1 Summary of the coursework

This coursework involves two parts: a 3–page written essay, and a 3–page scientific report with code implementation, discussed in the following two sections accordingly. The deadline for submission is on May 4th, 2021 at 22:00. You will submit all the deliverables in a single compressed file (preferably .zip format) through DUO. If there are any queries regarding this coursework, please do not hesitate to contact me: ehsan.toreini@durham.ac.uk

2 Essay Project Outline

In this part, you will prepare an essay on your understanding of bias and discrimination in artificial intelligence. Your essay should follow a survey-like approach based on the reading assignments I have suggested at the end of each lecture week. I would encourage you to search more on this topic and go beyond my suggested list of reading in your final essay. You can find the reading assignment at the last slide on each week lecture; however, the full list will be released in DUO.
This essay should demonstrate your point of view on the overall subject of algorithmic bias. As the sole–author of this essay, I will leave the final structure of the essay to you. However, here are my suggestions: As the author, you will discuss the justifications on the reasons that bias in AI–based solutions should be addressed. Also, your essay should demonstrate various ways to measure the fairness of a dataset and algorithm and discuss different ways to mitigate algorithmic bias. Moreover, you will also discuss what you expect to see in the fair machine learning solutions in the future.
You should appropriately use citations and back your proposed discussion with relevant literature. You can use the reading assignments or you can search your own. However, you should focus on peer-reviewed references with reasonably well–cited manuscripts.
Your final manuscript should be submitted in pdf format (maximum 3 pages in IEEE conference template [1]).

3 Implementation Project Outline

In this project, you will be analysing a human-centric [2] dataset and develop a fair machine learning ecosystem to detect, reduce and eventually mitigate different types of bias that exist in the final outcome of the algorithm in various ways.

3.1 Task 1: Project Proposal

The goal of the first sub-task in your coursework is to select the project you wish to work on during this term. In this stage, you are required to write maximum of one page project proposal describing what you intend to implement, and the reasons you believe that it suits this submodule as bias in artificial intelligence.
Please describe what you'd like to do for your course project. Describe the motivation for this project, what concrete tasks you plan to do, and what the final work product of these tasks will be.
Is there a particular context you're thinking about (e.g., hiring, advertising) as you formulate this project? If you're planning to do an empirical project, list the dataset(s) you will use or gather.
It isn't necessary that you do precisely what you outline above for the final project, but it should help you start thinking about your work to answer these questions. Also, please indicate what technologies you plan to use in your implementation (i.e. the programming language, packages that you use for data analysis and the implementation of AI algorithm).
Please consider one specific section on the progress of the project. Leave it blank for now. You will gradually complete this section by the end of this coursework. Remember the overall length of the project proposal document should not exceed 3 pages (1 page for project proposal and 2 pages for project progress report) in IEEE conference template [3]).
You can use the project suggestions in separate files available on DUO.

[1] You can find the word template file in a folder named *coursework* in DUO
[2] A dataset such that each entry represents one measurement of one person.
[3] You can find the word template file in a folder named *coursework* in DUO

3.2 Task 2 - Data Analysis

Find a human-centric dataset which is publicly available with demographic information (age, gender, race, sexual orientation, country of origin, etc). The dataset should be in the scope of your project proposal in Task 1. Please follow these steps:

1. Please include a link to the dataset you used, as well as any documentation that accompanied its release. Clearly describe any "cleaning", binning, bucketing, or discrete-to-continuous feature transformation you did in this process in the *project progress* section left blank in the project proposal document.
2. For one way of splitting the dataset into different demographic groups, write down the size of the groups, the average value for each (numeric) feature, the variance of each (numeric) feature, the mode for each categorical feature, and the three most frequent values for each categorical feature, each computed on the different demographic subgroups. If your dataset has more than 20 features, you can report this information only for 20 features. Do you observe any interesting differences between the different subgroups' statistics?
3. If you have observed any sort of bias, please describe it and explain the reasons why this bias has happened from your point of view in the *project progress* section in of your project proposal document.

3.3 Task 3 - Conventional implementation

At this stage of the project, you have a biased dataset. Now, you will implement a conventional ML algorithm that suits the project description. For instance, if you have chosen to solve a classification problem in your project proposal, you will implement a classification algorithm that is widely used (and is potentially biased). You can use the same conventional ML model as used in the suggested papers in the scope of your chosen project.

1. Describe your chosen algorithm and justification for selection in the *project progress* section in the project proposal document.
2. Naively split your dataset into training and testing sets by randomly sampling some of the data, for example, 70% train and 30% test. If your model has hyper–parameters that you wish to optimise, you may wish to create a separate validation dataset to optimise these, for example 70% training, 15% validation, and 15% testing.
3. Train your model and see how it generalises to the testing dataset. Explain your approach and findings.
4. Subsample a new testing dataset in an unbiased way and representative of the task, for example, you may wish to ensure gender and age diversity. Retrain your model and see how it generalises to these new testing conditions. Compare your findings with the results in 3 and explain your approach.
5. If you have observed any sort of bias, please describe it and explain the reasons why this bias has happened from your point of view in the *project progress* section in of your project proposal document.

