CS 4973: Responsible ML Final Quiz

Fall 2023 45 Points

This quiz has 26 questions. You have the whole class (90 minutes) to answer all the questions.

Your name:

_____

Technical Fairness Interventions: Conceptual

One helpful taxonomy for categorizing fairness interventions is the following:

- Pre-processing
- In-processing
- Post-processing

For each sentence, select the category that best describes the fairness intervention and briefly describe why.

1. You set different thresholds for different groups in a medical risk assessment algorithm, in order to equalize multiple fairness metrics simultaneously. 1 point

   Mark only one oval.

   - ◯ Pre-processing
   - ◯ In-processing
   - ◯ Post-processing

2. Briefly describe why: 1 point

   _____

3. During the training process, you add a constraint to the loss function that penalizes the correlation between predictions and protected attributes. 1 point

Mark only one oval.

- ◯ Pre-processing
- ◯ In-processing
- ◯ Post-processing

4. Briefly describe why: 1 point

_____

5. You re-sample individual data points with different probabilities, so that the training data reflects what a "fair process" would have generated. 1 point

Mark only one oval.

- ◯ Pre-processing
- ◯ In-processing
- ◯ Post-processing

6. Briefly describe why: 1 point

_____

7. You re-rank the top 20 candidate videos on TikTok's For You page so that they represent the demographics of TikTok's creator population. 1 point

Mark only one oval.

- ◯ Pre-processing
- ◯ In-processing
- ◯ Post-processing

8. Briefly describe why: 1 point

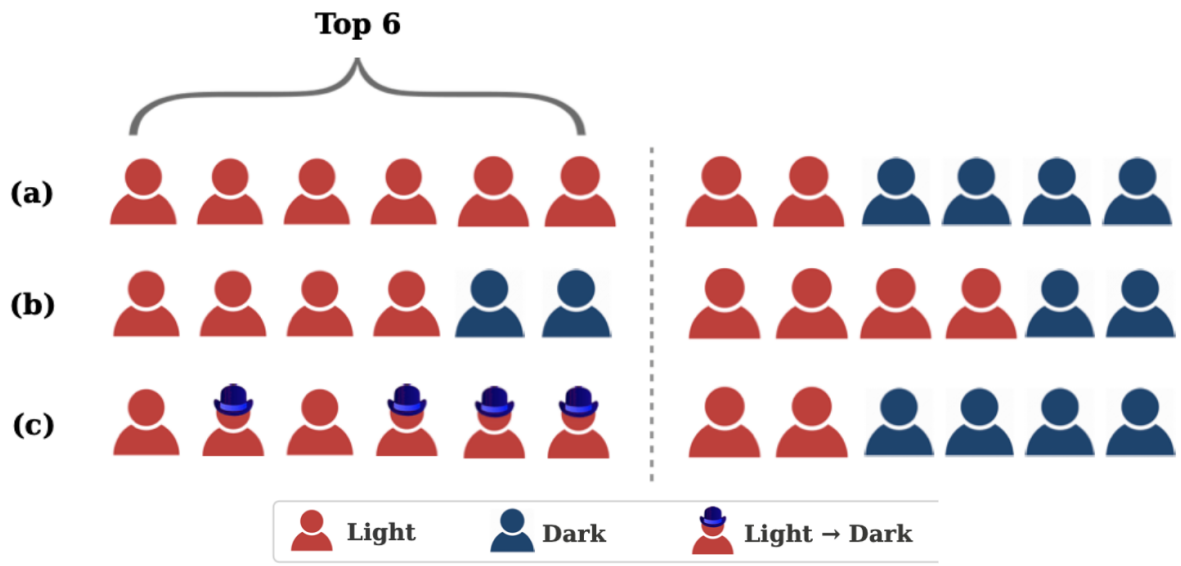
Real-World Fairness: Fair Ranking

The paper "When Fair Ranking Meets Uncertain Inference" discusses empirically what can happen when the demographic labels for people being ranked in a fair ranking system are incorrectly inferred.

The fair ranking metric Skew is defined in the paper as follows.

$$Skew_{group,k} = \frac{Fraction\ of\ group\ members\ in\ top\ K}{Fraction\ of\ group\ members\ overall}$$

In the image below, we have three different ranked lists. (a) shows the maximally unfair ranking, (b) shows a fair ranking, and (c) shows a situation where some light people have been mispredicted as dark. Calculate the real skew values for the following cases.



9. Skew (light, 6) for case a 1 point

10. Skew (dark, 6) for case b 1 point

11. Skew (light, 6) for case c 1 point

12. Skew (dark, 6) for case c 1 point

13. Explain the harm that arises in case (c) using the skew values you calculated. 2 points

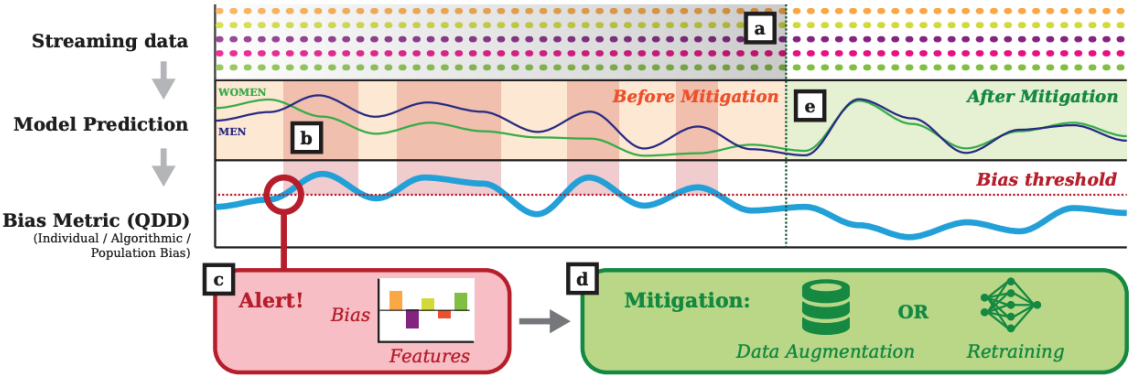
14. Which of these biases (also discussed in the paper) are specific to a top K ranking output, as opposed to a classifier model output? 2 points

Mark only one oval.

- Exposure Bias: The difference in attention received by different people in the output
- Utility Bias: The difference in the relevance/utility scores assigned by a classifier to each person
- Representation Bias: The difference in representation of people from different groups in the output
- Sample Bias: The data sample used to train the model is severely group imbalanced

Real-World Fairness: Drift

The image below is from the FairCanary paper. Answer the following questions in the context of the paper.



15. Describe what is happening in labels (b) and (c). 2 points

16. The paper discusses two types of drift: data drift and concept drift. Briefly discuss how the two types of mitigation shown in label (d) can be used to tackle either kind of drift. 2 points

17. Why was Quantile Demographic Disparity (QDD) chosen as the metric in the paper over other fairness metrics? 2 points

Check all that apply.

- [ ] QDD supports continuous outputs
- [ ] QDD supports feature-level explanations
- [ ] QDD ranges between 0 and 1, which is easy to understand
- [ ] QDD can be as granular as the operator wants, by adjusting the number of bins

AI Safety

18. In the context of AI safety challenges, give examples (hypothetical or real) of "Value Misalignment" and "Reward Hacking" failures. 2 points

19. Which of the following characteristics are concerns in a hypothetical AI model exhibiting deceptive alignment failure? 2 points

Check all that apply.

- [ ] Lack of general-purpose planning capability
- [ ] Clear alignment with human values during training
- [ ] Hard to detect malicious behavior during deployment
- [ ] Actively resisting shutdown during misaligned behavior

Memorization in Large Language Models

20. Which of the following contributes to more memorization in large language models? 2 points

Mark only one oval.

- Smaller model size
- De-duplication of training data
- Larger Model size
- GAN-based training

21. Researchers have been able to demonstrate attacks able to extract private training data from models like GPT-2. Discuss whether releasing detailed analyses of model memorization behaviors poses ethical risks. How might the AI community balance transparency around issues with preserving privacy? 3 points

Technology and Law

22. In her talk, Johanna mentions that the FTC's three-part test for "fairness" under consumer protection law considers substantial injury, unavoidability, and countervailing benefits. How might this definition of fairness differ from fairness as defined in the context of AI/machine learning systems? 3 points

23. The law often fails to adequately protect against privacy harms due to "death by a thousand cuts," where small encroachments normalize people to loss of privacy. Why might broadly framed regulations struggle to effectively address this type of cumulative harm? 3 points

AI Policy and Governance

24. Discuss two lessons that the UK CDEI learned from organizing the PETS (Privacy Enhancing Technologies) challenge. If you were a legislator organizing an event, how would you do things differently? 3 points

25. We have discussed concerns about companies running bias bounties for appearance rather than genuine problem solving. Give two ways in which the community participating can have more control and hold the companies accountable. 2 points

26. AI systems are currently used in content moderation, such as hate speech or NSFW content detection. Using content moderation as a case study, describe how a dual governance approach could help address these concerns more effectively compared to purely centralized or decentralized policy solutions alone. 3 points