541 Lab 3 - Bias and fairness

Required readings

We covered some of these videos in lab, but including them here so that you have everything in one place. There might be questions on specifics from the required readings whereas the optional readings are more of a general help, optional questions, and if you are interested to explore further.

- Princeton AI Ethics Case study "Hiring by machine" (35 min read)
- **Lec 5** Latanya Sweeney "How Technology Will Dictate Our Civic Future" (02:38 - 08:15 and 11:52 - 16:24)
- **Lec 5** Michele Gilman, Mary Madden, & Alicia Luchett "Privacy and Poverty" (06:48 - 12:56)
- **Lec 6** Ava Soleimany "AI Bias and Fairness" (04:23 - 08:32)
- **Lec 6 (partly)** Joy Buolamwini "The Coded Gaze: Bias in Artificial Intelligence" (13 min video)

Optional readings

▶ Click to show

Submission instructions

rubric={mechanics:20}

You receive marks for submitting your lab correctly, please follow these instructions:

- Follow the general lab instructions.
- Click here to view a description of the rubrics used to grade the questions
- Push your `.ipynb` file to your GitHub repository for this lab (make at least three commits).
- Upload your `.ipynb` file to Gradescope.
- Include a clickable link to your GitHub repo for the lab just below this cell
  - It should look something like this https://github.ubc.ca/MDS-2022-23/DSCI_541_labX_yourcwl.
  - If you are working in a group, you can create you own (public) repo in the UBC-MDS organization and link that instead.
- All your written answers must be in your own words.

- You are not allowed to use generative AI tools to write your answers for you or simply paraphrase answer that you generate from these tools (that will lead to a failing grade), but you can use them to further understand the topics you are learning about.

https://github.com/UBC-MDS/541-lab3-AEKZ

Overall writing quality

rubric={writing:20}

You will receive an overall writing grade for the entire lab instead of for each question. This is just a small part of your total grade, but please use the Jupyter Lab spell checker extension to catch typos and read through your text for grammatical errors before submitting (or paste it into Google Docs/MS Word/Grammarly. You don't need to type anything under this cell, it is just a placeholder to generate the grading rubric.

1. Short answer questions

Keep your replies brief, 1-3 sentences per question. Although these are short answer questions, don't copy answers from the readings, use your own words so that you practice learning these concepts. These will not be discussed during the lab.

Question 1.1

rubric={reasoning:60}

1. What is historical bias?
2. What is representation bias?
3. What is measurement bias?
4. What is aggregation bias?
5. What is evaluation bias?
6. What is deployment bias?

YOUR ANSWERS HERE

1. Historical bias is the bias that arises from prior in-equalities, and exists regardless of algorithmic efficiency and data quality. This historical record will be encoded

into a dataset and thus perpetuating existing beliefs regardless of the status quo. We can look into the poor candidate classification case in the PARiS’ training which used resumes from existing employees deemed exemplary or poor fits.

2. Representation bias arises from under-representation of relevant groups where in some exceptional cases, groups are entirely missing. For example, systems like facial recognition may perform poorly for underrepresented groups, like gender class or racial groups, leading to higher error rates and unequal outcomes.

3. This is a bias that arises from the metrics incompatibility or imperfect instrumentation of features or labels. For example, using arrest rates as a proxy for criminal activity can introduce bias because arrest rates are influenced by policing patterns, a confounding variable that reflects systemic inequalities rather than actual behavior. As a result, the labels used to train the model may be distorted and unfair.

4. Aggregation bias comes from applying a single model across all groups without accounting for group-specific differences. This implies where aggregate statistics can mask underlying patterns, Simpson’s Paradox. For example, if Vaccine A works better for men and Vaccine B works better for women, combining the results without accounting for gender can give a misleading overall conclusion.

5. This occurs when evaluation methods (like test datasets or metrics) do not reflect the diversity of the real world. This comes with a poor generalization. The bias happens when a model is assessed using the wrong metric or tested on a population that doesn't reflect how it will be used in the real world. For example, a speech recognition system which is evaluated on native English speakers will be constrained for an application in a real-world scenario; because the model may underperform on non-native speakers since it is not tested on this group.

6. Deployment Bias the bias that arises after the development of the model where the output of the model goes beyond its intended purpose. For example, heavy reliance on automated predictions without proper human oversight, leading to harmful or unfair outcomes.

2. Discussion questions

This section asks you to expand a bit on your reasoning, but still aim to write succinct replies around one paragraph per sub-question. The goal of lab discussions are not to provide you with the right answers, but to help your discussion along. Your TA will assist in this by bringing up topics that you might not have thought of, ask questions to break the silence or a dead end, and move the conversation along so that you have time to go through most questions. How useful the lab discussion is for your submission ultimately relies on that you actively contribute to the discussion and help your peers contribute and exchange ideas.