3.4 Task 4 - fair machine learning implementation

In this step, you will implement one of the fair ML methods of mitigating bias in the scope of your selected project (if you choose only one algorithm to implement, you should justify your choice in the *project progress* section). You will use the solutions that were provided in the suggested projects papers (if there is no paper suggested for your project, you can use the methods taught in the class.) Now, follow these steps:

1. Implement the fair ML solution used in your project (if there is more than one, implement only one. If there is no solution mentioned, you can use select your algorithm from the ones I taught in class or a similar research paper)
2. Describe your proposed algorithms in the *project progress* section.
3. Now test the performance (i.e. accuracy, sensitivity or similar criteria, can be found in the research paper for your suggested project) of your trained model for the minority groups (if you have more than one, choose only one). Compare it with the performance of your model over the majority group.
4. If you have observed any reduction in algorithmic bias, describe it and use appropriate plots to demonstrate it in your *project progress* section.
5. Do you get roughly the same results as your project paper? If not, reconsider your code or justify the reasons in your *project progress* section.

6. If you have observed any sort of reduction in accuracy, please describe it from your point of view in the *project progress* section in your project proposal document. Use proper plots and graphs where necessary.

4 Timeline and Project Marking Guidelines

4.1 Project Timeline

The deadline for submission is 22:00 on May 4th, 2021. You will submit the two projects in one single compressed file (preferably in zip format). Use your student ID as the name of the zip file. The essay and project proposal (and progress report) should be in pdf file format. Your final project submission should have the following items:

- Essay coursework, in PDF file format
- Project proposal, in PDF file format
- Implementation coursework, in a separate folder, in py or ipyb file format

4.2 Marking Guideline

Essay Coursework. [30 marks]
You submitted essay will be marked based on the quality of writing, relevance of the citations, depth of discussions on the topics above and the overall correctness, fluency, clarity and delivery.
Essay coursework will be assessed based on the following criteria:

- evidence of adequate and appropriate background reading
- a clear statement of aims and relevant selection of content
- sensible planning and organization
- evidence of systematic thought and argument
- clarity of expression
- careful presentation (e.g. accurate typing and proof-reading, helpful diagrams, etc.)
- observation of conventions of academic discourse, including bibliographic information
- observation of length requirements

Implementation Coursework. [70 marks]
The code and project proposal will be marked as follows:

- Implementation [35 marks]
  - Does your project work?
  - Correct implementation of fair ML solutions.
  - How effective and considerate of various biases is your dataset sampling strategy?
  - How well does your model generalise? Is it over–fitting or under–fitting?
  - The final project should mitigate existing bias in the dataset while maintaining an acceptable level of accuracy
- Sophistication and appropriateness of the solution [10 marks]
  - How well have you applied the relevant theory to the problem?
  - How hackish is your implementation, or is it robust and well-designed?
  - Have you just cited and pasted code, or is their evidence of comprehension with further study and novel design extending beyond the lecture materials?
- Project proposal [25 marks]
  - The reasoning of the project choice

- overall quality of writing (same criteria as essay coursework)
- clarity of the project proposal
- the completion of *project progress* section
- Does the discussions in the *project progress* sections match the source code results?
- Justifications of implementation choices in the *project progress* section.

5 Frequently Asked Questions

It is strongly recommended to read these common questions and answers carefully.
I just read a very interesting research paper that is not in the list of suggested projects. Can I use it?
Of course! I encourage everyone to find a project that excites them! However, make sure you can develop the algorithms used in the paper before committing to its implementation. You can get an idea of how difficult an algorithm is by googling or searching Github for its source code. In any case, if you feel confident, go for it!
I found code online which looks similar to what I need. Can I use it?
Yes, but you must cite the code in both the written report and in the comments at the top of the code. As a common practice in any software development, you first try to search and make sure you are not the first one who is trying to make it work. However, it is one of my tasks to make sure you are doing something original. So, please adapt the code and make sure you have cited it. Otherwise, it is very likely that you get caught (see the sub-mission Plagiarism and Collusion section on DUO to read about the tools used to detect this). This incurs a very severe departmental penalty.
Isn't the best strategy to just copy the state-of-the-art? Yes, if you notice from the list of suggested projects, you are already working on the state–of–the–art solutions in fair AI. If you know the literature, you will find the most reputable research in the field of fairness and algorithmic bias belongs to a conference called (ACM FAT, which recently got rebranded as ACM FAccT). So, keep looking in their accepted paper list to find the most recent developments in the field.
I'm struggling and feeling overwhelmed by all of this. The maths is too complicated and I don't know where to begin.
Try to read the paper for your suggested project carefully. It must contain a lot of implementation detail that you might have neglected. There might be a sentence somewhere in the paper that inspires you to have another brilliant idea! If you get errors, read them slowly, Google them. When you're confident enough, try to implement something a little bit more complicated. Do a slow, step–by–step implementation approach.
My writing is not as good, will it make my essay mark automatically low?
Not really, I understand for the majority of students (including myself), English is not their first language. Therefore, instead of focusing on using complicated words, try to stay as simple as you can in your writing. My first advice to everyone is to write simple. Avoid complicated sentences, words or grammatical combinations. Focus on the quality of discussions rather than making your essay look fancy. I would also recommend using online grammar check tools (such as Grammarly) to polish any mistakes. Also, you can always borrow technical words or some discussions from the papers you want to cite (just remember to cite them, then rephrase them in your wording to avoid plagiarism).
Can I use deep neural networks or just classical machine learning models?
I recommend you to follow the footsteps of the research paper you chose to implement. If there is no paper for your chosen dataset (or you just feel adventures), it is no problem to use more sophisticated algorithms.
Does the page limit include references?
No, you can have an unlimited number of references with a reference section starting on the 4th page.