSI model learns a fair representation of a specific task. It is ensured that the model learns to discriminate only based on features other than the sensitive attribute, which the GLOW model attempts to alter as little as possible. This adversarial training is performed by uniformly selecting points on the sensitive attribute vector with a maximum perturbation level, and train the model based on these images, to ensure fair treatment.

To balance fairness, accuracy, and the ability to transfer to unknown downstream tasks, an optimal value of different losses has to be found. The adversarial loss and the loss of a reconstruction network from the representation to the latent space are added to the classification loss which emerges from the classification of the original face. The overall loss of the training is therefore a weighted average of the three losses, with hyperparameters λ_1, λ_2 and λ_3 serving as the weights for the losses of the auxiliary classifier.

Naive — To compare the performance of LASSI to a baseline, the original authors also trained a fairness-unaware baseline model, denoted as the naive model. For this naive model, the representation is learned with the loss of adversarial training and the loss of

the reconstruction network turned off ($\lambda_1 = \lambda_2 = 0$), such that it only learns from the classification loss of the original, unaltered faces.

B Datasets

From the original studies we selected two datasets to use in this reproducibility studies. In table 6 we present a short overview of both datasets.

Dataset	Size	Features	Description
CelebA [14]	202,599	40	*A large-scale face dataset consisting of celebrity images annotated on 40 attributes. The images cover large pose variations and background clutter.*
FairFace [15]	97,698	3	*A large-scale face images dataset, annotated on three attributes and balanced on the race attribute.*

Table 6. Overview of the datasets used in this reproducibility studies.

CelebA — The CelebA dataset [14] is a large scale dataset containing over 200 thousand images of real world celebrities. Each image is stored as a jpg-file and is annotated with 40 features such as Pale_Skin, Big_Lips, Smiling and Wearing_Hat. The full list can be found on the dedicated website.[1] Each attribute is annotated with a score of 1 when the attribute is present in the image, or a -1 otherwise. The attractiveness score is determined by human input.

FairFace — The FairFace dataset [15] was created to mitigate the race bias problem in most public face image datasets. It contains of close-to 100 thousand images, balanced on the race attribute.[2] The images are stored as jpg-files and annotated with four features: age, gender, race and service_test, of which only the first three are used. The possible values of these annotated features are summarized in table 7.

Feature	Values
Age	[0-2], [3-9], [10-19], [20-29], [30-39], [40-49], [50-59], [60-69] or [70+]
Gender	Male, Female
Race	White, Black, Indian, East Asian, Southeast Asian, Middle Eastern or Latino

Table 7. Summary of the possible feature values in the fairface dataset.

C Outlier study

From the additional experiments presented in Section 4.1 we find a significant drop in the fairness score achieved by the LASSI model for the sensitive attributes `Chubby` (5.6%) and `Bald` (10.9%) on the task `Attractive`. To examine these outliers we look at the possible correlation between these sensitive attributes and the task; in addition, we visualize the input generated by the GLOW model for these attributes.

[1] CelebA dataset: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[2] FairFace dataset: https://github.com/joojs/fairface

Correlation — In table 8 we present the calculations demonstrating which percentage of a certain attribute positively corresponds with a task. For example, 55.5% of faces annotated as chubby are smiling, but only 3.3% of faces annotated as chubby are tagged as attractive in the CelebA dataset.

Interesting numbers here include that only 8.2% of faces with *Narrow_Eyes* are wearing a hat. However, it is important to note that the total number of people in the dataset tagged as *Wearing_Hat* is only 4.8%, indicating that this 'outlier' follows a general trend of only few people wearing a hat.

To account for this, in the 'ratio' column in table 8 we present the ratio with regards to the total amount of people annotated with a certain attribute. For example, 49.4% of people with *Big_Lips* is annotated to be smiling, which follows the trend of the whole database in which 48.2% of people is smiling. The ratio presented here is then calculated as 49.4 divided by 48.2 to retrieve a ratio of 1.02. A ratio close to 1 therefore indicates following a similar trend to the full database.

Task:	Smiling		Wearing_Hat		Attractive		Necklace	
Sensitive attrib.	%	Ratio	%	Ratio	%	Ratio	%	Ratio
Chubby	55.5	1.15	19.9	4.15	3.3	15.55	11.6	1.06
Bald	51.3	1.06	1.0	4.80	3.1	16.55	2.7	4.56
Big_Lips	49.4	1.02	8.7	1.81	56.8	1.11	42.0	3.15
Narrow_Eyes	59.1	1.23	8.2	1.71	41.0	1.25	30.2	2.46
Total:	48.2%		4.8%		51.3%		12.3%	

Table 8. The calculated percentages and ratios of attributes and corresponding tasks presented it table 4. Highlighted in green are the lower values that follow trends similar to the full dataset, highlighted in red are the higher values, following a different trend.

Two values in this table that stand out are *Chubby* and *Bald* on the attractiveness task; which are also the two attributes and task that the LASSI model could not achieve high individual fairness on (see table 4). The ratios here are 15.55 and 16.55 respectively, which means that for every chubby face and bald face that is tagged as attractive there are over 15 faces that are *not* chubby or *not* bald and tagged attractive.

These results indicate that a face being chubby or bald has too much influence on being tagged as attractive in this dataset, preventing LASSI to achieve high certified individual fairness while maintaining high accuracy.

