Capstone Progress Report 2

Bias Evaluation in Open Model Platform

April 17, 2025

Instructions: Each capstone team can use this template to capture and summarize progress. This should be shared with the sponsor and faculty mentor. When submitting the report during the course, a PDF file is preferable.

Stakeholder Names and Roles

Stakeholder	**Role**
Becky Desrosiers	*Team member*
Abner Casillas-Colon	*Team member*
Naomi Ohashi	*Team member*
George Shoriz	*Team member*
Phillip Waggoner	*Mentor*
Emanuel Moss	*Sponsor*
Elizabeth Watkins	*Sponsor*
Dawn Nafus	*Sponsor*

Project Title: Bias Evaluation in Open Model Platform

Project GitHub repository: https://github.com/oatmeelsquares/BiasOMZ

Abstract

This project seeks to evaluate one model in Intel Labs' Open Model Zoo for potential bias against protected characteristic(s). The team will start by researching bias metrics, identifying useful datasets, and choosing a model that will be feasible to evaluate, with appropriate metrics. Finally, the model's inference on the datasets will be evaluated using the chosen bias metrics.

Outline of the Project

Intel's Open Model Zoo (OMZ) is an offering that allows regular people to use pretrained AI models for their own purposes. This project is important because AI models can easily have bias inherent to their training, which can impact their performance and their impact on society, depending on for what purpose they are leveraged. Since the OMZ is publicly available, the potential for applications is

expansive and, in turn, so are the potential for consequences from biased training. Stakeholders could include anyone who employs the model, or anyone who could be affected by myriad applications of the model.

The scope of this project is one model and finding a metric or set of metrics that can quantify the bias in the model's training. We assume that we start with a possibly biased, pretrained model. A stretch goal of the project will be to create a pipeline that will facilitate future bias detection in other models.

Success Criteria

SC1	Summary of existing bias metrics/evaluation methods for AI models
SC2	Identified dataset(s) with ground truth that can be used with the chosen model
SC3	One model summarized in terms of chosen bias metrics
SC4	(Stretch goal) Initial development of testing pipeline to follow the lead of this project

Progress report

SC1 – we have completed our research and lit review

SC2 – we have identified a model to test and a reasonable dataset

SC3 – we have our final bias report from Aequitas and a preliminary one from AIF360

SC4 – This goal will be modified to essentially documentation of our progress and publishing of clean code to GitHub. It should be usable by any following groups for continued experiments.

Model: face-detection-0200

We decided to go with a face detection model largely arbitrarily. Facial recognition is a known hot area for bias and image recognition is well-known even to laypeople. Once we found the FairFace dataset, it confirmed our choice.

Dataset: FairFace (described below in Section **Summary of the Data**)

We chose to use the FairFace dataset because it works for the model we're testing and it comes pre-labelled with age, gender, and face data. It is big enough to get generalized statistics, and a subset that is uniformly distributed across gender and race, which may be helpful to our metrics. This dataset will also be useful for future testing on facial recognition models.

Metrics:

- **Equal Parity** – ensure every group is represented equally in the dataset
- **False Negative Rate Parity** – same false negative rate between groups, important when the model is used for benefitting people
- **False Omission Rate Parity** – same false omission rate between groups, important when the model is used for benefitting people
- **AUC** – same model performance between groups, useful for overall fairness
- False Positive Rate and False Discovery Rate parity were also considered, but turned out to be undefined in a dataset full of only positives

We consider an 80% default disparity intolerance for all quantitative metrics for baseline analysis. This should vary based on the metric applied and based on the implementation of the model.

Data Assumptions and Limitations

Identifier	Description
[A]	The FairFace Dataset is supposed to be balanced in terms of race, gender, and age. We found that a partial subset of the data is balanced by race and gender, but there is a nonuniform distribution of ages.
[L]	FairFace is full of only faces, so any false-positive-based metrics will be of limited usefulness

Summary of the Data

We will be using data that is publicly available online. There will be no need to store it long-term.

We recognize there are three main biases around ethical concerns in facial recognition technology. We plan to focus on the race bias (target variable) in the first round, followed by potentially age and gender.

1. Racial bias towards certain races
2. Gender bias towards certain genders
3. Age bias towards certain age groups

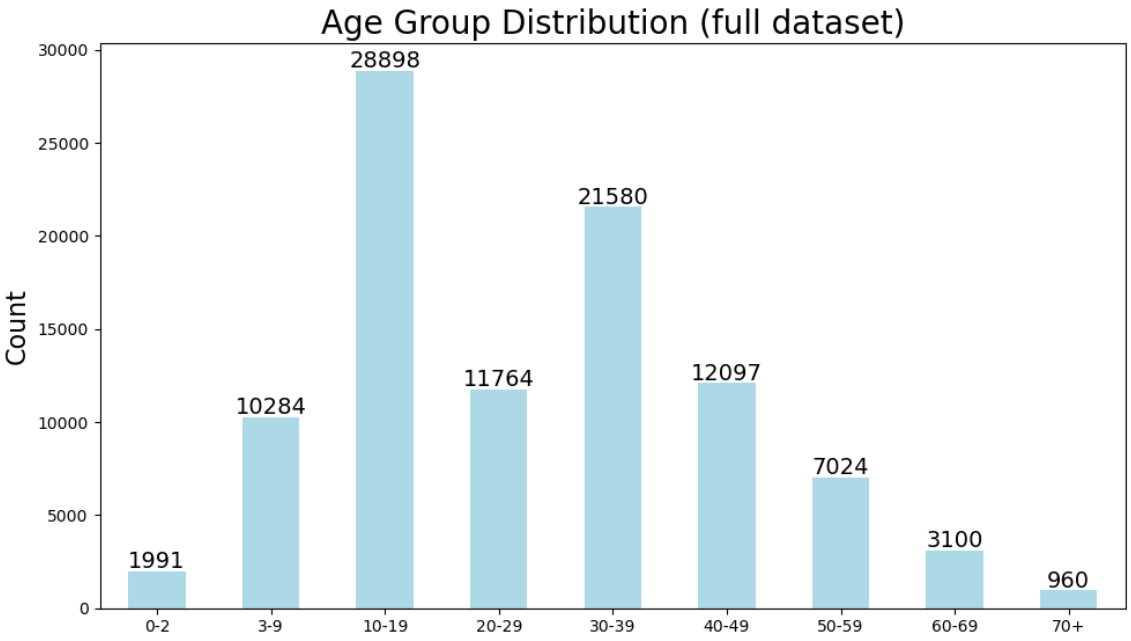
We found FairFace, a novel face image dataset including 97,698 collected from the YFCC-100M Flickr dataset and labeled with race, gender, and age groups. Each image has 5 labels:

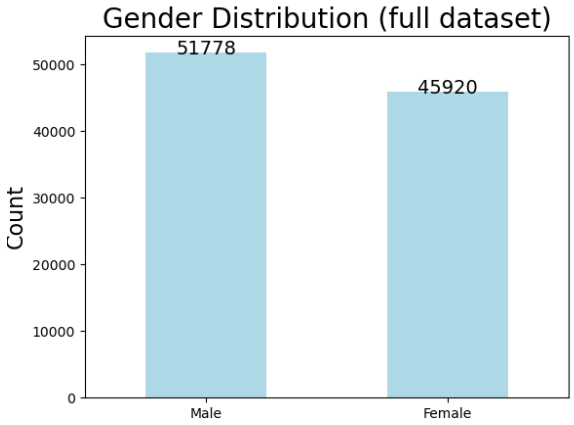
1. file
2. age
3. gender
4. race
5. service_test

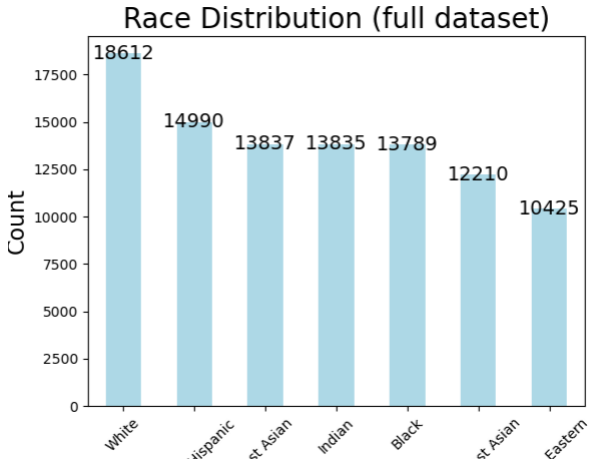
Feature categories:

age	gender	race	service_test
0-2	Male	East Asian	True
3-9	Female	Indian	False
10-19		Black	
20-29		White	
30-39		Middle Eastern	
40-49		Latino_Hispanic	
50-59		Southeast Asian	
60-69			
70+			

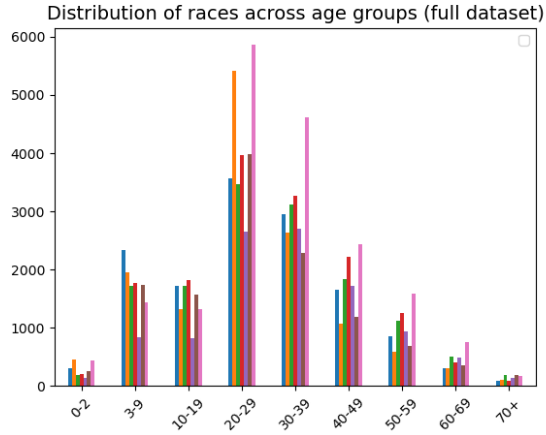
`file` is the filename of the associated image. `age`, `gender`, and `race` represent the demographic categories of each image, and their respective distributions in the dataset are shown below:

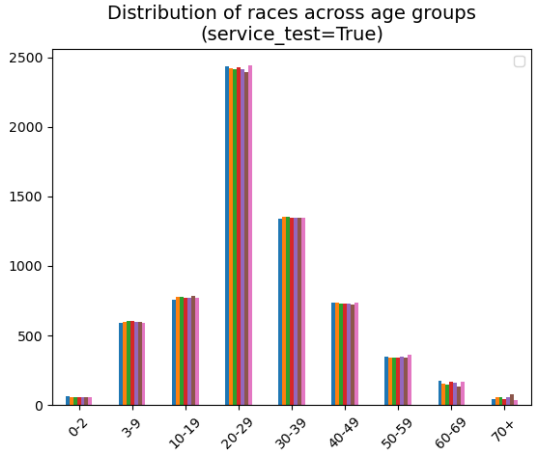

Age Group Distribution (full dataset)
Count
30000
25000
20000
15000
10000
5000
0
28898
21580
12097
11764
10284
7024
3100
1991
960
0-2
3-9
10-19
20-29
30-39
40-49
50-59
60-69
70+


Gender Distribution (full dataset)
Count
50000
40000
30000
20000
10000
0
51778
45920
Male
Female


Race Distribution (full dataset)
Count
17500
15000
12500
10000
7500
5000
2500
0
18612
14990
13837
13835
13789
12210
10425
White
Indian
Black

The meaning of the boolean `service_test` column is somewhat ambiguous, but based on an issue[1] on the FairFace GitHub page, it is an indicator of a subset of 40,252 rows where the dataset is balanced by gender and race within each age group as well as across all age groups, as can be seen in comparing the figures below:

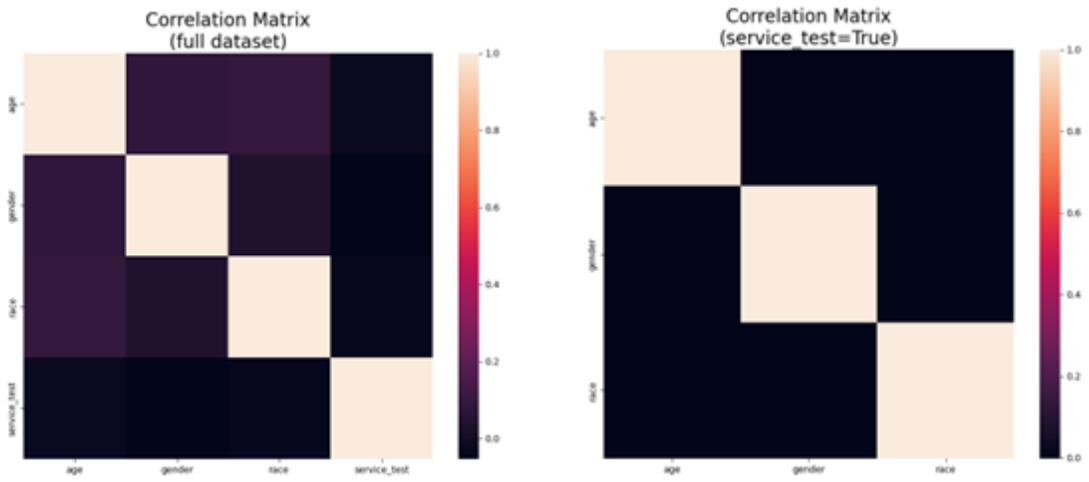




We find the same alignment in gender across age ranges when applying the service_test=True filter. This also results in a uniform distribution across gender and race for the sub-dataset overall, while the age distribution remains about the same. We will experiment with using both the full dataset and the subset where service_test=True.

Correlation

We found no/negligible correlation between the labels (max magnitude 0.088) in the full dataset, and even less in the age-balanced data subset, as shown in the heatmaps:



Summary of Modeling and Analysis

Getting model predictions

To evaluate bias a model and dataset were needed that would provide a reasonable output that could be reformatted and compared between groups. The face-detecting-00200 model that was chosen is an object detection model based on Mobilenetv2 (a CNN) using a multiple SSD head for face detection.

After importing the dataset that would be used for testing, the next steps involved the setup of the model and preprocessing of data for the correct input formats. Model setup and deployment was done using OpenVINO's python API wrapper in a Jupyter notebook. Once the model had been installed, configured, and compiled, the data was preprocessed to fit input requirements. Preprocessing included resizing the image, reordering dimensions to account for batching, and ensuring the correct pixel order of blue, green, and red.

Challenges did arise upon implementation of the model and how it tied to output evaluation. To control our labeling, images that were passed all contained a face and only one face for detection. This led to concerns if the bounding boxes were accurately accounting for the resized image that took up almost the entire frame. As part of the processing steps code was used to evaluate and determine the bounding boxes were acting as expected. The dataset utilized had only faces which also introduced potential issues with recognition and accuracy rates that needed to be addressed in the evaluation. This was done utilizing thresholds later on but was a potential concern that was explored during the modeling steps.

Getting metrics from Aequitas

Objective and Dataset

This evaluation investigates whether the Intel Labs OpenZoo Face 0200 detection model demonstrates any demographic bias in detecting human faces. The analysis used the FairFace dataset, which includes diverse demographic labels across three attributes: gender (Male, Female), race (seven groups), and age (nine categories ranging from “0–2” to “more than 70”). The dataset comprises 86,744 labeled face entries after merging with prediction data.

Methodology Overview

After aligning prediction outputs with ground truth demographic labels, three detection confidence thresholds were initially defined: 95%, 90%, and 80%. Each threshold was used to create a binary score representing whether a face was successfully detected (1) or not (0). For the purposes of evaluating model bias, all faces in the dataset were labeled as positive ground truth (label_value = 1) since every image contains a face.

While a 90% threshold was briefly tested, it offered limited analytical value and is therefore excluded from final reporting. The 95% threshold was selected to reflect a statistically significant alpha level of 0.05, providing a strict baseline for detection accuracy. In contrast, the 80% threshold was chosen based on the general face detection confidence standard suggested by AWS, serving as a practical and industry-aligned lower bound. This dual-threshold design allows comparison between strict statistical rigor and widely used commercial detection criteria.

To assess detection disparities, Aequitas was used to generate group-based fairness metrics such as False Negative Rate (FNR), False Positive Rate (FPR), and their relative disparities compared to reference groups (Male, White, and age 20–29). Additionally, the Area Under the Curve (AUC) for each group was computed to measure separability in model confidence.

Summary of Detection Outcomes

At the 95% threshold, 7,226 faces (8.33%) went undetected:

- Gender: 4,359 Males and 2,867 Females undetected.
- Race: White (1,659) and Black (1,292) had the highest undetected counts.
- Age: 20–29 (1,996) and 30–39 (1,800) had the most misses.

At the 80% threshold, detection improved markedly, with only 1,312 faces undetected (1.51%):

- Gender: 803 Males and 509 Females undetected.
- Race: White (327) and Black (252) remained the most undetected.
- Age: 20–29 (405) and 30–39 (331) again had the highest omissions.

Fairness Analysis Using Aequitas

To assess fairness in the Intel Labs OpenZoo Face 0200 model, we applied Aequitas bias metrics at two key thresholds: a 95% and an 80% threshold. The evaluation focused on disparities in prediction performance across gender, race, and age groups, relative to the reference groups: Male, White, and age 20–29.

Summary of 95**%**, Aequitas:

- **FNR Disparity**: Younger age groups (e.g., 3–9, 10–19) showed lower FNRs than the reference (20–29), while older groups (e.g., 50–59, 60–69) had up to 34% higher FNRs.
- **Race Disparities**: Southeast Asian, Indian, and Latino_Hispanic groups experienced 26–35% lower false negative rates compared to Whites, while Black and White groups had similar FNRs.
- **Gender Disparity**: Female faces had a 26% lower FNR than the male reference group.

