CS 4973-05: Responsible Machine Learning — Fall 2023 — Avijit Ghosh

Day 2 — Preparation Questions For Class
Due: Thursday 9/21/2023 at 9:00 am on Canvas

Name: [Put your name here]

You may consult any and all resources in answering the questions. Your goal is to have answers that are ready to be shared with the class (or on a hypothetical job interview) as written. Your answers should be as concise as possible. When asked to explain a figure, your response should have the following structure: provide context (state what experiment was being run / state what problem is being solved), state what has been plotted, remark on what we observe from the plots, and interpret the results.

We recommend you use Overleaf for easy editing of this TeX document.

Directions: Read 'Can you make AI fairer than a judge? Play our courtroom algorithm game'

- Read the whole article

Question 1. *The graphics in the article illustrate a tension between equalizing error rates across groups and choosing a single threshold for all people. Why was it impossible to achieve both of these at the same time?*

Response:
Read 'Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency'

- Read Sections 1–4. (Though you're welcome to read the whole paper if you like!)

Question 2. *What fairness metric did the Twitter team use to measure disparate impact? What is a non-technical interpretation of this metric?*

Response:

Question 3. *Explain Figure 2 (only the sub-figure on the left).*

Response:
What experiment produced these results:
What is plotted:
What we observe:

Question 4. *Explain Figure 5.*

Response:
What experiment produced these results:
What is plotted:
What we observe:

Question 5. *What is "argmax bias"? What are the effects of "argmax bias"? How might you mitigate "argmax bias"?*

Response:

Question 6. *What are some of the inherent limitations of formalized fairness metrics (i.e. the demographic parity metric used by the Twitter team and the metrics used by ProPublica)?*

Response: