MODULE ANSWER | Foundation

Address “Fairness”

Communicating “Fairness” in the Context of Your Dataset.

A dataset can be unfairly biased in any step of the dataset lifecycle, owing to upstream and downstream factors.

Upstream factors have to do with how the data are collected. These could look like missing data points or aggregated data – those that reflect the source producer’s individual biases, historical inequities from social constructs, and conceal important differences between subgroups. As we move downstream, human discretion by ways of selection, inclusion, filtering and sampling can further shift the representation of specific subgroups. Even further downstream, biases can be introduced due to the limitations of tools and technologies – what AI systems can and cannot do, and the kinds of feedback loops that emerge.

Fairness in Machine Learning has many definitions and criteria, each addressing a different set of implications for downstream users of ML-powered products. You’ll find some fairness criteria more relevant to your dataset than others, given its use case(s) and that different criteria often paint very different pictures of your dataset. Note that this space is rapidly evolving – we all are learning new techniques to improve the fairness in our datasets and their applications every day.

A Data Card that is comprehensive has essential clues and context from across your dataset’s lifecycle to help readers determine how to responsibly use your dataset in their contexts. In the Data Card, describe the framing of fairness in how your dataset is used for the development of ML models and the steps that you, as dataset creators, have undertaken to address or mitigate known unfair biases. In contrast, vague descriptions and metadata without context can reduce trust in your dataset, or worse, introduce assumptions in decisions about dataset use that can lead to undesirable, unfairly biased outcomes. While your Data Card may not be the appropriate surface to go into the details of each aspect of fairness in ML as they relate to your dataset, you

should be able to eloquently walk readers through the most important and objective manifestations of bias, skews, and imbalances in your dataset.

Note that this space is rapidly evolving – we all are learning new techniques to improve the fairness in our datasets and their applications every day.

Be cautioned: do not treat your Data Card as a final word on fairness-related implications in your dataset. This is just the start of a journey.

Key Takeaways

- There are many different definitions of fairness, ranging from social to mathematical. Fairness analyses for different definitions can paint a very different picture of your dataset.
- Biases, skews and imbalances can come from many different parts of the dataset lifecycle.
- An unfairly biased dataset can be one that is unrepresentative of society, or your target population – but a biased dataset can also be one that is accurately representative of a biased population.
- Running a fairness analysis is not a one-time undertaking – datasets are after all, used in technological systems that respond to dynamic sociocultural landscapes.

Actions For Your Team

1. **Discuss your system.** Invest time and effort in objective discussions of fairness and bias in your datasets or AI systems early on in the creation of your dataset. Reach out to experts who can guide a discussion of how fairness manifests in your dataset and how to design your fairness analyses – from setting meaningful criteria to determining follow up actions.

2. **Keep it comprehensive.** Work backwards from the potentially harmful outcomes to document decisions and actions from across the dataset lifecycle that could contribute or exacerbate these outcomes. Use methods like root cause analysis and futures wheel to determine undesirable outcomes and their causes, which can help you determine the necessary information to communicate for responsible use of your dataset.

3. **Context is critical for Responsible AI.** When providing the results of a fairness analysis, include the contexts in which the analysis was conducted, the intended practical application or expected impact of the analysis, and goals that reflect desirable outcomes of the use of the dataset in the development of AI systems. If you've modified definitions or made a set of assumptions to run your analysis, be sure to document those too. These are crucial actions for the responsible development of a Data Card.

4. **Connect thresholds to the real world.** Set thresholds and criteria based on limitations, failure modes, adversarial product tests, and the outlier utilities of your higher-value system features.

5. **Really explore your data and model.** Use methods like Normalized Pointwise Mutual Information (NPMi) and counterfactual analyses to explore your systems. Where possible, link to notebooks and visualizations to show your findings from fairness analyses.

Considerations

→ Can a reader understand your fairness analysis in terms of the real-world implications for individuals, communities, and societies?

→ What are the known limitations as a result of upstream and downstream factors in your dataset? Are biases, skews, imbalances, and impact from these factors documented in your Data Card?

→ What about your dataset, its creation, and use can lead to undesirable outcomes, even when used as intended? Are these cautioned against in your Data Card?