Summary of **80% threshold**:

- All disparities decreased significantly.
- FNRs dropped overall to near or below 2% for most groups.
- While relative disparity values persisted, their practical significance diminished due to the improved overall detection.

95% Confidence Threshold

At the stricter 95% threshold, the model demonstrates greater selectivity, which amplifies disparity in detection rates across demographic groups. The False Negative Rate (FNR) is the primary fairness indicator here, as it represents how often the model misses detecting a face when one exists.

- *Gender:*
  The model misses fewer female faces relative to males. The FNR for females is 7.0%, while for males it is 9.5%. This translates to an FNR disparity of 0.74, indicating the model is approximately 26% less likely to miss a female face compared to a male face. The corresponding True Positive Rate (TPR) for females is also slightly higher than for males.
- *Race:*
  Disparities among racial groups are more nuanced. While White individuals are the reference group, several others show better performance. For instance, Southeast Asians (FNR disparity: 0.65) and Indians (0.67) are missed significantly less often than Whites. In contrast, Black individuals (1.05) are slightly more likely to be missed. This suggests that White and Black faces experience relatively higher false negatives, while most other racial groups benefit from more favorable detection behavior. Notably, the Prevalence Rate (pprev) and Prediction Rate (ppr) are consistent with the distribution in the data, indicating no severe sampling imbalance.
- *Age:*
  Age-based disparities are pronounced. The model is most accurate with children ages 3–9 (FNR disparity: 0.73) and 10–19 (0.83). However, accuracy declines sharply for older groups: individuals aged 50–59 (1.32), 60–69 (1.34), and more than 70 (1.25) are significantly more likely to be missed compared to young adults aged 20–29. These disparities imply that as age increases beyond 30, the likelihood of detection failure increases, which raises fairness concerns for older populations.

80% Confidence Threshold

At the 80% threshold, the model accepts a broader range of predictions, which generally reduces FNRs across the board. Nonetheless, disparities in group-wise performance persist, though their magnitude is reduced.

- *Gender:*
  The FNR for females drops to 1.2%, compared to 1.7% for males. The disparity is consistent with

the 95% results, with women less likely to be missed than men (FNR disparity: 0.72). This again suggests slightly better model performance for females.

- *Race:*
  Racial disparities shrink but remain in line with earlier patterns. Latino_Hispanic and Southeast Asian groups benefit most, with FNR disparities of 0.50 and 0.55, respectively meaning they are 45–50% less likely to be missed than Whites. East Asian (0.73) and Indian (0.68) groups also outperform the reference. However, Black individuals (1.04) remain marginally more likely to be missed, again signaling a consistent pattern across thresholds.
- *Age:*
  The disparity between younger and older age groups diminishes but is still evident. Children aged 3–9 (FNR disparity: 0.57) and 10–19 (0.70) maintain better performance than adults aged 20–29. Older adults, particularly those aged 50–59 (1.11) and 30–39 (1.09), remain more likely to be missed, though the gap is narrower than at the higher threshold

AUC Observations

The AUC (Area Under the Curve) scores represent how well the model distinguishes between detected and undetected faces across demographic groups. At the 95% confidence threshold, AUC scores ranged from approximately 0.89 to 0.94, indicating high overall separability in prediction confidence between detected and undetected samples. While most groups performed similarly, Southeast Asian and Indian subgroups showed slightly stronger AUCs, suggesting clearer confidence margins in those predictions.

At the 80% threshold, AUC values became even more consistent across all gender, race, and age groups, with most scores clustering above 0.93. This reflects a general smoothing effect in the score distribution, where relaxed detection criteria result in more equitable model behavior. The narrowing spread of AUCs at 80% reinforces the earlier observation: lower thresholds produce fairer and more consistent detection confidence across demographics.

Conclusion

This analysis shows that at a high threshold (95%), the Intel Labs OpenZoo Face 0200 model underperformed in detecting older age groups and certain racial categories, particularly White and Black individuals. At a general-purpose threshold (80%), detection improved for all demographics, with greatly reduced disparities. This highlights a trade-off between conservative detection confidence and equitable performance. These findings can guide operational threshold choices depending on application sensitivity to fairness and precision.

<u>Getting metrics from AIF360</u>

The AIF360 Python package offers 71 bias detection metrics, 9 bias mitigation algorithms, and a unique, extensible metric explanation facility to help users understand the meaning of bias detection results.

The following results are based on the preliminary experiments of AIF360 for demonstration purposes.

Datasets

The requirements for the dataset objects need to be met for AIF360.

1. No null values
2. Defined “Protected attributes (i.e., race, gender, etc.) that are encoded as integers

3. Categorical features that are not protected attributes are One-Hot-Encoded or ordinal encoded.
4. Define "Privileged groups" (i.e., male, white, etc.) and "Unprivileged groups" (i.e., female, non-white, etc.). With the FariFace dataset, we encoded "race" and "gender" columns as binary, protected attributes, such as white or non-white (white: 0, others: 1). Similarly, we encoded "gender," male or female (male: 0, female: 1).

Results of the preliminary experiments on AIF360

- Best balanced accuracy without fairness constraints = 1.0000. The score is extremely high with 100% true positive rate and 100% true negative rate. This is an unusual outcome, probably because the FairFace dataset is simple and perfectly balanced. This was achieved without fairness constraints, meaning the model was optimized solely to separate the classes, regardless of how different groups are treated.
- Optimal classification threshold without fairness constraints = 0.0100. Even a slight probability >1% leads to predicting the positive class. This suggests the model is highly confident in its negative predictions.
- Optimal classification threshold with fairness constraints = 0.0100. Even when fairness constraints are applied, the same threshold still gives the best balance between fairness and accuracy. This might mean the model was already behaving fairly at this threshold, or the fairness constraints weren't strong enough to shift decision boundaries.
- Optimal ROC margin = 0.0000. This suggests no difference between the model's prediction and a random guess. When fairness constraints were applied, the model lost all class separability under fairness. This might indicate a) the data is strongly biased, b) the fairness constraint is too strict, c) the groups being compared are fundamentally different in terms of class distribution. The preliminary results suggest that we need to review the FairFace dataset closely and conduct a further investigation of the results.
- Disparate Impact Ratio (DIR) = 1.4118. The DIR indicates disparity in favor of the privileged group. A ratio of 1 suggests no disparity, less than 1 indicates a positive outcome for the unprivileged group, and greater than 1 implies a higher likelihood for the privileged group. In this case, the model predictions on the validation and the test sets show a disparity where the privileged group receives positive outcomes, like loan approvals or job offers, 1.14 times more often than the unprivileged group, raising fairness concerns.

Future Work Plan

Apr 17: Report 2 DUE

Week 15: (Apr 17-23)

- Make sure all code is clear and readable
- Make clear and thorough README for GitHub
- Make sure everything is well-documented
- Start final report writeup

Week 16: (Apr 24-30)

- Finalize report
- Design presentation

May 1: Presentation and Final Deliverables DUE

Potential Concerns [C] and Blockers [B]

Identifier	Description
[C]	Keep in mind the nuance of metrics, that rarely can things be encompassed in a single number

Questions for Reflection

Think back on your capstone experience and discuss these questions as a team. You may include perspectives from the different team members where appropriate.

Reflection Question 1: What was the biggest challenge that you faced with this project?

For this project the open-ended nature of the request represented a challenge in both planning and determining what would be an appropriate goal to set for the client. Determining ethical bias in models is a vast scope especially paired with the sheer volume of models that were provided for investigation. We partially were able to overcome this challenge via combining research acquired from white papers and attempting to narrow the scope of what was needed for a deliverable after continued dialogue with the client and our mentor. Even with this however, the research domain still presented a vast area for research that continued to introduce refocusing to combat scope creep as we moved onto the model deployment and testing stages.

Reflection Question 2: Did this project stretch you to grow? If so, how?

Yes, this project stretched the whole team to grow both technically and analytically. Working with Intel labs open zoo 0200 face detection model, evaluating the model in Aequitas, and AIF360 was new to the team. It challenged us to deeply understand how to integrate and interpret the model and metrics like false negative rate and disparity and AUC scores across demographic subgroups. We had to think critically about how to structure the model so it can be evaluated, defining the thresholds in the evaluations, and making fair comparisons. This also pushed us to communicate complex findings clearly, especially in writing the report for a non-technical audience. This project helped us improve our confidence in bias evaluation and the ability to translate technical insights into meaningful insights.

Reflection Question 3: Do you believe the capstone experience will be helpful for your career? If so, how?

Paired with the ethics class, this capstone specifically will help us go about our career with a much broader mindset about data, machine learning models, and how they impact fairness and can have real consequences in peoples' lives. Furthermore, we feel that this project will be great for our resume/portfolio and will show an interest in machine learning in general, and bias specifically, that can help us stand out from the crowd when looking for a new job in data science.

