

Using the Interactive Notebook: Fairness in Machine Learning

1 Getting Started

This notebook is designed as an interactive teaching tool to help computer science students explore ethical challenges in machine learning. It combines theoretical background with code, data exploration, and interactive tasks.

To use the notebook:

1. Required Files:
 - Clone or download the GitHub repository:
<https://github.com/LukasWel/ethical-challenges-in-ml>
 - Inside the repository, you will find the following structure:
 - 'Fairness_Notebook/Notebooks/'
Contains the seven Jupyter notebooks (Part 1 – Part 7) and all required images.
 - 'Fairness_Notebook/Data/'
Contains all datasets needed for the practical studies.
 - 'Fairness_Notebook/User_Guide/'
Contains this user guide.
 - 'Fairness_Notebook/Thesis/'
Contains the corresponding master's thesis to this this notebook.
2. Open the Notebooks:
 - Use Jupyter Lab or Jupyter Notebook (e.g. via Anaconda)
 - Start with "Part1_Motivation" and proceed in numerical order
 - Each notebook page builds on concepts from the previous one
3. Run Cells Step-by-Step:
 - Execute each cell in order, from top to bottom
 - Markdown cells should already be rendered, but code blocks must be executed one by one
 - The blocks in each notebook depend on previous ones, so execution order matters
4. Additional Notes:
 - Some visualizations may take a few seconds to load
 - Make sure to maintain the original folder structure after downloading or cloning.
 - For the practical studies, make sure the CSV files and image paths in the notebooks are correctly set to your folder structure.
 - The UTKFace dataset was split into two parts due to file size limitations on GitHub. Please combine the contents of both folders (UTKFace 1 & UTK Face 2) into a single folder named UTKFace. This new folder should be placed in the same location where the two parts were located:
'Fairness_Notebook/Data/GenderClassification_Study/'
 - Notebook 5 uses TensorFlow, which is officially supported only for Python versions **3.8 to 3.11**. Please ensure you are using a compatible Python version when running this notebook.

2 Solutions to Quizzes

Page 1

1. False
2. It showed different error rates between racial groups
3. Carefully auditing how features correlate with sensitive attributes

Page 2

1. False
2. Statistical Parity
3. If predictions are imperfect and sensitive attributes influence the outcome, different fairness goals can be in conflict

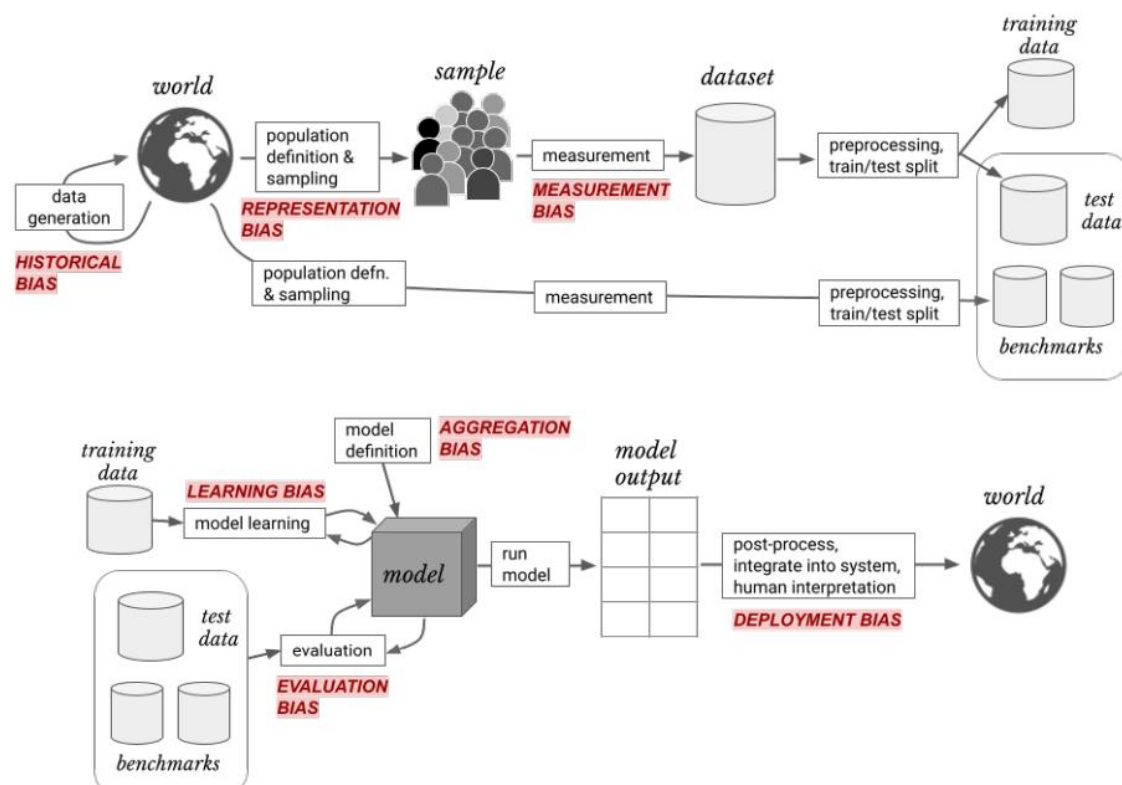
Page 3

1. True
2. Because it equalizes access to support by reducing false negatives
3. Different fairness metrics reflect different values and may conflict with one another

Page 4

1. False
2. A recidivism model uses "arrest record" as a label for "criminal behavior"
3. It can occur even in small datasets due to overfitting

Solution to the Exercise: Localize Forms of Bias in ML-Lifecycle:



Note. From Suresh & Guttag, 2021.

Page 5

1. False
2. It asks whether people are fairly represented, not fairly treated
3. Women were more often misclassified than men, especially Women of Color

Page 6

1. True
2. Misallocation of resources and distorted crime patterns
3. Down-weighting discovered incidents during training

Page 7

1. False
2. Ignoring fairness concerns at the decision boundary
3. An outcome where a better alternative exists for all groups

3 References

- Amazon Web Services. (n.d.). *Amazon SageMaker Clarify*. Retrieved May 21, 2025, from <https://aws.amazon.com/de/sagemaker-ai/clarify/>
- Amt für Statistik Berlin-Brandenburg. (2021). *Lebensweltlich orientierte Räume (LOR), Version 2021*. <https://www.berlin.de/sen/sbw/stadtdaten/stadtwissen/sozialraumorientierteplanungsgrundlagen/lebensweltlich-orientierte-raeume/>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. <https://fairmlbook.org/>
- Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy*. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp.149–159). PMLR. <https://proceedings.mlr.press/v81/binns18a/binns18a.pdf>
- Boonstra, M., Bruneault, F., Chakraborty, S., Faber, T., Gallucci, A., Hickman, E., Kema, G., Kim, H., Kooiker, J., Hildt, E., Lamadé, A., Mathez, E. W., Möslin, F., Pathuis, G., Sartor, G., Steege, M., Stocco A., Tadema, W., Tuimala J., ... Zicari, R. V. (2024). *Lessons learned in performing a trustworthy AI and fundamental rights assessment* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2404.14366>
- Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR. <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>
- Calegari, R., Castané, G. G., Milano, M., & O’Sullivan, B. (2023). Assessing and enforcing fairness in the AI lifecycle. In *Proceedings of the Thirty-Second International Joint Conference in Artificial Intelligence (IJCAI-23)* (pp. 6554–6562). <https://doi.org/10.24963/ijcai.2023/735>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163. <https://doi.org/10.1089/big.2016.0047>
- Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24, 1–117. <http://jmlr.org/papers/v24/22-1511.html>

- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). *Runaway feedback loops in predictive policing* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1706.09847>
- Fairlearn Organization. (n.d). *Fairlearn*. Retrieved May 21, 2025, from <https://fairlearn.org/>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for datasets* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1803.09010>
- Hall, M., Jenni, S., Raji, I. D., & Bethge, M. (2022). *A systematic study of bias amplification* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2201.11706>
- Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, 24, 1521–1536. <https://doi.org/10.1007/s11948-017-9975-2>
- Karkkainen, K., & Joo, J. (2021). 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* (pp. 1548–1558). <https://doi.org/10.48550/arXiv.1908.04913>
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.1609.05807>
- Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 4069–4079). <https://dl.acm.org/doi/10.5555/3294996.3295162>
- Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3), 252–260. <https://doi.org/10.1108/JICES-06-2018-0056>
- Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 34–42). IEEE. https://talhassner.github.io/home/publication/2015_CVPR/2015_CVPR.pdf
- LF AI & Data Foundation. (n.d.). *AI Fairness 360*. Retrieved May 21, 2025, from <https://ai-fairness-360.org/>
- Lum, K. & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19. <https://doi.org/10.1111/j.1740-9713.2016.00960.x>

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115.
<https://doi.org/10.1145/3457607>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 220–229).
<https://doi.org/10.1145/3287560.3287596>
- Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdli, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), Article e1356. <https://doi.org/10.1002/widm.1356>
- Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, Article 13.
<https://doi.org/10.3389/fdata.2019.00013>
- Osoba, O. A., & Welser, W. IV. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. RAND Corporation. <https://doi.org/10.7249/RR1744>
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 5680–5689). https://papers.nips.cc/paper_files/paper/2017/hash/b8b9c74ac526fffb2d39ab038d1d7-Abstract.html
- Polizei Berlin. (2023). *Kriminalitätsatlas Berlin 2023*.
<https://www.berlin.de/polizei/service/kriminalitaetsatlas/>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 33–44).
<https://doi.org/10.1145/3351095.3372873>
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, 7(11), Article 146. <https://doi.org/10.3390/data7110146>
- Robinson, D., & Koepke, L. (2016). *Stuck in a pattern: Early evidence on "predictive policing"*

- and civil rights*. Upturn. <https://www.upturn.org/work/stuck-in-a-pattern>
- Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), Article 58. <https://doi.org/10.1145/3392866>
- Schmidt, J., Pietsch, V., Nocker, M., Radar, M., & Montuoro, A. (2024). Navigating the tradeoff between explainability and privacy. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024)* (Vol. 1, pp. 726–733). <https://doi.org/10.5220/0012472200003660>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 59–68). <https://doi.org/10.1145/3287560.3287598>
- Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Vol. 17, pp. 1–9). <https://doi.org/10.1145/3465416.3483305>
- Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). DeepFace: Closing the gap to human level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1701–1708). IEEE. https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (pp. 1–7). <https://doi.org/10.1145/3194770.3194776>
- Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 5365–5372). University of Hawai'i at Mānoa. <https://scholarspace.manoa.hawaii.edu/items/bc761360-eae84bb7-ba2e4fdd8e9bcd6d>
- Zicari, R. V., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Hickman, E., Gallucci, A., Gilbert, T. K., Hagendorff, T., van Halem, I., Hildt, E., Holm, S., Kararigas, G., Kringen, P., Madai, V. I., Mathez, E. W., Tithi, J. J., Vetter, D., Westerlund, M., & Wurth, R. (2022). *How to assess trustworthy AI in practice* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2206.09887>

Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., Gerke, S., Gilbert, T. K., Hickman, E., Hildt, E., Holm, S., Kühne, U., Madai, V. I., Osika, W., Spezzatti, A., Schnebel, E., Tithi, J. J., Vetter, D., Westerlund, M., ... Kararigas, G. (2021). On assessing trustworthy AI in healthcare: Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics*, 3, Article 673104. <https://doi.org/10.3389/fhumd.2021.673104>