AI, Ethics, and Society
Final Exam

General information

The test is open-book, open-note, open-internet: you may consult any materials you want as long as you do NOT interact live with another human being. This means you may not post about the exam on Ed, text others about the exam, email others about the exam, talk on the phone about the exam, or otherwise gain live support from another person by any other means.

Do not just copy and paste information from articles. If you copy and paste information, you will be flagged for plagiarism and may receive a 0 for the exam. The final exam should be submitted as a PDF in JDF format. The file name for submission is GTuserName_FinalExam. Reports that are not neat and well organized will receive up to a 10-point deduction.

Professor Woke decided to create a new course around the theme of Ethical AI. As she browsed the internet for stories about the abuse and misuse of AI, she became saddened by all of the problems out there that developers didn't even seem to be aware of. She worried that the tech backlash against technologists because of this misuse would override the potential benefits that AI could have for society. She also worried that developers, even if they hear about this abuse, would not be mindful enough to want to fix the problem. She wonders if she should just give up and take those offered jobs in corporate management, which actually pays much better with less worries.

It is your task as a student to convince Professor Woke that, as a technologist, you understand some of these problems with AI misuse and have ideas about how to address some of the corresponding bias and unfairness issues.

Task 1 (25 pts): You must find a public artifact (newspaper/magazine article, blog post, YouTube video, etc.) that has been released within the last six months (November 1, 2022 – May 1, 2023). The public artifact must identify some aspect of AI misuse as applied to an application/scenario/domain; the misuse must impact a regulated domain and/or a legally recognized protected class; the public artifact must have associated with it, some form of data evidence (either a research publication, a released dataset, results from a survey, etc.). *Note: The evidence does not need to have been released within the last six months, only the public artifact.* Provide the following information related to your artifact: Title of artifact, release date, link to artifact, application/scenario/domain of misuse; regulated domain/protected class impacted; link to evidence. As an example (caveat: this public artifact was released outside of the six-month time period):

- Public Artifact:
    - Title - "Can you make AI fairer than a judge? Play our courtroom algorithm game"
    - Released - October 17, 2019
    - Link - https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm/
- Application/Scenario/Domain of Misuse: Criminal Risk Assessment (Predictive Algorithm)
- Regulated Domain/Protected Class: Education/Race
- Evidence: Dataset - https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

Task 2 (20 pts): Provide a 1-2 paragraph summary of the bias that is identified by the public artifact. In your description, your summary should be specific and reference definitions, concepts, ideas discussed during this course (i.e. in lectures, assignments, cases, etc.). Please BOLD all references to definitions, concepts and ideas discussed in class.

Task 3 (30 pts): Provide specific details (in bullet or table format) on any and all quantifiable metrics that are available or can be derived from compiling together information from the artifact and associated evidence. There should be enough metrics (at least 6) and details provided for us to validate your ability to synthesize course concepts based on the overarching topics:

- Privileged/unprivileged groups
- Misleading graphs
- Sources of Data Bias
- Sources of Sampling Bias
- Sampling Methods Used to Collect Data
- Correlations found in the data
- Outcome measures: Averages, Standard Deviations, Quartiles, Frequency Distributions, Margins of Error
- Bias & Fairness (or other) metrics used to identify differences in outcomes

Task 4 (25 pts): Identify an issue related to one of the quantifiable metrics listed above (Task 3) that, if addressed, might help mitigate bias and/or unfairness. Design a method to help address the issue identified. The method should relate to a concept discussed in the lectures. You can explain the method using pseudo-code with an explanation or a python script with comments. Remember to identify the issue, the data inputs and outputs (based on the evidence), and the anticipated change in outcomes.
Note: You do not need to have a working code or quantitative results.