Reflection Question 4: Anything else that you would like to share?

The evaluation of bias in existing models is critical efforts across industries, governments, and AI community, fundamentally supporting the principles of Explainable AI (XAI). Algorithmic bias detection approaches not only promote the best practices but also establish explainability as a foundational

requirement for trustworthy, evidence-based AI deployment. While fairness auditing should be considered essential for AI products, we face significant challenges in establishing appropriate thresholds and weights based on existing standards – many of which warrant scrutiny themselves such as the 4/5 rule[9]. We should question implicit assumptions such as the automatic categorization of white or male population as privileged groups. Such categorizations may reflect unconscious biases among developers themselves. I would advocate for contextual fairness assessment that considers both a model's technical design and its intended application environment. Meaningful bias evaluation much be nuanced, recognizing that fairness definitions vary across different contexts and use cases.

References

[1] A. A. Kuriakose, "Algomox Blog | Bias in Generative AI: Detection, Mitigation, and Management through MLOps," Bias in Generative AI: Detection, Mitigation, and Management through MLOps, Apr. 16, 2024. https://www.algomox.com/resources/blog/bias_generative_ai_detection_mitigation_mlops/ (accessed Feb. 07, 2025).

[2] A. Chinchure et al., "TIBET," https://github.com/TIBET-AI/TIBET. (Accessed Apr. 11, 2025).

[3] A. Chinchure et al., "TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models," arXiv, 2023, doi: 10.48550/arxiv.2312.01261.

[4] B. Moses and A. Dhinakaran, "ML Observability Overview | Machine Learning Observability Resources," Beyond Monitoring: The Rise of ML Observability, May 19, 2021. https://www.montecarlodata.com/blog-beyond-monitoring-the-rise-of-observability/ (accessed Feb. 07, 2025).

[5] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," arXiv, 2019, doi: 10.48550/arxiv.1902.11097.

[6] C. Liu et al., "Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions," arXiv, 2021, doi: 10.48550/arXiv.2111.09983.

[7] C. R. Sugimoto and V. Larivière, *Equity for Women in Science: Dismantling Systemic Barriers to Advancement*, Harvard University Press, Cambridge, MA, USA, 2023.

[8] D. DeAlcala, I. Serna, A. Morales, J. Fierrez, and J. Ortega-Garcia, "Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma," in 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Jun. 2023, pp. 1167–1172, doi: 10.1109/COMPSAC57700.2023.00176.

[9] D. Paperno, M. Marelli, K. Tentori, and M. Baroni, "Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood.," Cogn. Psychol., vol. 74, pp. 66–83, Nov. 2014, doi: 10.1016/j.cogpsych.2014.07.001.

[10] D. Roselli, J. Matthews, and N. Talagala, "Managing bias in AI," in Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19, New York, New York, USA, May 2019, pp. 539–544, doi: 10.1145/3308560.3317590.

[11] E. A. Watkins, M. McKenna, and J. Chen, “The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness,” arXiv, 2022, doi: 10.48550/arxiv.2202.09519.

[12] G. Barcelos, “Understanding Bias in Machine Learning Models - Arize AI,” Understanding Bias in Machine Learning Models, Mar. 15, 2022. https://arize.com/blog/understanding-bias-in-ml-models/ (accessed Feb. 04, 2025).

[13] G. Montavon, W. Samek, and K.-R. Müller, “Methods for Interpreting and Understanding Deep Neural Networks,” arXiv, 2017, doi: 10.48550/arxiv.1706.07979.

[14] H. Suresh and J. V. Guttag, “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle,” arXiv, 2019, doi: 10.48550/arxiv.1901.10002.

[15] J. Finocchiaro et al., “Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness,” arXiv, 2020, doi: 10.48550/arxiv.2010.05434.

[16] J. Himmelreich, A. Hsu, K. Lum, and E. Veomett, “The Intersectionality Problem for Algorithmic Fairness,” arXiv, 2024, doi: 10.48550/arxiv.2411.02569.

[17] K. Kärkkäinen, 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, 2019, pp. 1548-1558, arXiv, doi: 10.48550/arXiv.1908.04913.

[18] L. Rice, “Missing Credit: How the U.S. Credit System Restricts Access to Consumers of Color,” in *Who’s Keeping Score? Holding Credit Bureaus Accountable and Repairing a Broken System: Hearing before the U.S. House Committee on Financial Services*, 116th Cong., Feb, 2019, https://nationalfairhousing.org/wp-content/uploads/2019/04/Missing-Credit.pdf (Accessed Apr. 04, 2025).

[19] M. Gray et al., “Measurement and mitigation of bias in artificial intelligence: A narrative literature review for regulatory science.,” Clin. Pharmacol. Ther., vol. 115, no. 4, pp. 687–697, Apr. 2024, doi: 10.1002/cpt.3117.

[20] M. Shah and N. Sureja, “A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions,” Arch. Computat. Methods Eng., vol. 32, no. 1, pp. 255–267, Jan. 2025, doi: 10.1007/s11831-024-10134-2.

[21] O. Aka, K. Burke, A. Bäuerle, C. Greer, and M. Mitchell, “Measuring Model Biases in the Absence of Ground Truth,” arXiv, 2021, doi: 10.48550/arxiv.2103.03417.

[22] P. Saleiro et al., “Aequitas: A Bias and Fairness Audit Toolkit,” arXiv, 2018, doi: 10.48550/arxiv.1811.05577.

[23] P. Saleiro et al., “Bias and Fairness Audit Toolkit,” Center for Data Science and Public Policy, University of Chicago, http://aequitas.dssg.io/ (Accessed Apr. 08, 2025).

[24] P. Stock and M. Cisse, “ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases,” in Computer vision – ECCV 2018: 15th european conference, munich, germany, september 8–14, 2018, proceedings, part VI, vol. 11210, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 504–519.

[25] R. K. E. Bellamy et al., “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” arXiv, 2018, doi: 10.48550/arxiv.1810.01943.

[26] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” arXiv, 2018, doi: 10.48550/arxiv.1802.04422.

[27] S. Dehdashtian et al., “Fairness and Bias Mitigation in Computer Vision: A Survey,” arXiv, 2024, doi: 10.48550/arxiv.2408.02464.

[28] S. Ritter, D. G. T. Barrett, A. Santoro, and M. M. Botvinick, “Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study,” arXiv, 2017, doi: 10.48550/arxiv.1706.08606.

[29] S. Simpson, J. Nukpezah, K. Brooks, and R. Pandya, “Parity benchmark for measuring bias in LLMs,” *AI Ethics*, Dec. 2024, doi: 10.1007/s43681-024-00613-4.

[30] S. T. Erukude, A. Joshi, and L. Shamir, “Identifying bias in deep neural networks using image transforms,” *Computers*, vol. 13, no. 12, p. 341, Dec. 2024, doi: 10.3390/computers13120341.

[31] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” arXiv, 2016, doi: 10.48550/arxiv.1607.06520.

[32] T. P. Pagano et al., “Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods,” *BDCC*, vol. 7, no. 1, p. 15, Jan. 2023, doi: 10.3390/bdcc7010015.

[33] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, “A survey on datasets for fairness-aware machine learning,” *WIREs Data Min & Knowl*, vol. 12, no. 3, May 2022, doi: 10.1002/widm.1452.

[34] V. Eubanks, Automating Eligibility in the Heartland. In *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. St. Martin’s Press, 2018, https://doi.org/10.5204/lthj.v1i0.1386

[35] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” arXiv, 2017, doi: 10.48550/arxiv.1708.08296.

[36] X. Ferrer, T. van Nuenen, J. M. Such, M. Cote, and N. Criado, “Bias and Discrimination in AI: A Cross-Disciplinary Perspective,” IEEE Technol. Soc. Mag., vol. 40, no. 2, pp. 72–80, Jun. 2021, doi: 10.1109/MTS.2021.3056293.

[37] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, and M. R. Lyu, “BiasAsker: measuring the bias in conversational AI system,” in Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, New York, NY, USA, Nov. 2023, pp. 515–527, doi: 10.1145/3611643.3616310.

[38] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, and M. R. Lyu, “BiasAsker,” https://github.com/yxwan123/biasasker (Accessed Apr. 12m 2025).

[39] Y. Yang et al., “Enhancing fairness in face detection in computer vision systems by demographic bias mitigation,” in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, Jul. 2022, pp. 813–822, doi: 10.1145/3514094.3534153.

[40] Y. Zhou, M. Kantarcioglu, and C. Clifton, “Improving Fairness of AI Systems with Lossless De-biasing,” arXiv, 2021, doi: 10.48550/arxiv.2105.04534.