Causation — A possible explanation for this, is that when the LASSI model is trained on these tasks and attribute perturbations, the latent representations of faces that differ in the specific attribute are likely to diverge to a significant extent. This is done to ensure a higher prediction accuracy, because the faces that differ in the attribute must also differ in the class of the task, given the above-mentioned strong correlation.

This leads to low fairness as the perturbation results in vastly divergent data representations. In contrast, attributes that are ethically neutral, such as smiling, do not pose a concern in this regard. From an ethical perspective however, examples such as perturbations in chubbiness should not affect predictions of attractiveness.

We conclude that our experiments with LASSI produced poor individual fairness under certain settings, due to the highly biased relation between some attributes and tasks.

Corrupted input — Another explanation for the low individual fairness achieved by LASSI under certain settings, is possibly corrupted input generated by the GLOW model. To do this, we select a random sample of faces, visualize the input generated by GLOW and calculate the certified individual fairness scores achieved by the LASSI model. The code to do this is presented on the project GitHub repository.[3]

Similar to the visualizations we present in 4.2.2 we find that LASSI also achieves 0% fairness scores for this random sample of faces varied on the 'Bald' sensitive attribute. The figure shows that the faces are not only altered in their baldness, but also in various other facial features, likely impacting the classification task at hand.

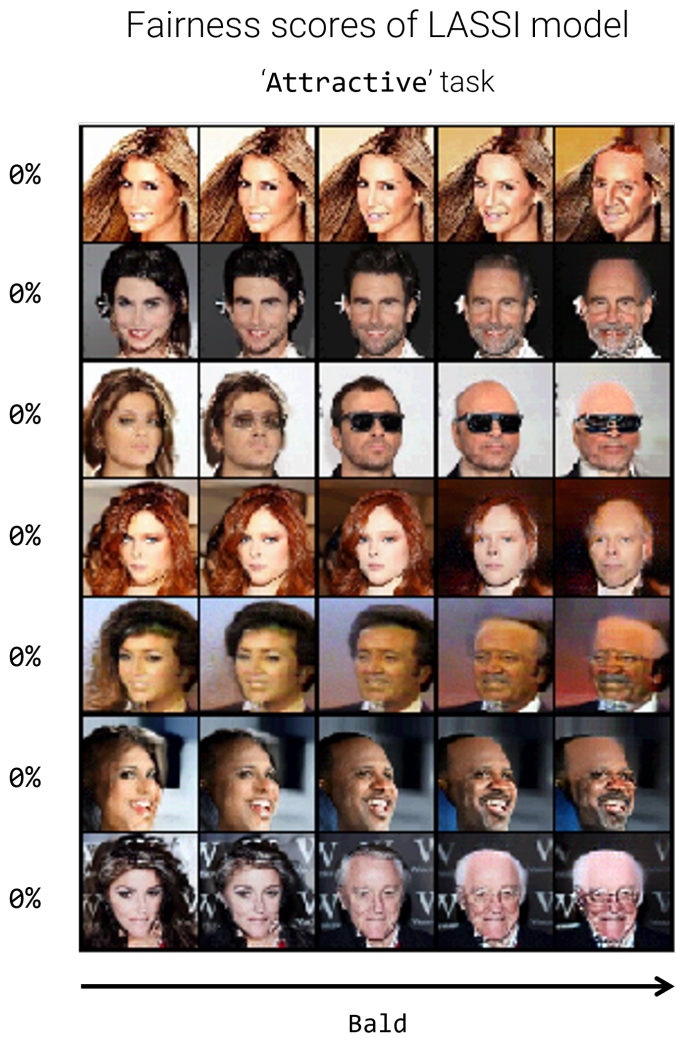


Figure 4. Random sample of faces generated by the GLOW model, varied on the sensitive attribute baldness. The individual fairness achieved by LASSI is 0% for all faces in this random sample, likely caused by the highly altered input data.

We conclude that the two settings in which LASSI achieves a low certified individual fairness do not comprise the robustness of LASSI, but are likely caused by highly correlated data and corrupted input data generated by the GLOW model.

[3]Reproducibility studies GitHub page: https://mametchiii.github.io/lassi-reproducibility/

D Visualizations of additional experiments

(a) Big_Lips

(b) Narrow_Eyes

(c) Chubby

(d) Bald

(e) Race=Indian

Figure 5. Visualizations of the input generated by the GLOW model for a face from the CelebA dataset and a face from the FairFace dataset. The attributes in these figures are the sensitive attributes we use in our additional experiments.

As described in 4.2 we performed additional experiments with new tasks and sensitive attributes. Since we have seen that the outcome of the GLOW model can be unpredictable under some settings, the output of the GLOW model with the used sensitive attributes is visualized, to evaluate if this output is not corrupted. This is important in the analysis of the LASSI model

In figure 5 we observe two visualizations which are not only altered in the corresponding sensitive attribute, namely Chubby and Bald. Their visualizations looks similar, and a change in multiple different attributes can be observed. A possible explanation for this result is that the GLOW model lacks data of these attributes, and therefore creates an attribute vector which does not correspond to the desired attribute vector. The visualizations of Race=Indian, Big_Lips and Narrow_Eyes correspond with our expectations.