Student Name: **Shonil Dabreo**

Student ID: **s3835204**

- What is the state-of-the-art on evaluation of fairness in machine learning?
- What are the main challenges on defining 'fairness' and fairness-aware measures in this context?
- Is it possible to measure fairness in the context of your WIL project? If so, how?

Discussion

The authors conducted an interview and a survey to get a broad sense of challenges faced in defining fairness measures. Survey respondents were asked questions about their team's current practices, challenges, and needs for support around fairness.

Many interviewees reported that their teams look into their training datasets and not ML models to improve the fairness in their products. Moreover, several survey respondents mentioned that they tried to address the fairness issues; the most commonly attempted strategy was **"collecting more training data"**. However, for example, in an automated scoring, to score African American students fairly, the team need the data of their highly scores (data collection team) which is very rare. Sampling high-scorers without scoring all the essays was a challenge. An interviewee recommended for an effective communication between data collectors and model developers. Following a survey was taken where majority of the respondents agreed to facilitate the communication. There may be challenges when specific user populations are less engaged with the product. So to balance the data get the more data with less engaged users to improve fairness.

The **Scaffolding fairness-aware test set design** must be well constructed and not biased. For example, the gender biases, images of female doctors often mislabeled as nurses. The test set design should capture potential fairness issues.

Another challenge was to estimate the minimum number of additional data points per sub population or which sub populations' needs to be considered when developing specific kinds of ML applications to ensure sufficient data from those sub populations or balance across them while curating existing datasets. People mostly choose attributes like age and gender but as interviewee suggests these cohorts should be defined based on the domain and problem. For example, in automated writing evaluation, it should be based on whether a person is a native speaker. However, if targeting specific sub populations improves fairness then to do it in most morally, ethically and vaguely responsible way was yet another concern. The interviewees mentioned that their practice of getting together and trying to imagine everything that could go wrong with their products, so that they can **proactively monitor for those issues**. They also mentioned that despite efforts of gathering training data to address fairness issue were hampered by teams' blind spots. This means that no one person on the team has expertise in all types of bias especially when different cultures are taken into account. For example, an image captioning system correctly identified celebrities of some countries and mislabeled some celebrities which were noticed by the customers. It is possible that a particular team member may know a popular celebrity (e.g. Selena Gomez) and so could fix that but not with a celebrity from another random country. The authors suggested needs for **implementing domain-specific proactive auditing processes**. There are **fairness metrics/key performances indicators (KPIs)** through which performance and progress are monitored. Interviewee teams executed automated tests but it's really hard for them to fix things that we can't measure. **Fairness auditing without access to individual-level demographics** is an auditing method used without access to individual demographics. However, some interviewees reported that they could only use coarse-grained demographic information (e.g., region or organizational-level demographics) for fairness auditing. These methods were abandoned by the teams citing limited time and resources to spend on building their own solutions. For example, companies working on k-12 student populations were strictly prohibited from collecting such demographics. Also, with the demographics of school the product could only be useful for gifted students or remedial students. Another example is, while working on chatbots, the challenges were to recruit a sizeable, diverse sample of user-study participants. There is a challenge in diagnosing whether specific issues (e.g., complaints from customers) are broader, systemic problems or just "one-offs". There should be an automated system to gives us data points of where we mess up.

Authors mentioned that several interviewees reported that their teams struggle to **isolate the causes** of unexpected fairness issues. It was often difficult for teams to decide where to focus their efforts-switching to a different model, augmenting the training data in some way, collecting more or different kinds of data, post-processing outputs, changing the objective function, or something else. Moreover, there were side effects when **making changes to datasets or models** to improve fairness.

Fail-soft strategy could be used to ensure that the worst-case harm is minimized. Actions (personalized messages) are designed in response to particular model outputs by which they could try to imagine the impacts these actions might have in specific false-positive and false-negative scenarios. In applications involving richer, complex interactions between the users and the system, fairness can be context dependent. **Prototyping ML systems** such as chatbots aided by **simulation tools** to see if certain forms of personalization might be harmful with respect to equity by finding ways to automate the identification of risky conversation patterns that emerge.

Lastly, several interviewees considered biases that maybe present in the humans during various stages of ML development.

Fairness in WIL Project

In our WIL Project, we are analyzing the impact of covid on economies of the countries. We will be comparing the impacts according to the current situation in terms of covid cases in that particular country and whether their measures are good enough to contain the impact of covid on the national economy.

Fairness auditing without access to individual-level demographics method is already in effect as our dataset only includes country-level non-sensitive information i.e. countries with their cases as well their quarterly GDP.

Fail-soft strategy measure could be used to minimize the harm by designing actions i.e. messages with respect to the responses to ML model outputs to try to imagine the impacts these actions might have in specific false-positive and false-negative scenarios. For example, it could be dangerous where a patient has covid but reported as negative (False negative) and a patient is covid-free but reported as positive (False positive).

References

[1]"Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?", *Arxiv.org*, 2019. [Online]. Available: https://arxiv.org/pdf/1812.05239.pdf.
[Accessed: 20- Sep- 2020].