Some tips to make your discussions in lab more effective

It is easy to overlook the flaws of our own reasoning, so having a discussion with colleagues is an excellent opportunity to develop your thinking and receive feedback from someone who can provide an alternative perspective from your own. Nevertheless, many people don't know how to have an effective discussion, so I am sharing a few tips for you to be able to make the most out of this opportunity:

- Commit to learning, not "winning" debates.
- Comment in order to share information and develop arguments further, not to persuade.
- Listen respectfully, without interrupting, to try to understand each others' views.
  - Don't focus on what you are going to say next while someone else is talking.
- Challenge ideas, not individuals.
  - And be open to having your own ideas challenged.
- Think about as good arguments as possible against your position.
  - This is especially useful if many of your peers have the same opinion, help your group find angles that you might otherwise be missing.
- Allow everyone the chance to speak.
  - Politely ask members of your group about their opinion.
- Avoid assumptions about any member of the class or generalizations about social groups.
  - Be careful about asking individuals to speak on the behalf of their (perceived) social group.
- Be aware of logical fallacies, but avoid pointing them out in rude or disrespectful ways.

Question 2.1 -- Hiring by machine

rubric={reasoning:100}

> 1. PARiS' lists of suggested applicants closely resembled the lists that would have been drafted by Strategeion's human HR team. To the extent that PARiS was biased towards a particular kind of applicant, this suggests that the human HR workers were as well. Are there any additional considerations/dangers of algorithmic biases compared to human ones in a scenario like this? Could

having a "human in the loop" mitigate these biases and how could that be implemented effectively given the large number of applicants?

2. Biased data pose a problem for ensuring fairness in AI systems. Given the company’s demographics, what could Strategeion’s engineers have done to counteract the skewed employee data? To what extent do you think such proactive efforts are the responsibility of individual engineers or engineering teams?
3. Job interview companies Pymetrics and Hirevue implement games from psychology research to more accurately determine the qualities of the applicants. They can also use facial recognition software to detect and judge emotional reactions during interviews. Both companies have statements around their focus on fairness in the interview process and work to eliminate human bias. What do you see as the main potential advantages and possible dangers with hiring interviews conducted this way and do you think such companies will improve the hiring process overall?

YOUR ANSWERS HERE

1. Although PARiS's results mirrored human HR decisions, algorithmic bias can be more dangerous. It scales quickly, affects more people, and is harder to detect due to its opaque nature. People may falsely assume algorithms are objective, hiding underlying discrimination, and accountability is often unclear. A "human in the loop" can help, but only if well-designed; humans must review select cases, monitor fairness, use tools to interpret decisions, and be trained to spot bias. Effective oversight requires real authority to question and adjust algorithmic outputs, not just approve them.

2. To counteract biased employee data, Strategeion’s engineers could have rebalanced the dataset, used fairness-aware algorithms, or audited model outcomes across demographic groups. These steps help reduce the risk of reinforcing existing inequalities. While individual engineers should raise concerns, real responsibility lies with teams and leadership. Ensuring fairness in AI requires collective action, not just isolated efforts.

3. Psychology-based games and facial recognition in interviews can standardize assessments and reduce some human bias, offering a more consistent view of candidates’ traits. However, these tools carry risks; emotion detection may misread cultural differences, and biased training data can reinforce inequality. Without transparency or strong safeguards, such systems may replace human bias with algorithmic bias. Their overall benefit depends on responsible design and ongoing oversight.

Question 2.2 -- Bias trade-offs and considerations

rubric={reasoning:100}

1. We covered some ways of how algorithmic bias has the potential to inflict more harm than human bias (scalability, automation, self-reinforcement, etc). Can you identify some properties of algorithmic bias that are favorable compared to human bias for addressing the root cause of the problem? In other words, could a biased algorithm be more useful than a biased human for understanding where the bias is coming from and how to reduce it?
2. A company has created a software solution to predict the prevalence of a rare genetic disease based on automated analysis of skeletal structure from X-ray images. They claim that their algorithm can detect a skeletal structural defect which is a common symptom of the rare disease and they report a 99.9% accuracy on their training and evaluation data. You are tasked with auditing this algorithm to determine whether it has any potentially harmful biases or if it can safely be employed in the national health care system. Go through the six biases from question 1.1 and explain which ones you would take into account when evaluating the company's product. For each bias either explain why you want not include it or give a specific example for how this bias could apply to the scenario here and what questions you might want to explore/ask the company during your assessment?

YOUR ANSWERS HERE

1. A biased algorithm could be more useful than a biased human in some ways. For example, to reduce bias, the engineering team can look into system logs or use certain performance metrics to detect the bias, adjust the model and minimize bias. In contrast, we all know it's not easy to change a human being's mind, especially when they are adults who believe themselves to be professional. What’s more, people are more likely to hide their bias because sometimes they know it is wrong.

   When it comes to understanding the source of bias, algorithmic bias always comes from human bias embedded in the data that is being used for training the model. In other words, algorithms reflect human biases. These biases become more visible and thus easier to detect and control. However, to truly understand a bias and where it comes from, we still need to dive deeper into human psychology.

2. **Historical bias**: Not included, or we can say it’s similar to representation bias in this case.

Representation bias: Where do the training and evaluation data come from? Do these data have a good representation of different races, age, gender and regions? If not or the subgroups are not balanced, i.e. it was trained only on children data from certain hospitals, this model will have representation bias and won't have good generalization.

Measurement bias: Not included, as this company uses typical X-ray images. Aggregation bias: Do you have the same 99.9% accuracy for all the subgroups as in the overall population? Can you use different models for different subgroups? It is related to the representation bias, since we have different subgroups, the difference between groups could be a concern.

Evaluation bias: What is the evaluation metric? 99.9% accuracy could be overfit. As this model is used for disease detection, we would prefer to reduce the false negative and provide further examination to the positive samples, therefore, the recall score should be provided as well to avoid evaluation bias.

Deployment bias: Not included. We are auditing this algorithm to minimize the deployment bias.

Question 2.3 (Challenging)

rubric={reasoning:15}

1. Representation bias and evaluation bias can both lead to poor performance on specific subgroups due to under-representation in the sampled data. In many cases it is possible to raise the model performance both for these subgroups and as a whole by balancing the data set. However, sometimes making the performance more equal between subgroups could come at the expense of overall model performance. What are good guidelines for when this trade-off is worthwhile?
2. Read an article of your choice (not from the lecture slides) relating to one of the biases that we discussed during lecture (or another important machine learning bias if you think something was left out). Provide a link to the article and explain the type of bias that occurred in the situation the article is describing as well as provide solutions for how you think that bias could have been mitigated.

1. It's good to be fair by balancing performance across subgroups, but sometimes that means compromising overall performance. So when is that a good trade-off?

- **How bad is the harm**: If performance is poor for a subgroup and that leads to real harm, like someone not getting proper medical care or being treated unfairly, then improving performance for that group should be a major concern.
- **Whether the subgroup has faced historical discrimination**: If the subgroup has experienced discrimination in the past, it is important that the model captures them accurately, even if it means sacrificing some overall performance.
- **How the model will be used**: In areas like healthcare, hiring, or education, fairness is extremely important. In less critical applications, like recommending ads or music, the trade-off may not be necessary.
- **How much performance would be lost**: If making the model fairer only slightly affects performance, it is usually worth it. But if the loss is significant, it is important to consider how much that would reduce the model's usefulness.
- **What the affected communities think**: This is not just a technical issue. It is also a social one. Getting feedback from the people who are impacted is important, and the process should be open and transparent.
- **Legal and ethical demands**: Sometimes fairness is required by law or by an organization's values. In those cases, it has to be the top priority, regardless of the performance cost.

In general, improving fairness is worth it when it reduces harm, addresses past injustices, or meets ethical and legal responsibilities, even if the overall performance drops a bit.

2. **Article Title**: The Case for Globalizing Fairness: A Mixed Methods Study on Colonialism, AI, and Health in Africa

   Authors: Mercy Asiedu, Awa Dieng, et al. **Link**: https://arxiv.org/abs/2403.03357

   Type of Bias: The article explores representation bias, historical bias, and systemic bias in the development and deployment of AI systems in African healthcare. The authors argue that many machine learning models are developed based on Western contexts and fail to consider the historical and socio-political realities in Africa, particularly the legacy of colonialism. This leads to models that often underperform or behave unfairly when applied in African countries.

   Explanation of the Bias: Because AI systems are often trained on data from high-income countries, they may not accurately reflect the needs, conditions, and diversity of African populations. For example, health models may be biased if they lack data from rural areas, darker skin tones, or underrepresented ethnic groups. The authors support their findings using a mix of literature review and qualitative research, including input from over 700 participants across Africa. Their results show that colonial legacies and existing global power imbalances can contribute to the poor performance and lack of fairness in AI tools deployed on the continent.

How the Bias Could Have Been Mitigated:

- Develop fairness criteria that reflect African realities rather than relying on Western definitions of fairness.
- Engage African experts and communities in identifying relevant health problems and co-developing AI solutions.
- Avoid direct use of Western-trained models without adjustment. Instead, adapt or retrain models with African-specific data.
- Invest in local infrastructure, such as electronic health records and medical imaging tools, to build relevant datasets.
- Build trust by being transparent and inclusive, recognizing historical injustices and ensuring communities are involved from the beginning.

Conclusion: This article shows that fairness in AI cannot be separated from history and context. To reduce bias in African health applications, developers must account for colonial legacies, invest in local systems, and involve African voices in every step of the AI pipeline.

Help us improve the labs

The MDS program is continually looking to improve our courses, including lab questions and content. The following optional questions will not affect your grade in any way nor will they be used for anything other than program improvement:

1. Approximately how many hours did you spend working or thinking about this assignment (including lab time)?

Ans: 4 hrs

2. Were there any questions that you particularly liked or disliked?

Ans: Q 2.2

Ans: [Questions you disliked]

In [ ]: