Egaleco: Advancing Fairness in Machine Learning

Final Submission Paper from the Product and Policy Team

Alan Kyle, Gurpreet Kaur Khalsa, Mudit Mangal and Orissa Rose

May 2023

Abstract

This paper is prepared for review by the UC Berkeley School of Information faculty in satisfaction of the MIMS capstone project requirements. It discusses the Product and Policy team's contributions to the Egaleco project, a Machine Learning (ML) fairness toolkit focused on evaluating fairness outcomes for ML applications in healthcare. The paper is sectioned by: Problem Statement, Project Team, Methods, Tool Development, Discussion, Features for the Future, and Conclusion.

Problem Statement

Decisions are Increasingly Informed by Machine Learning Models

Machine Learning (ML) algorithms increasingly dictate opportunities and outcomes for individuals and groups across economic, social, political, and medical contexts. This is especially true in healthcare, where algorithms are being introduced to support practitioners and professionals as they predict[1] and diagnose disease,[2] allocate resources, manage patient health records and evaluate public health trends.[3]

While the use of ML has contributed to important advancements in the field, some applications have created disparate impacts that replicate discriminatory societal and institutional inequities along race, gender, age and ability categories. Notable recent examples include:

- 2019: Obermeyer et al. found racial bias in a widely used healthcare algorithm meant to identify and help patients at highest risk of serious illness. It used healthcare expenditures, not medical visit data, as an (improper) proxy for determining patient

[1] (Woldaregay et al., 2019)
[2] (Ullah et al., 2020)
[3] (Kumar et al., 2022)

treatment needs. Because of this the algorithm assigned Black patients the same level of need as White patients when the Black patients were in fact sicker.[4]

- 2022: An algorithm built to predict liver disease from blood tests was found to miss cases of liver disease among women twice as often as it did for men (44% missed in women compared to 23% missed in men).[5]
- 2023: Insurer Medicare Advantage used algorithms to determine the amount of recovery time senior patients needed after the procedure and repeatedly underestimated time needed, prematurely terminating senior's covered care and leaving them with large bills.[6]

Companies expected to check for model bias don't have clear solutions

As acknowledgement of the impacts of ML bias grows, and Fairness, Accountability, Transparency, and Ethics (FATE) principles are embraced by industry[7] and academia,[8] companies are expected to evaluate their own products and surface harmful biases before models go to market. Accordingly, numerous commercial and open-source "AI Fairness Toolkits" have emerged to help ML practitioners evaluate their models for undue bias and proactively reduce algorithmic harms. The challenge with any tool for ML fairness is that the value of "fairness" is highly varied across disciplines and is context-dependent.[9] Consequently, existing ML fairness toolkits are either too broad or lack the necessary context to support comprehensive bias identification, mitigation, and stakeholder education in a given product vertical or industry segment.

A 2022 study, *Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits*[10] found that industry professionals use existing AI fairness toolkits inconsistently and encounter shortcomings that limit their usefulness. The study demonstrates that thoughtful evaluation of model bias requires a combination of technical assessments of models and datasets and an understanding of socio-cultural patterns to guide and justify why particular data treatments are necessary or should be avoided. The average ML practitioner is unlikely to know where to start, have expertise in these areas, the time to learn an open-ended toolkit to target a highly specific fairness question, and produce accessible explanations for non-technical decision makers.

Our project, Egaleco—meaning equity in Esperanto—seeks to fill these gaps in AI fairness toolkits. Egaleco identifies problematic disparity in machine learning model performance, teaches users the legal and ethical reasons why resolving it matters, and offers practical advice for how to advance fairness in existing workflows. For this project we chose to narrow our scope to evaluating fairness at the group level, meaning models are assessed on how similar its outcomes are for different groups of people. Our demo test dataset demonstrates this through

[4] (Obermeyer et al., 2019)
[5] (Straw et al., 2022)
[6] (Ross, 2023)
[7] (Microsoft Research, 2023)
[8] (Human-Computer Interaction Institute, 2023)
[9] (Mulligan, et al., 2019)
[10] (Deng, et al., 2022)

the use of COVID-19 data and an analysis of the disparate outcomes across race and sex categories.

Project Team

Egaleco is the result of a collaborative effort between eleven Masters students at the UC Berkeley School of Information. Our four-person Product and Policy team collaborated closely with four graduate students from the MIDS program. A separately advised UX team of three MIMS colleagues were also instrumental in conducting design research. All teams worked together through biweekly all-hands meetings, exchanging deliverables for feedback, and integrated each other's contributions into our respective products.

On our Product and Policy Team, Mudit Mangal led development of the ML evaluation model and creation of fairness visualizations, Gurpreet Kaur Khalsa led back-end and front-end architecture implementation (in collaboration with MIDS teammates) and served as the product manager, and Orissa Rose and Alan Kyle led usability research and the development of educational content throughout the tool as well as a white paper, best practices document, and policy spreadsheet that appear on the Resources page of the tool.

In addition to conducting research, the UX team designed our first MVP as well as an idealized prototype with the same core product scope but without the constraints of our engineering resources this semester. The idealized prototype serves as an inspiration for future development.

Methods

In this section we detail the mixed-methods research and development approach of our team, presented in order of occurrence.

Literature Review

Our literature survey revealed divergent approaches to fairness. Beyond Deng et. al's paper about the ways practitioners try (and fail) to use existing fairness toolkits, our understanding of fairness in technical systems was heavily shaped by Mulligan et al's *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*. The authors frame four approaches to fairness (Quantitative, Law, Social Science, Philosophy) and provide a fairness analytic that informed our approach to crafting probing questions for tool users and developing white paper content.

We synthesized that framework into three areas to guide our literature review:

1. Quantitative fairness → how do data scientists talk about fairness work?
2. Legal fairness → what laws define how and for whom fairness must be realized?

3. Social fairness → how have Science and Technology Studies (STS) scholars problematized these prescriptive approaches to fairness? What other angles must be considered when attempting to build responsible AI?

Each team member was responsible for sourcing and annotating different papers, all of which were added to our literature repository with a tag indicating which team would most benefit from a review. Findings from the papers that discussed fairness metrics, accuracy trade-offs and visualization techniques informed the data work described in the Tool Development section. Literature exploring the many ways we must challenge and re-define fairness informed our White Paper and are summarized there.

Policy Survey

From the beginning, we knew it would be a good idea to think about the policy implications of AI in US healthcare. The US healthcare sector is a highly regulated environment, and the prospect of AI regulation is more present than ever. It stands to reason then that an AI evaluation tool for healthcare should at least be informed by the public policy landscape. To start thinking about ways we might incorporate public policy details into Egaleco, we conducted an in-depth survey of existing and proposed Healthcare and AI policies. We also included prominent AI ethics frameworks to aid in our own conceptualizing of fairness in AI. The output of this effort became the Legal and Ethical Frameworks Spreadsheet, which ultimately became one of the resources linked in the toolkit.

While this spreadsheet is by no means exhaustive, it did reinforce the need to educate machine learning practitioners in healthcare about the many ways their work is regulated and fit within ethics frameworks.

Comparative Analysis of Existing Fairness Toolkits

To understand trends of existing fairness toolkits we surveyed the marketplace and selected 10 popular and commonly cited toolkits for a systematic review of attributes.

1. Fiddler
2. AI Fairness 360 (AIF360)
3. Arize
4. Aequitas
5. Fairlearn
6. Audit AI
7. FairModels
8. Responsible AI Toolbox (Microsoft)
9. TruEra
10. What-if

Each tool was evaluated through these research questions:

- Who was this built for?
- How are fairness concepts taught or offered to users?
- Does the product's design and functionality support an easeful fairness assessment?
- Is the code open-source or closed-source?
- What shortcomings of this tool are things we want to address in Egaleco?

Trends we noted:

- *The majority of tools require data science training.* A minority of the toolkits were B2B tools marketed as off-the-shelf fairness tools. If a company were to purchase a use license for one of these tools, they would need someone with a data science background on staff to understand the process. All other tools besides Google's "What-if" tool were built in a way that limits their usability to an experienced data scientist.

- *There is no standard for educating users about fairness and when informational content does appear it is siloed or sparse*. The amount of educational policy and fairness content varied widely across tools. When it was present it was typically posted in a separate section from the metrics sections, meaning users would have to self-motivate to learn more. This observation was a core inspiration in our product scoping. We wanted to embed language and features in the tool that would turn data scientists into fairness advocates without requiring them to exit the assessment process.

- *Open versus closed source coding was mixed*. Closed source fairness tools like Arize don't share their code repository publicly, while Fiddler shares no part of its tool. Other toolkits like Fairlearn, AIF360, and Aequitas are open source with the tool easily accessible to the public. By granting access to the code and user documentation, these open source tools made it easier for us to understand the capabilities of these tools in-depth. This was important because in some cases just the tool itself didn't provide that additional level of detail which was needed to properly understand the implementation of a method or technique. Similarly, without the code and any documentation for the closed source tools like Arize some aspects were ambiguous and hard to evaluate in their entirety.

- *Educational guidance absent in "code interfaces"*. Many toolkits like Fairlearn and AIF360 provide tutorial python notebooks that serve as the primary guide for users seeking to conduct any sort of fairness analysis. This format inherently limits the accessibility of these tools to users with technical acumen. If any non-code guidance was provided, it was a separate asset that did lend itself well to quick use or comprehension.

Semi-structured Interviews with ML professionals and fairness experts

Through a collaborative needfinding and interview process with the UX team, we conducted a total of twelve 1-hour long semi-structured interviews with two categories of people. The first are four fairness experts, people with extensive experience assessing fairness in ML. The second are eight ML practitioners, people with some exposure to ML but limited or functional understanding of ML fairness.

Our interviews revealed that the greatest challenge of the project would be to provide a sufficiently satisfying assessment that engages users while also introducing sufficient friction for them to grapple with a topic as complex as fairness. Although the perspectives of our participants were varied, the recurring themes were:

- Need for a common fairness vocabulary
- The importance of thinking about model performance in terms of the people impacted (instead of just thinking about it like numbers or statistics)
- A final product or resource that tool users can bring to their teammates

Privacy Impact Assessment

To map potential privacy risks to users of the tool and data subjects, we conducted a privacy impact assessment (PIA) with Jared Maslin, Sr. Director of Data Privacy at Good Research and instructor at UC Berkeley. When working with US health-related data, it is possible that Egaleco users are subject to domestic health data regulations (e.g., HIPAA) and/or consumer data regulations (e.g. CPRA). The discussion section addresses the practices employed to guide users data upload behaviors.

Mixed-Methods Usability Testing

Usability testing for Egaleco assets happened in multiple ways throughout the semester. The first category of usability testing was targeted towards target user feedback on tool prototypes, with a specific focus on the flow of the assessment, instruction formulation and the wording of questions. The second category of usability focused on fairness visualizations with a specific focus on displaying multiple fairness disparities simultaneously.

Category 1 Usability Testing

In consultation with our team and based on the wireframes we created, the UX team conducted two rounds of usability testing for the first category of inquiry. They tested two versions, an idealized prototype that includes design elements beyond our engineering capability, and an implementable version that was more realistic. For the first round they showed interviewees both versions to compare and contrast them for further iteration. While the second round of testing focused only on the idealized version, our team was able to include updates to language from the first round of testing and got valuable feedback in both rounds.

The UX team recruited novice ML practitioners with limited ML fairness exposure who resemble the target user base for Egaelco—ML practitioners who know enough about ML fairness to seek a resource to address it but need assistance to use fairness metrics effectively. Five ML practitioners participated in each round of usability testing.

We got encouraging signs of improvement to the tool's language between testing rounds. In response to the first round, we made changes by paring back formal and dense language, clarified privacy considerations, and shared more reasoning behind the questions the tool asks users. After the second round we didn't see any of these concerns repeated by interviewees.

Category 2 Usability Testing

The product and policy team led this round of usability testing, which focused on novel decision tree representations and fairness disparity and intersectionality charts according to demographic subgroups. Testing was conducted with graduate students who self-identified as data scientists and have been working and/or studying the field for 2-4 years. Testing involved pre-and post visualization interaction questioning, task observation, semi-structured interviews, a demographic survey, and a Likert-style evaluation. Findings are presented in the Discussion section.

Tool Development

The team chose to focus on healthcare because the life or death implications of ML in the medical field felt like an area where Egaleco would be most impactful. To innovate on existing fairness evaluation models we sought to enliven decision tree logic through visualizations and present users with a collection of applicable fairness metrics, instead of the singular metrics that most other tools offer. In this section we detail some of the major tasks we carried out to build the Egaleco web application and policy tools.

Data, Metrics and Processes

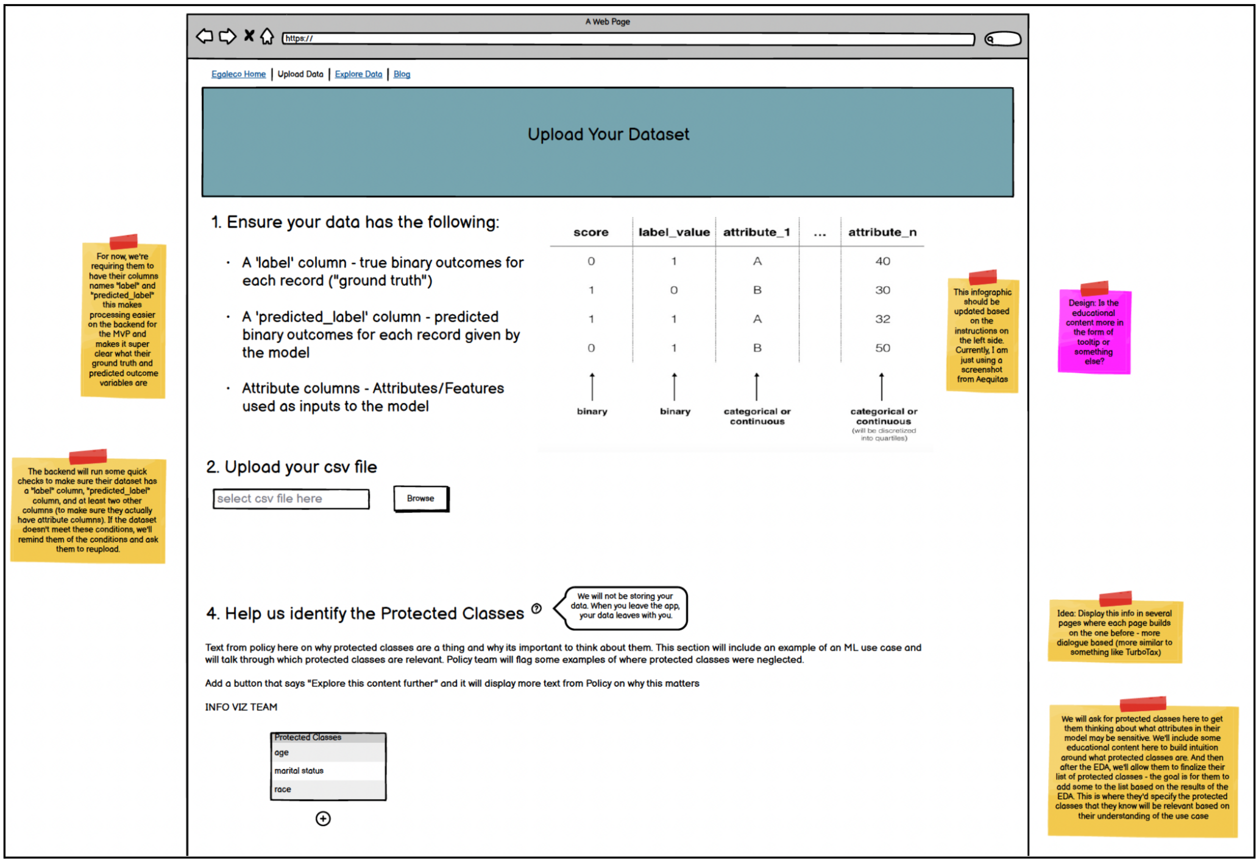
- Holistic Fairness Metric Research – Informed by our literature review, comparative tool assessment, and qualitative interviews with ML practitioners, we developed a list of the various fairness metrics we wanted to research (shown below). Fairness metrics offer different lenses through which to examine different kinds of fairness across groups and are a major feature of any fairness toolkit. A glossary defining all of these fairness metrics and their conditions of use is provided in the Appendix.

Category	**Description**	**Metrics**
Group Metrics	These metrics measure disparity between different groups based on the actual values and/or predicted values (see Appendix A). Generic ratio and/or difference will be used to compare two groups. For example, we'll show TPR Parity between the Reference Group and each of the other groups (if there's 5 groups they'll see 4 TPR Parities).	TPR Parity, FPR Parity, TNR Parity, FNR Parity, PPV Parity, FOR Parity, FDR Parity, NPV Parity, Demographic Parity, Accuracy Parity
Aggregated Group Metrics	These are metrics that aggregate two group metrics (or require that both metrics have parity in order to satisfy this group metric). For example, for Predictive Value Parity we will show them PPV Parity and NPV Parity and explain that to satisfy Predictive Value Parity, both PPV Parity and NPV Parity have to be 0. Generic ratio and/or difference will be used to compare two groups.	Equalized Odds, Predictive Value Parity, ABROCA (will also require users to upload predicted probabilities)

Individual Fairness Metrics	Measures the notion that people similar with respect to the classification task be treated similarly	Generalised Entropy Index
Meta Metrics	These metrics that summarize between group disparities that the group metrics capture, when there are more than two groups, including intersectional groups. These will be applied over the group metrics. For example, if there's 5 subgroups, 'max-min difference' will be the difference between the highest TPR and the lowest TPR.	Max-Min Ratio, Max-Min Difference, Variance
Intersectionality Metrics	We'll first ask users which two protected classes they want to look at intersectionality for; then we'll show them the reference groups they had selected for those protected classes and ask if they want to change them. We'll then calculate their group metrics: if white and male are the reference groups, then we'll show TPR Parity for white male vs white female, white male vs black male, and white male vs black female. Only allowed to run intersectionality once. Also run the meta metrics for these. Generic ratio and/or difference will be used to compare two groups.	Group Metrics, Aggregated Group Metrics, Meta Metrics applied on subgroups formed by combining two protected variables i.e. subgroups like (male, white); (male, black) etc

- Selection of Egaleco's fairness metrics - When it came time to finalize the metrics that Egaleco would include, we excluded individual fairness metrics and some of the aggregated group metrics. This decision is addressed further in our discussion section. The fairness metrics that are functional in this version of Egaleco are:
    - *Group metrics:* TPR Parity, FPR Parity, TNR Parity, FNR Parity, PPV Parity, FOR Parity, FDR Parity, NPV Parity, Demographic Parity, Accuracy Parity
    - *Aggregated Group Metrics:* Equalized Odds
    - *Meta Metrics:* Max-Min Difference, Variance
    - *Intersectionality Metrics:* Group Metrics and Meta Metrics applied on subgroups formed by combining two protected variables.
- Mapping critical stages of the fairness audit process (the steps users take in the tool). This guided the development of the wireframes and the first set of UX prototypes.

- Wireframing – Our balsamiq wireframe (pictured) allowed policy and UX teammates to comment on functionality and suggest design elements and educational content elements. This wireframe was what the UX team used to generate their first prototype for usability testing. Sticky notes shown in the image below display feedback and suggestions made by our cross-functional team.



Upload Your Dataset

1. Ensure your data has the following:

- A 'label' column - true binary outcomes for each record ("ground truth")
- A 'predicted_label' column - predicted binary outcomes for each record given by the model
- Attribute columns - Attributes/Features used as inputs to the model

score	label_value	attribute_1	...	attribute_n
0	1	A		40
1	0	B		30
1	1	A		32
0	1	B		50
binary	binary	categorical or continuous		categorical or continuous (will be discretized into quartiles)

2. Upload your csv file

select csv file here

Browse

4. Help us identify the Protected Classes ⓘ

We will not be storing your data. When you leave the app, your data leaves with you.

Text from policy here on why protected classes are a thing and why its important to think about them. This section will include an example of an ML use case and will talk through which protected classes are relevant. Policy team will flag some examples of where protected classes were neglected.

Add a button that says "Explore this content further" and it will display more text from Policy on why this matters

INFO VIZ TEAM

Protected Classes
age
marital status
race

⊕

For now, we're requiring them to have their columns names "label" and "predicted_label" this makes processing easier on the backend for the MVP and makes it super clear what their ground truth and predicted outcome variables are

The backend will run some quick checks to make sure their dataset has a "label" column, "predicted_label" column, and at least two other columns (to make sure they actually have attribute columns). If the dataset doesn't meet these conditions, we'll remind them of the conditions and ask them to reupload.

This infographic should be updated based on the instructions on the left side. Currently, I am just using a screenshot from Aequitas

Design: Is the educational content more in the form of tooltip or something else?

Idea: Display this info in several pages where each page builds on the one before - more dialogue based (more similar to something like TurboTax)

We will ask for protected classes here to get them thinking about what attributes in their model may be sensitive. We'll include some educational content here to build intuition around what protected classes are. And then after the EDA, we'll allow them to finalize their list of protected classes - the goal is for them to add some to the list based on the results of the EDA. This is where they'd specify the protected classes that they know will be relevant based on their understanding of the use case

- Create Data Exploration Tab – While guiding the user through a comprehensive EDA on their uploaded dataset was beyond the scope of what we could build in this version, we did build a mini-EDA. The data exploration sections include “number of unique categories”, “% of missing observations”, “least/most frequent category”, “Skewness”, “Outliers” etc. We explored ways to incorporate statistical tests like the Chi-Square Test of Independence, Levene’s Test and other metrics like KL Divergence and JS Divergence. These tests and additional metrics are useful for measuring bias in the testing dataset but incorporating them into the current iteration of Egaleco proved too complex, thus we’ve marked it as a feature for implementation in future versions of the tool.
- Formulate End to End Demos on chosen datasets. We identified a set of datasets from around the healthcare space which had specific usage in a ML Healthcare application. Then, we created python notebooks to run our proposed EDA process and applied a few

fairness metrics to identify bias in line with our fairness audit process. This helped us identify loopholes in our process, gave us more grounding and familiarity with the fairness metrics and finally helped us replicate and modularize the pipeline for our web app.

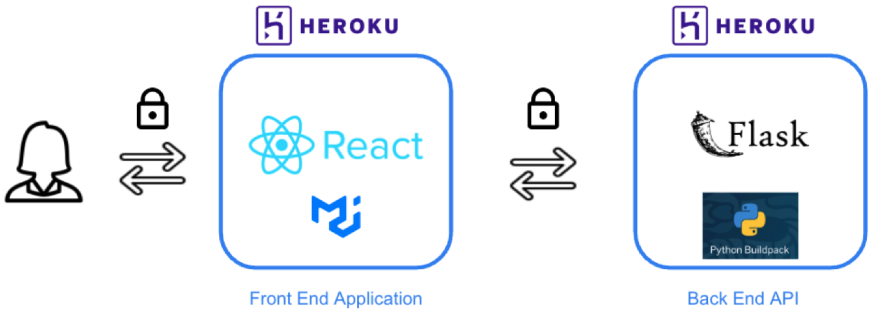
- Design decision flow of fairness metrics – We wanted to design a decision flow that helped users choose the relevant set of fairness metrics for their use case in a dynamic way, as opposed to showing them static decision tree logic. To create what is now the interactive question series we began by mapping the range of paths a given user could travel through the decision tree depending on their answers. We then formulated that logic into a series of questions that would be embedded in the scrolling visualization.

- Visualizing Fairness Metrics – Product and Policy worked together on visualizations in Figma and D3 to depict the flow of the fairness metrics section (described above) and help users choose a fairness metric. Leveraging progressive disclosure, scrollytelling and multiple-choice questions, our visualizations educate users on key concepts of fair ML to help them determine which ML fairness metric(s) would be appropriate to evaluate their use case. This choice was made in response to Deng et al.'s findings and expert interviewer feedback that fairness nomenclature in toolkits was often unintuitive to participants. We hope the innovative presentation of decision tree logic and interactivity of the charts will engage users more than static imagery or a code-only interface.

System Architecture

Egaleco is a full-stack application that employs authentication tokens to guarantee secure access to the tool. Additionally, all communication between the client and the server is routed through HTTPS, creating an encrypted channel that ensures the confidentiality and integrity of the data being transmitted.

The user flow begins with the user accessing our front-end website, which creates a user session. Once logged in, the user can upload their data, which is securely transmitted to the server and is temporarily stored on our backend application. To protect the user's sensitive health-related data, their session data is permanently destroyed upon exiting the application.



To provide a seamless user experience, we used React.js, the MUI React component library, and hosted our app on Heroku. Our Python and Flask backend manages information transfer between metric calculations and user interface.

API calls are made to conditionally display information from the user's uploaded dataset. Upon providing all necessary inputs, fairness metrics are calculated in the backend. Functionality is further explained in the forthcoming User Journey section. Future iterations of Egaleco will move towards a microservices architecture to allow greater scalability and flexibility.

Policy Assets

From the interviews with experts and literature surveys, we knew that to promote the use of Egaleco we had to improve the way fairness work gets prioritized and advocated for within organizations. Between this goal and the increased regulatory scrutiny of AI, we believed that incorporating educational elements of policy and best practices for responsible ML into Egaleco we could realize the following benefits:

1) Provide guidance for practitioners in designing and deploying fair ML products that are sensitive to market and consumer expectations as well as make it easier for people to take ownership of fairness work.

2) Give ML practitioners the language of public policy so they can be more persuasive with cross-functional colleagues and decision makers. Increased awareness of potential liability can raise the level of priority given to fairness matters.

3) Increased awareness of regulatory trends and non-quantitative understanding of fairness will help practitioners anticipate and build for the future.

A first iteration of our tool attempted to include notes and graphics about relevant laws and forthcoming policy alongside the recommended fairness metrics. But we discovered that the inputs that lead to a metric being recommended are quite different than the inputs needed to surface relevant laws. The broad complexity of healthcare and AI laws was simply not generalizable in the logic of our tool.

This spurred us to look for ways to incorporate policy and ethics outside of Egaleco's direct workflow. The result is a collection of resources housed in the top right Resources section. They are:

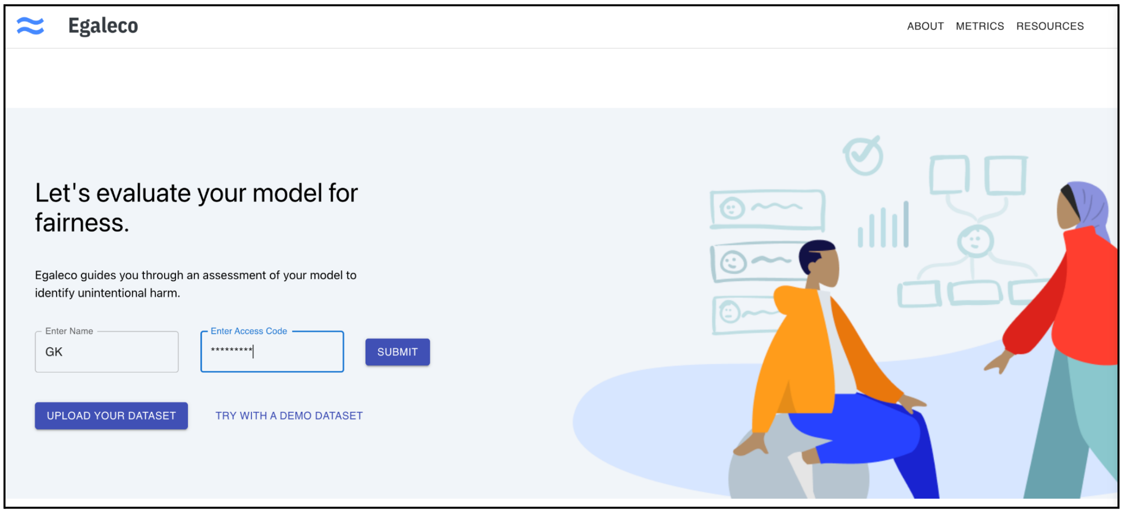
- A Legal and Ethical Frameworks spreadsheet. A curated, non-exhaustive collection of current and proposed policies and ethics frameworks for US healthcare. Users can learn about these different forms of guidance at a glance through the spreadsheet's notes.

- A white paper titled, *Looking Beyond Quantitative Fairness to Build Responsible AI Systems*. It provides a framework for thinking critically about ML fairness and promoting organizational buy-in for work that is commonly deprioritized.

- A Best Practices for Responsible ML guide that gives concrete steps ML practitioners can take to level up their efforts to mitigate algorithmic harms.

User Journey Overview

Egaleco came together as a result of the efforts detailed above. Below, we share the instructions that users see in the seven-steps of Egaleco's model assessment process. Users who don't have their own test dataset can still see these features in action and in context by opting to use Egaleco's demo test dataset on the introductory page.



Let's evaluate your model for fairness.

Egaleco guides you through an assessment of your model to identify unintentional harm.

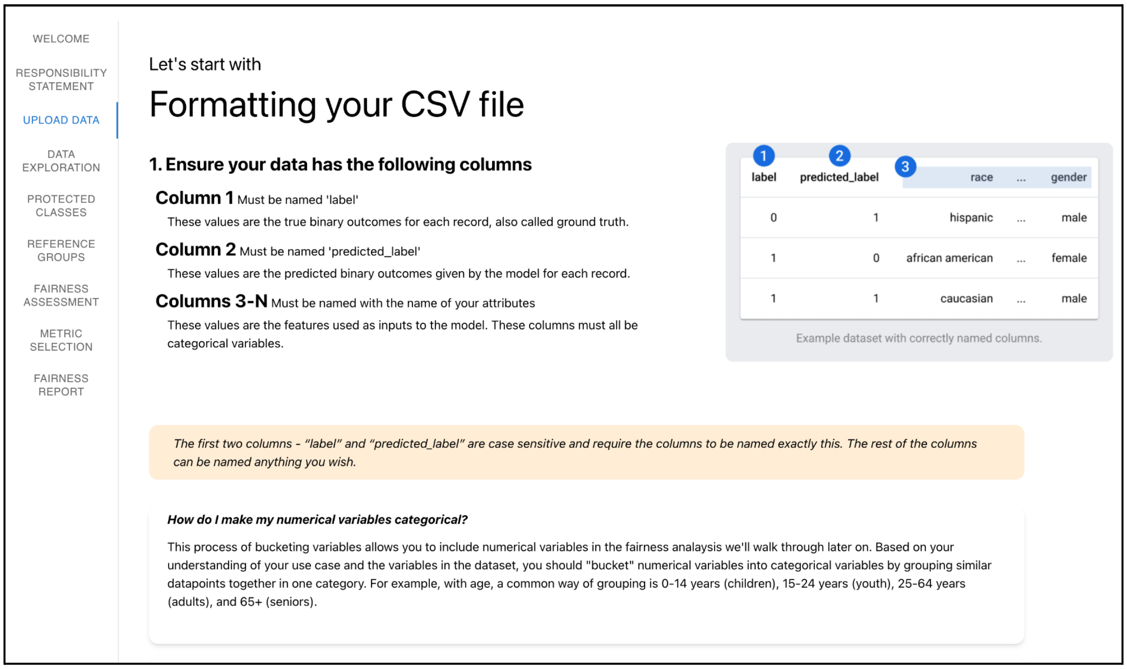
Enter Name: GK

Enter Access Code: •••••••••

1. **Upload Test Dataset:** There are a few pre-processing steps that you need to check for before using Egaleco. The tool prompts you to check that your test dataset includes the label (ground truth), predicted label (what your model predicted), and the predictor variables (features). Keep in mind that Egaleco currently only supports classification models with binary outcome variables. Check that the CSV file to be used matches the required formatting convention, using proper column names and data types.

   A note on privacy and security: When you upload your test dataset it is passed to our backend application and stored there temporarily. There is an authentication token so that our React and Flask apps can communicate securely. Finally, once you exit the application your session data is destroyed so that we aren't retaining any sensitive health related data.

If you are using your own test dataset, and not our demo dataset, you will first be shown a prompt to consider the implications of your data use and ML project.



Let's start with

Formatting your CSV file

1. Ensure your data has the following columns

Column 1 Must be named 'label'

These values are the true binary outcomes for each record, also called ground truth.

Column 2 Must be named 'predicted_label'

These values are the predicted binary outcomes given by the model for each record.

Columns 3-N Must be named with the name of your attributes

These values are the features used as inputs to the model. These columns must all be categorical variables.

1 label	2 predicted_label	3 race	...	gender
0	1	hispanic	...	male
1	0	african american	...	female
1	1	caucasian	...	male

Example dataset with correctly named columns.

The first two columns - "label" and "predicted_label" are case sensitive and require the columns to be named exactly this. The rest of the columns can be named anything you wish.

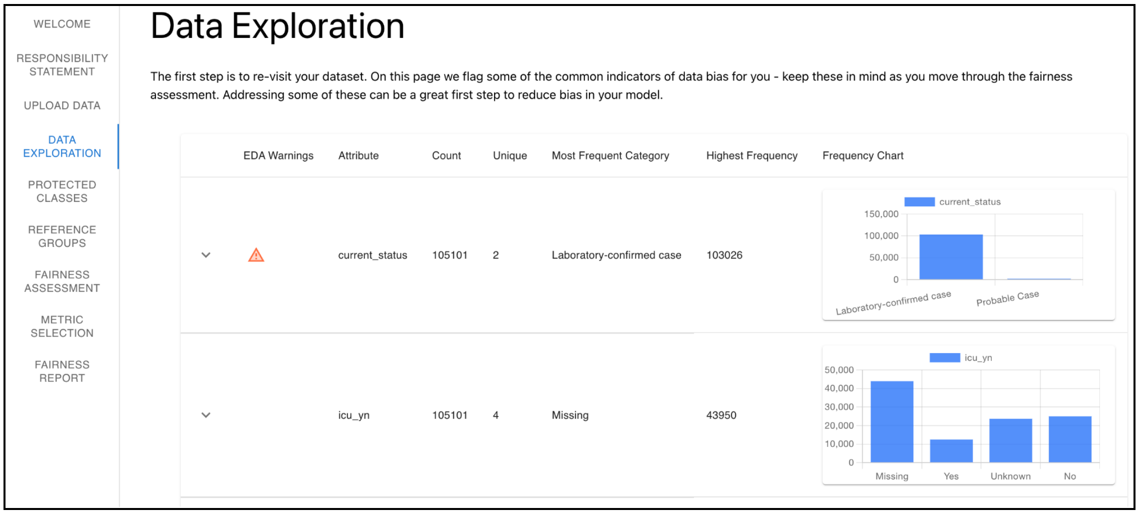
How do I make my numerical variables categorical?

This process of bucketing variables allows you to include numerical variables in the fairness analaysis we'll walk through later on. Based on your understanding of your use case and the variables in the dataset, you should "bucket" numerical variables into categorical variables by grouping similar datapoints together in one category. For example, with age, a common way of grouping is 0-14 years (children), 15-24 years (youth), 25-64 years (adults), and 65+ (seniors).

2. **Data Exploration:** This step raises potential distortions in the test data, which may contribute to model bias. For example, it may flag if a variable makes up a large proportion of the total distribution in the dataset, which may be something you choose to address in your training dataset.

Look for an orange warning triangle where the toolkit has identified a common indicator of bias and click on it for more information.

We're not able to embed a complete Exploratory Data Analysis (EDA) process into Egaleco at this stage, so we recommend making a healthsheet as a complement to tool use. You can learn more about this in our best practices resource.



Data Exploration

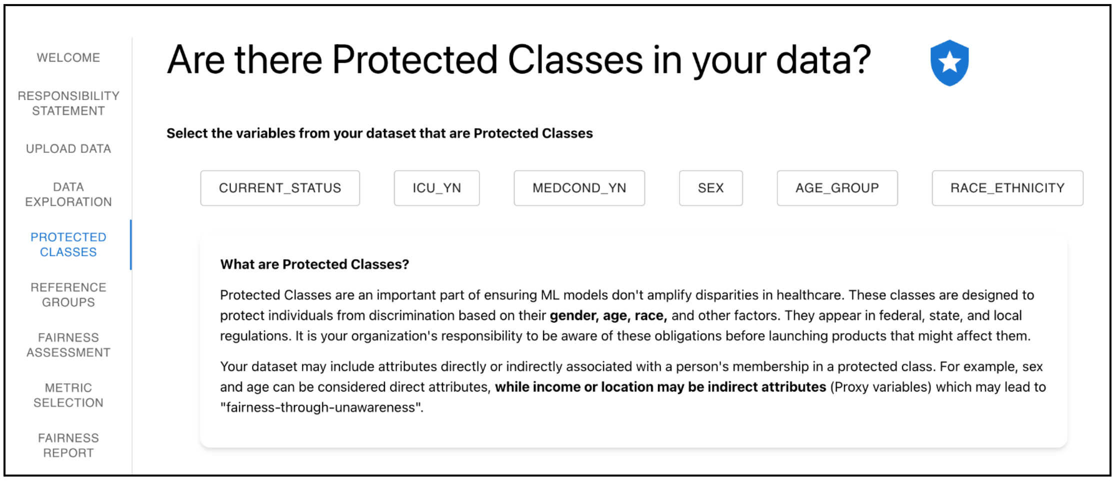
The first step is to re-visit your dataset. On this page we flag some of the common indicators of data bias for you - keep these in mind as you move through the fairness assessment. Addressing some of these can be a great first step to reduce bias in your model.

	EDA Warnings	Attribute	Count	Unique	Most Frequent Category	Highest Frequency	Frequency Chart
⌄	⚠	current_status	105101	2	Laboratory-confirmed case	103026	
⌄		icu_yn	105101	4	Missing	43950	

3. **Select Protected Classes:** Protected Classes are an important part of ensuring ML models don't amplify disparities in healthcare. They are designed to legally protect individuals from discrimination based on their gender, age, race, and other factors. It is important to select Protected Classes in order to tell Egaleco which variables to analyze fairness for.

   Although Egaleco provides limited guidance for proxy variables, they should be selected at this stage. This is a good opportunity to conduct research about how your variables are interrelated or seek expert consultation. For example, protected class attributes that often get disguised within a proxy are:

- Geographic location as a proxy for race (due to histories of redlining and segregation)
- Education as a proxy for race (due to histories of schooling and residential segregation)
- Insurance type as a proxy for disability status or medical condition
- Preferred Language as a proxy for national origin



Are there Protected Classes in your data?

Select the variables from your dataset that are Protected Classes

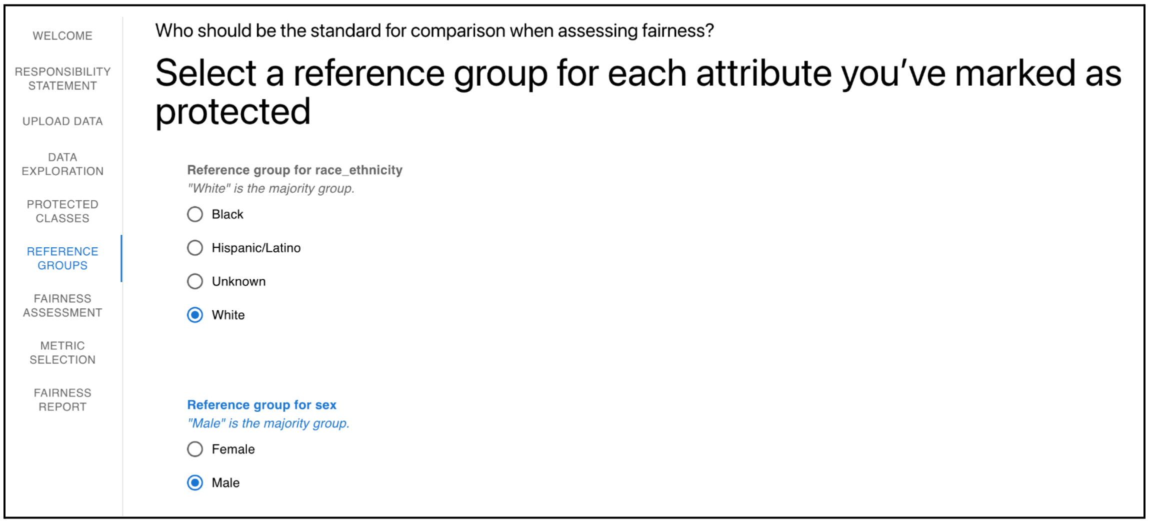
What are Protected Classes?

Protected Classes are an important part of ensuring ML models don't amplify disparities in healthcare. These classes are designed to protect individuals from discrimination based on their **gender, age, race,** and other factors. They appear in federal, state, and local regulations. It is your organization's responsibility to be aware of these obligations before launching products that might affect them.

Your dataset may include attributes directly or indirectly associated with a person's membership in a protected class. For example, sex and age can be considered direct attributes, **while income or location may be indirect attributes** (Proxy variables) which may lead to "fairness-through-unawareness".

4. **Select Reference Groups:** The reference group refers to a group that is used as a baseline or comparison for evaluating fairness in the algorithm's outputs.

   In your use case, if it is obvious that one group has consistently had more opportunities/resources as compared to other groups, it may be advisable to select that group as the reference group. In this situation, fairness would be seen as uplifting some groups by bringing the "performance" of the model for these groups closer to that of the reference group.



Who should be the standard for comparison when assessing fairness?

Select a reference group for each attribute you've marked as protected

Reference group for race_ethnicity

"White" is the majority group.

○ Black

○ Hispanic/Latino

○ Unknown

● White

Reference group for sex

"Male" is the majority group.

○ Female

● Male

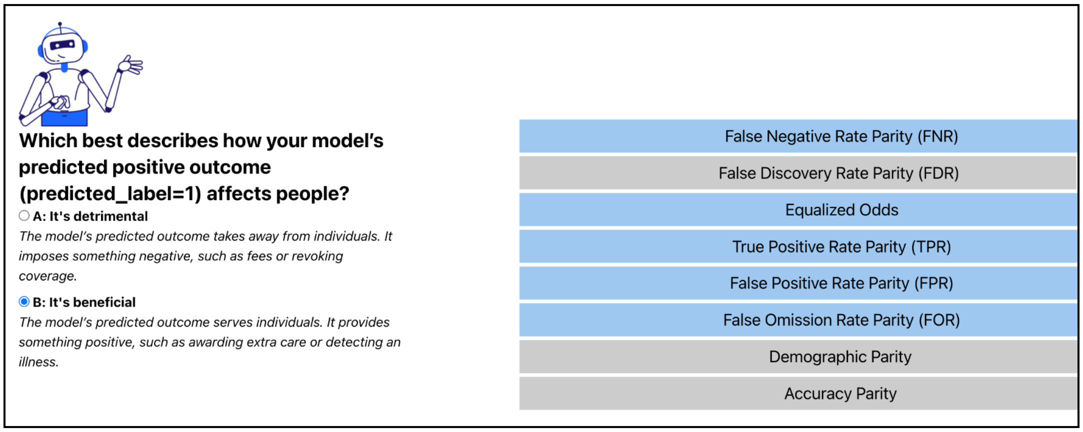
At the moment, our tool does not help users choose their reference group. Selecting reference groups is context-dependent and isn't always desirable or possible. Egaleco's demo dataset uses White data subjects as the reference group to benchmark other groups against because they were the most represented in the CDCs COVID-19 dataset.

5. **Fairness Assessment:** Here we provide information about how you might conceptualize fairness in your use case and select a representation of fairness to run in the assessment. The two options provided for this are:

   1) Fairness based on a group's representation in a given population; or

   2) Fairness based on how the model's errors adversely harm or give opportunity to one group over another.

   Then you will select if your intervention is punitive or assistive. In other words, the question could be framed as, is your algorithm providing care or a benefit or removing a care or benefit?



Which best describes how your model's predicted positive outcome (predicted_label=1) affects people?

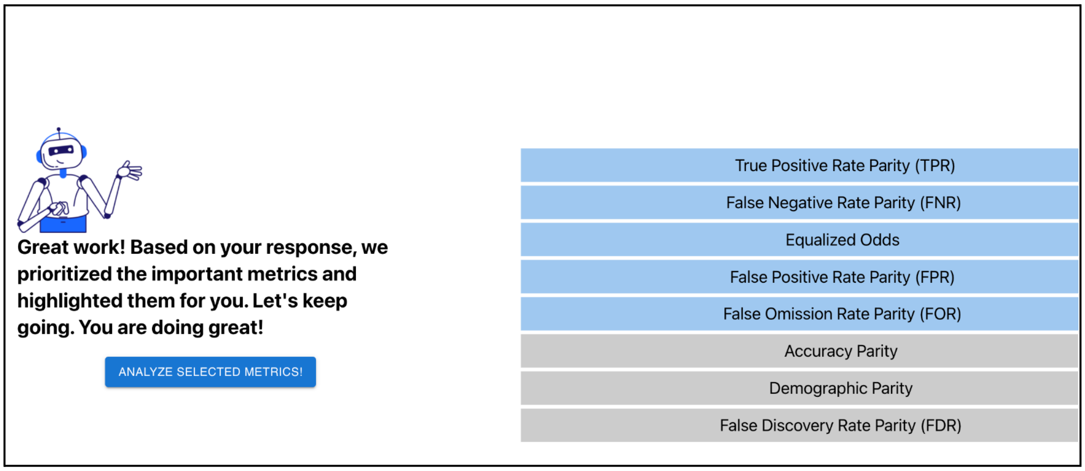
○ **A: It's detrimental**

The model's predicted outcome takes away from individuals. It imposes something negative, such as fees or revoking coverage.

● **B: It's beneficial**

The model's predicted outcome serves individuals. It provides something positive, such as awarding extra care or detecting an illness.

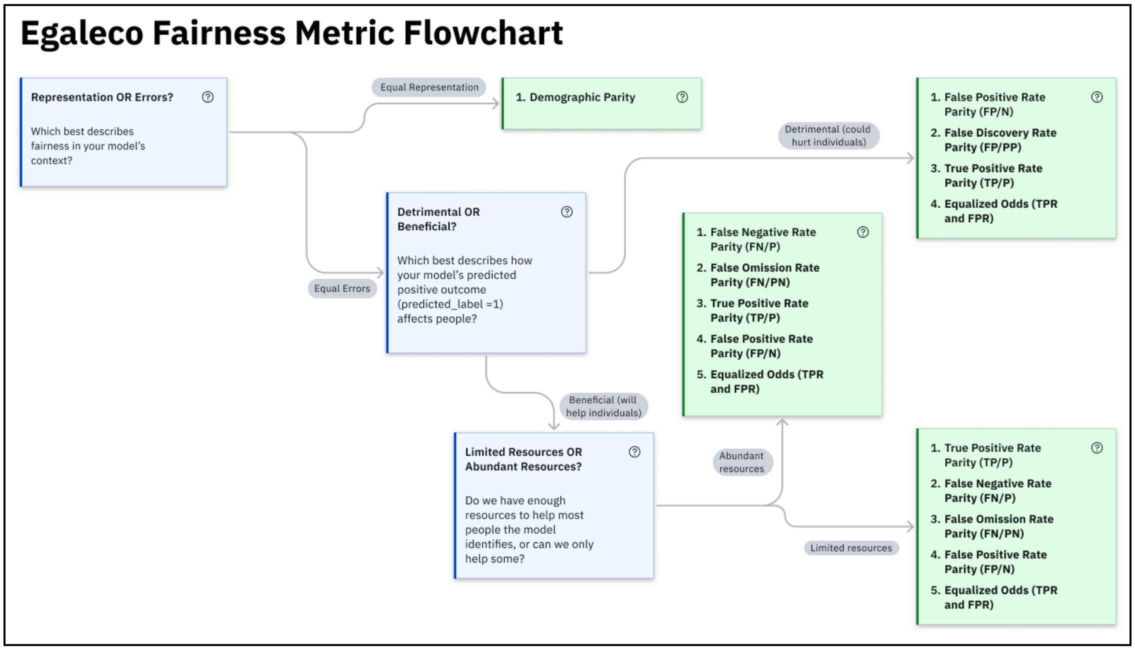
False Negative Rate Parity (FNR)
False Discovery Rate Parity (FDR)
Equalized Odds
True Positive Rate Parity (TPR)
False Positive Rate Parity (FPR)
False Omission Rate Parity (FOR)
Demographic Parity
Accuracy Parity

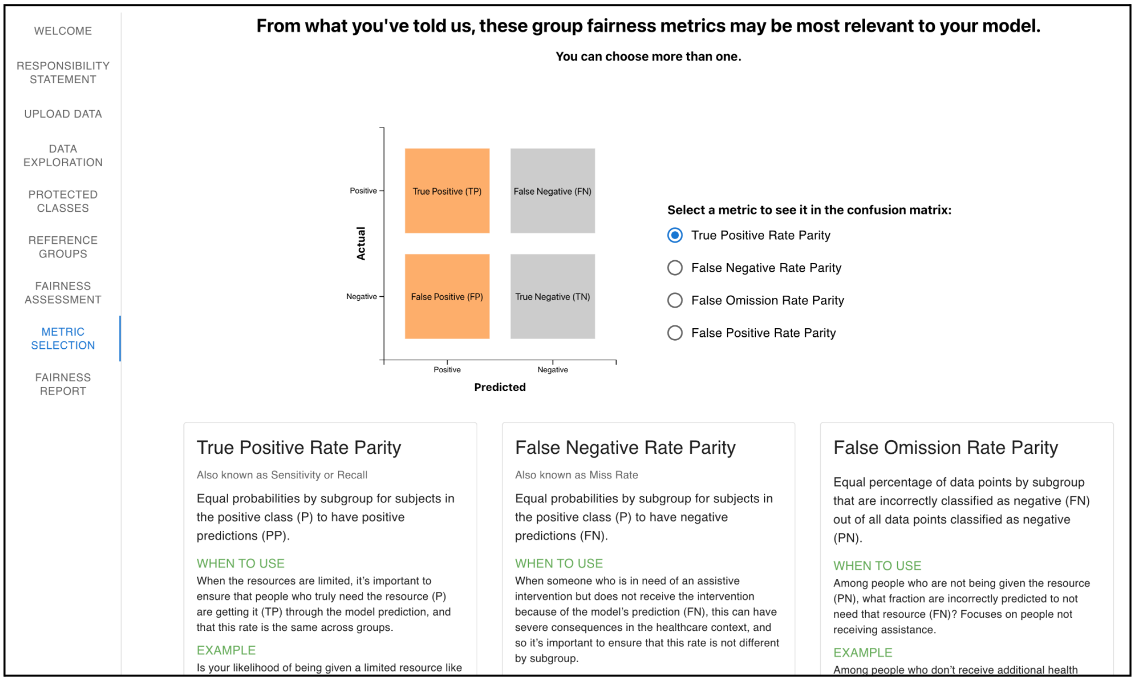


Great work! Based on your response, we prioritized the important metrics and highlighted them for you. Let's keep going. You are doing great!

True Positive Rate Parity (TPR)
False Negative Rate Parity (FNR)
Equalized Odds
False Positive Rate Parity (FPR)
False Omission Rate Parity (FOR)
Accuracy Parity
Demographic Parity
False Discovery Rate Parity (FDR)

6. **Metric Selection:** In this section we ask you some questions that help us filter for the most relevant fairness metrics. We use decision tree logic from a modified version of the Aequitas fairness toolkit and your answers to questions posed by the AI bot to identify the metrics most appropriate to your use case. The flowchart below shows the pathways for how fairness metrics are surfaced through the question and answer scrolly interaction. Note that the results don't produce a hard ranking, but rather surface metrics that may be most relevant to your project. You can always click the *see more* button to view information about all the metrics.


Egaleco Fairness Metric Flowchart
Representation OR Errors?
Which best describes fairness in your model's context?
Equal Representation
1. Demographic Parity
Equal Errors
Detrimental OR Beneficial?
Which best describes how your model's predicted positive outcome (predicted_label =1) affects people?
Detrimental (could hurt individuals)
1. False Positive Rate Parity (FP/N)
2. False Discovery Rate Parity (FP/PP)
3. True Positive Rate Parity (TP/P)
4. Equalized Odds (TPR and FPR)
Beneficial (will help individuals)
Limited Resources OR Abundant Resources?
Do we have enough resources to help most people the model identifies, or can we only help some?
Abundant resources
1. False Negative Rate Parity (FN/P)
2. False Omission Rate Parity (FN/PN)
3. True Positive Rate Parity (TP/P)
4. False Positive Rate Parity (FP/N)
5. Equalized Odds (TPR and FPR)
Limited resources
1. True Positive Rate Parity (TP/P)
2. False Negative Rate Parity (FN/P)
3. False Omission Rate Parity (FN/PN)
4. False Positive Rate Parity (FP/N)
5. Equalized Odds (TPR and FPR)



From what you've told us, these group fairness metrics may be most relevant to your model.

You can choose more than one.

Select a metric to see it in the confusion matrix:

● True Positive Rate Parity

○ False Negative Rate Parity

○ False Omission Rate Parity

○ False Positive Rate Parity

True Positive Rate Parity

Also known as Sensitivity or Recall

Equal probabilities by subgroup for subjects in the positive class (P) to have positive predictions (PP).

WHEN TO USE

When the resources are limited, it's important to ensure that people who truly need the resource (P) are getting it (TP) through the model prediction, and that this rate is the same across groups.

EXAMPLE

Is your likelihood of being given a limited resource like

False Negative Rate Parity

Also known as Miss Rate

Equal probabilities by subgroup for subjects in the positive class (P) to have negative predictions (FN).

WHEN TO USE

When someone who is in need of an assistive intervention but does not receive the intervention because of the model's prediction (FN), this can have severe consequences in the healthcare context, and so it's important to ensure that this rate is not different by subgroup.

False Omission Rate Parity

Equal percentage of data points by subgroup that are incorrectly classified as negative (FN) out of all data points classified as negative (PN).

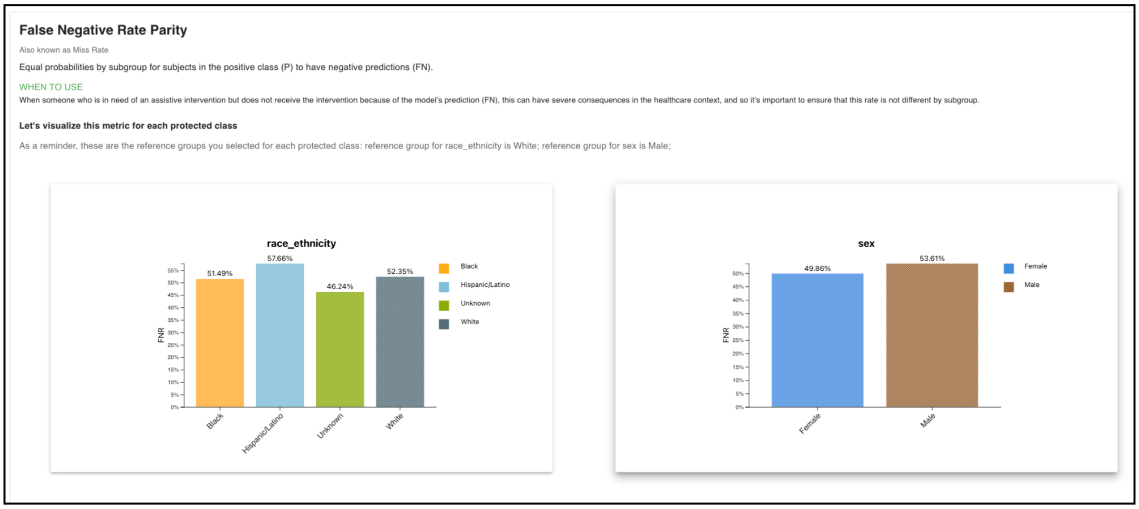
WHEN TO USE

Among people who are not being given the resource (PN), what fraction are incorrectly predicted to not need that resource (FN)? Focuses on people not receiving assistance.

EXAMPLE

Among people who don't receive additional health

7. Fairness Report: This final section provides information about selected fairness metrics including a description, explanation of when it's helpful to use them, and visualizations of the metrics for each protected class.



False Negative Rate Parity

Also known as Miss Rate

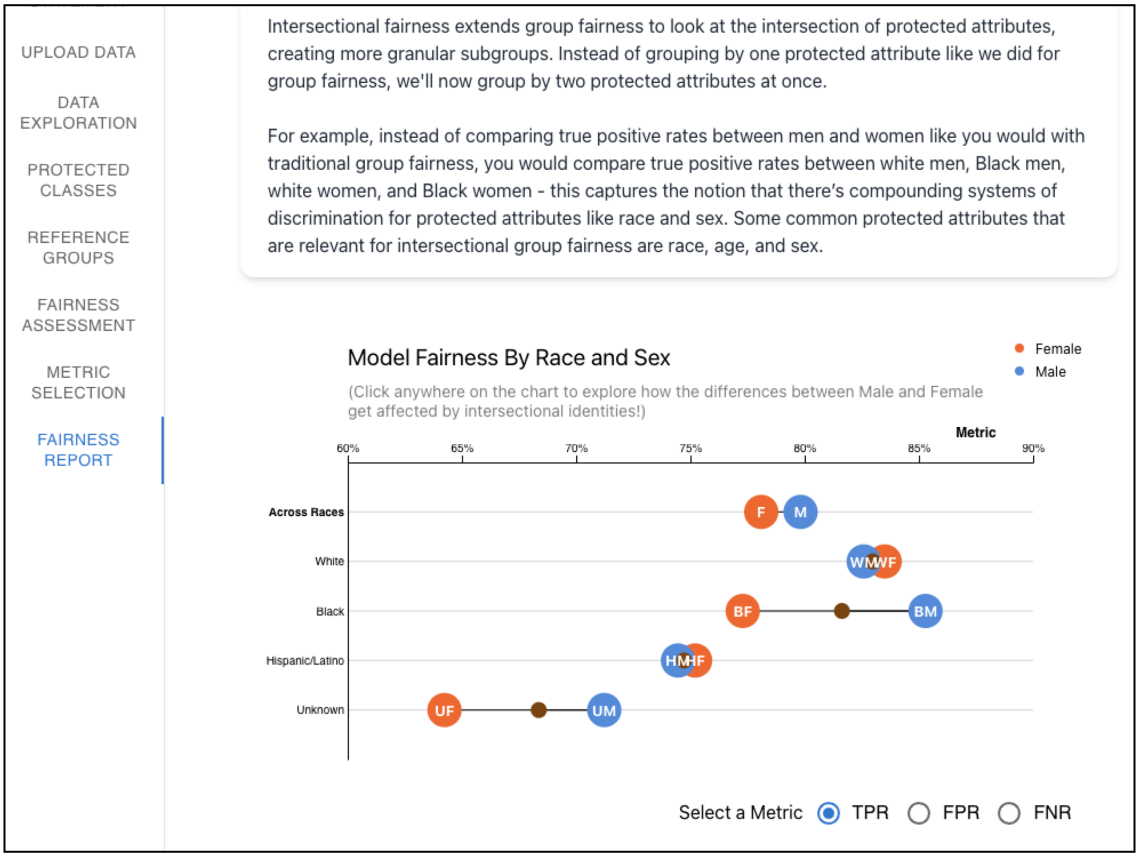
Equal probabilities by subgroup for subjects in the positive class (P) to have negative predictions (FN).

WHEN TO USE

When someone who is in need of an assistive intervention but does not receive the intervention because of the model's prediction (FN), this can have severe consequences in the healthcare context, and so it's important to ensure that this rate is not different by subgroup.

Let's visualize this metric for each protected class

As a reminder, these are the reference groups you selected for each protected class: reference group for race_ethnicity is White; reference group for sex is Male;



Intersectional fairness extends group fairness to look at the intersection of protected attributes, creating more granular subgroups. Instead of grouping by one protected attribute like we did for group fairness, we'll now group by two protected attributes at once.

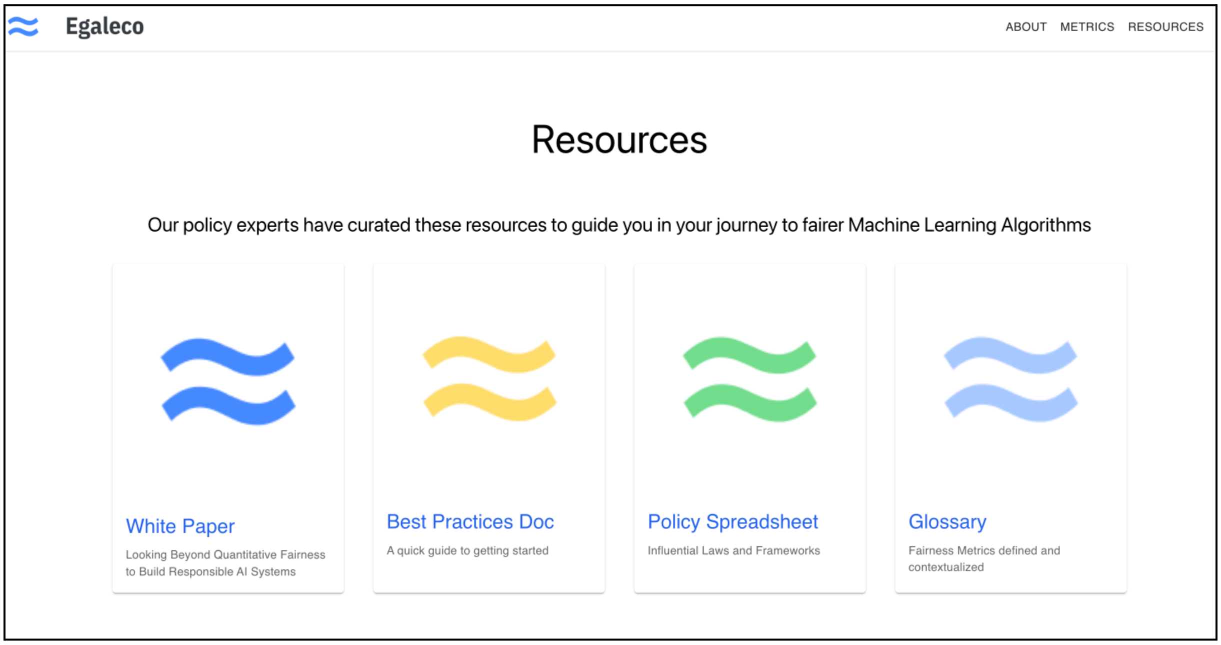
For example, instead of comparing true positive rates between men and women like you would with traditional group fairness, you would compare true positive rates between white men, Black men, white women, and Black women - this captures the notion that there's compounding systems of discrimination for protected attributes like race and sex. Some common protected attributes that are relevant for intersectional group fairness are race, age, and sex.

Model Fairness By Race and Sex

(Click anywhere on the chart to explore how the differences between Male and Female get affected by intersectional identities!)

Select a Metric ● TPR ○ FPR ○ FNR

At the conclusion of the assessment users are prompted to visit the Resources page. The assets on the Resources page are all designed to help users put assessment results into action.



Resources

Our policy experts have curated these resources to guide you in your journey to fairer Machine Learning Algorithms

White Paper
Looking Beyond Quantitative Fairness to Build Responsible AI Systems

Best Practices Doc
A quick guide to getting started

Policy Spreadsheet
Influential Laws and Frameworks

Glossary
Fairness Metrics defined and contextualized

Discussion

Decisions Taken Consciously

Throughout the semester, many meetings were spent discussing the objectives of our tool and how to balance competing priorities. This section narrates the decisions we made very intentionally about what this version of Egaleco does and does not do.

- This iteration of the Egaleco does not address some measures of aggregated group fairness, individual fairness and counterfactual fairness. We decided to focus on group fairness metrics because group fairness is often the most popular and the first type of fairness evaluated in assessment efforts – e.g. is this model unfair to females or to the black group? It is also easier to scale across larger impacted populations or datasets. Defining individual fairness for different healthcare contexts requires more context, specifically on how to define the notion of similarity among individuals. We also didn't implement aggregated group metrics like ABROCA because it required additional user input in the form of probabilities; we will pursue this in a future version of the tool. Other measures like Entropy Indices were not implemented in the current version of the tool because we discovered that they correlated to other measures and are not actively considered useful by the ML fairness community.

- This iteration of the Egaleco only examines classification models with binary outcome variables. This means that dataset variables can only be categorized as one of two options, for example: at-risk/no-risk; or positive/negative. The limitation of binary classification models is that they risk oversimplifying conditions that actually occur over a spectrum of outcomes. We talk to users more about the implications of this in the White Paper.

- Egaleco does not promise to fix models for users. Tweaking users' proprietary models in order to mitigate identified unfairness is not a feature we enabled in this version of Egaleco. We advise users that bias mitigation is the next step after evaluation that may involve modifying the training data to reduce bias or adjusting the algorithm to account for different subgroups in a population. Mitigations may also involve creating transparency and accountability mechanisms like the ones detailed in our Best Practices for ML document to ensure that the ML system is operating fairly. We encourage users to repeatedly revise and run their models through Egaleco to see how particular adjustments are advancing fairness efforts.

- This iteration of the Egaleco does not assist users in the identification of all of the relevant proxy variables. We made this choice because most fairness experts we interviewed were uncomfortable identifying proxies without additional time for research and consultation. External consultation is especially critical when working outside of one's own context. Thus we chose to put the responsibility of proxy consideration and selection on the user so they take it seriously. One risk of improper proxy identification is "unfairness through unawareness". Another risk is that proxy choice makes a subjective value judgment.

- Egaleco does not evaluate the existence of label choice bias. Educating data scientists who are working with training dataset and building models on the importance of thoughtful data labels (assigning "true value" labels) is a key step in minimizing algorithmic bias. For example, when we use healthcare costs as a proxy for healthcare need, it values the people who receive care (hence cost) not those who need care (which is often disconnected from whether or not someone ill is treated promptly, or at all). We are exploring ways to integrate related questioning into future versions of the tool. For the moment, we surface label choice bias as an area warranting attention in the Best Practices guide.

- Egaleco is not explicit about every circumstance in which it's best to use a certain fairness metric. We made the conscious decision not to be prescriptive about a singular, catch-all metric because ML fairness is very complex and context-sensitive. Different fairness metrics may prioritize different types of fairness, and it may not be possible to optimize for all of them simultaneously. The fairness report provided to users at the conclusion of the assessment suggests the 5 most relevant fairness metrics based on what we've learned about the user's model. We believe providing multiple is the responsible thing to do because considering multiple metrics is an important part of investigating the different dimensions of ML fairness.

- Egaleco's Best Practice recommendations were developed in response to interview feedback that there is potential for ML fairness to be overlooked in larger companies because of a lack of clarity, ownership and stewardship. We also heard that one's capacity to explore fairness issues are often constrained by time, so we kept the best practices brief. The document is not meant to be a definitive guide to responsible ML but a starting point for people to begin their fairness journey.
- Our data governance solution was to keep the data on the server for the duration of the Egaleco assessment, and delete it once the assessment is complete. In future iterations of the tool, we will attempt to not store any data on our servers since we know that information privacy is a significant concern and determinant of user adoption.
- Egaleco's demo uses a COVID-19 dataset sourced from the CDC, for which we could not create in-depth datasheets. There is a risk that we are unaware of any biases that are introduced during the data collection, cleaning, and processing. We chose this dataset despite those risks because COVID-19 is a subject everyone is familiar with, which lessens comprehension barriers for people new to the concept of ML fairness assessments.

Lessons Learned

1) ***Not all fairness issues can be addressed through binary question formulation.***

   Our tool cannot cover all the different combinations of fairness questions. Nor can our tool resolve all the fairness debates that might happen around a given product or across teams.

2) ***Measures of fairness can be at odds with each other.***

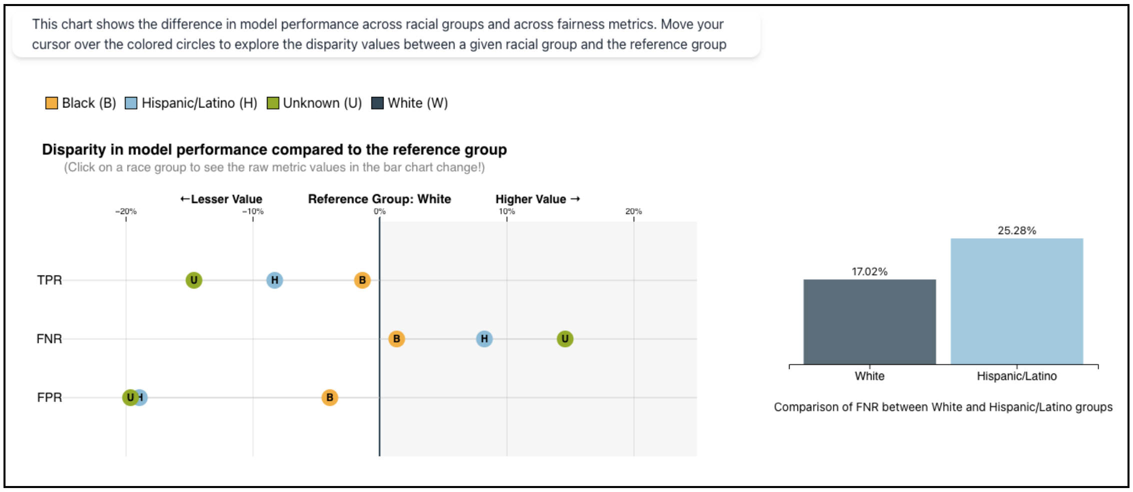
   Many times, you cannot satisfy all metrics simultaneously, you have to make tradeoffs.

3) ***It is critical to define fairness in the context of one's work but it is very hard to develop a definitive definition***.

   The meaning of fairness changes according to the situation and impacted stakeholders. Our interviews with fairness experts and ML practitioners revealed this as well as our attempts to prescribe a fairness solution to tool users when we lacked detailed information about their use case.

4) ***There are many nuances of contemplating fairness in a healthcare context that make it impossible to build a one size fits all approach, and very challenging to clearly represent through visualizations***.

The below charts demonstrate some of these complexities. Firstly, whether or not a metric carries a negative or positive association depends on the use case and type of intervention. Accordingly, when displaying multiple fairness metrics simultaneously, the overall fairness implications cannot be bulked together and defined as positive or negative. In the case of our demo data COVID-19 resource allocation model, a higher rate of false positives (FPR) for White patients means they are most likely to be given access to limited resources unnecessarily. In this same scenario, a lower rate of false negatives (FNR) for White patients means that the model is most accurate at predicting risk of death, and thus need for resources, for them. The direction of the disparity percentages are oppositional but the overall significance is improved performance for White patients over all others.



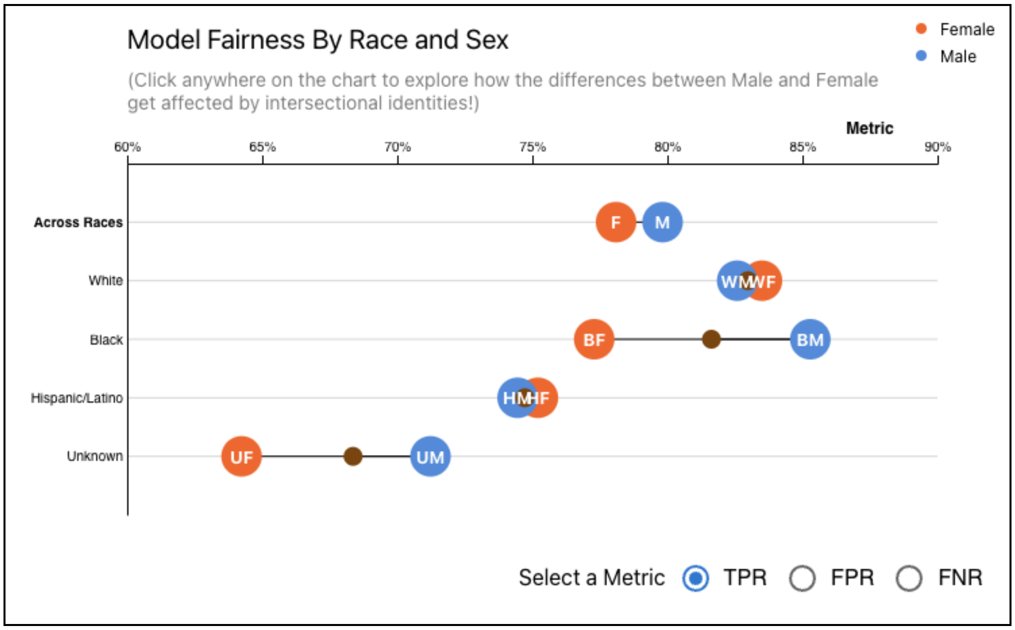
This chart shows the difference in model performance across racial groups and across fairness metrics. Move your cursor over the colored circles to explore the disparity values between a given racial group and the reference group

Disparity in model performance compared to the reference group

(Click on a race group to see the raw metric values in the bar chart change!)

Comparison of FNR between White and Hispanic/Latino groups

Secondly, there are a lot of reasons why a given model performs much more poorly on certain intersectional groups that cannot be explained through visualizations alone. The Intersectional Disparity Dot Plot (below) uses animation to show users how the fairness of a model changes when you consider the intersection of two protected attributes at once, e.g. race and sex. When a given intersectional group is small (e.g. few data points) the ML models naturally fall short in performing well on them. This makes it difficult to address poor performance via technical solutions like parameter adjustment and feature amendment because what is really needed is more and better quality training data. Providing explanations like this within the visualization would create a lot of visual clutter, hence their omission. We are considering separate, complementary visualizations or text features that would communicate this information for future iterations of the tool.


Model Fairness By Race and Sex

(Click anywhere on the chart to explore how the differences between Male and Female get affected by intersectional identities!)

Select a Metric ● TPR ○ FPR ○ FNR

5) It's hard to effectively incorporate substantive educational content into assessment tools.

Our comparative analysis of other tools revealed that policy language and fairness education content is often missing or siloed from the metrics sections of tools. Accordingly, our policy team developed longer narrative sections and examples about fairness to be embedded within the metric interface as tooltips, which are examples that appear on mouseover, and as introductory copy for each new page or tool frame. However, the first round of usability testing with our prototype revealed that users found this to be information overload or they ignored it. This is what led to the creation of the Resources page that tool users could reference before and after applying their metric assessments.

6) There are many types of fairness a tool cannot evaluate.

To ensure that Egaleco does not endorse the notion that fairness can be measured entirely through binary classifications and measures of statistical parity, we created the accompanying Resources page on the website. Furthermore, we hope the layering of modeling techniques and provision of narratives explanations alongside fairness metrics will provoke users' interest in pursuing quantitative and qualitative meanings of fairness.

7) Technical expertise does not translate to rapid comprehension of ML fairness subjects.

Our interviews were conducted with ML experts and our usability testing was conducted with experienced data scientists, yet the core decision tree logic and themes of fairness were confusing even after we'd distilled them. This suggests that quantitative tooling is not enough and future innovations must find a way to engage users with theoretical, ethical and conversational explorations of what fairness means in practice.

8) Practitioners bring their own biases to metric selection.

It takes time to change or inspire a nuanced fairness awareness because some see fairness as a matter of legal compliance while others have a preferred set of quantitative metrics they always use. Both approaches prevent practitioners from considering the full spectrum of fairness metrics that are valuable to consider when seeking to mitigate algorithmic harms.

9) The privacy risks of our tool change significantly depending on user behaviors.

While Egaleco is built with healthcare in mind, which has its own privacy sensitivities, there's nothing preventing users from uploading a CSV with other kinds of sensitive information. The privacy impact assessment identified that the most critical determining factors for regulatory and legal exposure will initially stem from the following variables:

1. Data subject populations included in uploaded data
2. Potential for persistent storage raises liability
3. Accuracy or reliability of platform outputs, metrics, and/or guidance
4. Data types/categories included in uploaded data

We are trying to get ahead of improper data usage by limiting tool access to those with a verified access code. Future versions of Egaleco will need a carefully crafted privacy notice and terms and conditions to prevent legal exposure.

10) It's challenging to balance the competing interests of product, policy, and design teams.

One of the central organizing factors for our project was that several School of Information students with a diversity of specializations saw the importance of this work and found ways to make a meaningful contribution. Indeed, the three broad areas of product, policy, and design all have important implications for algorithmic fairness in healthcare. But this rich topic also led to some challenges when it came to planning and executing such a complicated project. Some we found were:

- A tension between balancing rich educational content from the policy team and the design team principles that seek a brevity and casualness atypical of legal and STS literature.

- Logistical issues of waiting on one another, or proceeding to build some tool elements because of time constraints only to realize later on that more effort was required to reunite them.

- Knowledge sharing across domains was time consuming (but worthwhile!).

Features for the Future

The nuanced nature and scope of building an AI fairness tool means there are features we couldn't build during the course of this project. The following list shares features and activities we next steps for building Egaleco into the future:

- *Offer mitigation strategies* – Enable data scientists to implement bias mitigation techniques directly into their model within the tool interface.
- *Broaden types of model evaluations* – Future versions of the tool will conduct fairness evaluation for a regression model.
- *Evaluate user's training dataset for bias* – Future versions of the tool would give users a specific measure for training data bias before beginning the model assessment.
- *Embed Datasheets and Model Cards within the tool* – We currently recommend use of these resources to users but would like to embed the creation of these assets as a part of the tool.
- *Construct a fairness report suitable for sharing with non-technical team members.* We imagine this report will include various visualizations and language from the white paper that explains nuanced fairness concepts in more approachable language.
- *Emphasize the humans behind the data* – Embed real world healthcare examples and leverage iconography.
- *Enable dynamic educational features* – Our literary survey on the ways that people best learn and retain information found that friction (desirable difficulties) and problem-based learning (interactivity) are two highly effective educational methods. Future iterations of the tool will explore how we can use these methods in combination with visual, verbal and text content to deliver nuanced information about ML fairness.
- *Enable PDF download of Fairness Report* – This will allow tool users to easily save their work and share it with colleagues.
- *Refine decision tree questions* – Distilling the decision tree questions into laymen prose that feel accessible to unfamiliar users is an ongoing challenge. We spent a significant amount of time rewording and testing different iterations of the questions, however usability testing revealed that data scientists were still having trouble making the appropriate selection for the scenario we'd tasked them with. We plan to continue testing different words and soliciting expert user feedback.
- *Additional Visualization Testing* – Future work will expand the visualization usability testing we conducted to further explore the efficacy of including fairness visualizations alongside decision tree questions.

- *White paper workflow piloting* – The policy team members would like to test the utility of the whitepaper in an industry setting to understand barriers to adoption.

Conclusion

In the time since we began working on Egaleco, there has been a significant increase in mainstream interest in AI and in attempts to insert algorithmic decision making into many facets of daily life. We want to be a part of developing and deploying tooling that supports thoughtful and responsible technology use. Accordingly, we intend to continue developing Egaleco for healthcare, as well as explore other industry-specific applications.

Acknowledgements

We are grateful to our advisors Professor Daniel Aranki, PhD and Professor Marti Hearst, PhD for their guidance, feedback and support. We are also grateful to the data scientists and ML experts that volunteered their time for our interviews.

References

Deng, W. H., Nagireddy, M., Lee, M. S. A., Singh, J., Wu, Z. S., Holstein, K., & Zhu, H. (2022). Exploring how machine learning practitioners (try to) use fairness toolkits. *arXiv Preprint arXiv:2205.06922*.

Fairness, accountability, transparency, and ethics (FATE). Retrieved 2023, from https://www.hcii.cmu.edu/research-areas/fairness-accountability-transparency-and-ethics-fate

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022). Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing,* , 1-28. 10.1007/s12652-021-03612-z

Mulligan, D., Kroll, J., Kohli, N., & Wong, R. (2019). This thing called fairness. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1-36. 10.1145/3359221

Microsoft Research.*FATE: Fairness, accountability, transparency & ethics in AI* . Retrieved 2023, from https://www.microsoft.com/en-us/research/theme/fate/

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science (American Association for the Advancement of Science), 366*(6464), 447-453. 10.1126/science.aax2342

Ross, C. (2023, March). Denied by AI: How medicare advantage plans use algorithms to cut off care for seniors in need. *Stat,* https://www.statnews.com/2023/03/13/medicare-advantage-plans-denial-artificial-intelligence/

Selbst, A., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (Jan 29, 2019). Fairness and abstraction in sociotechnical systems. Paper presented at the 59-68. 10.1145/3287560.3287598 http://dl.acm.org/citation.cfm?id=3287598

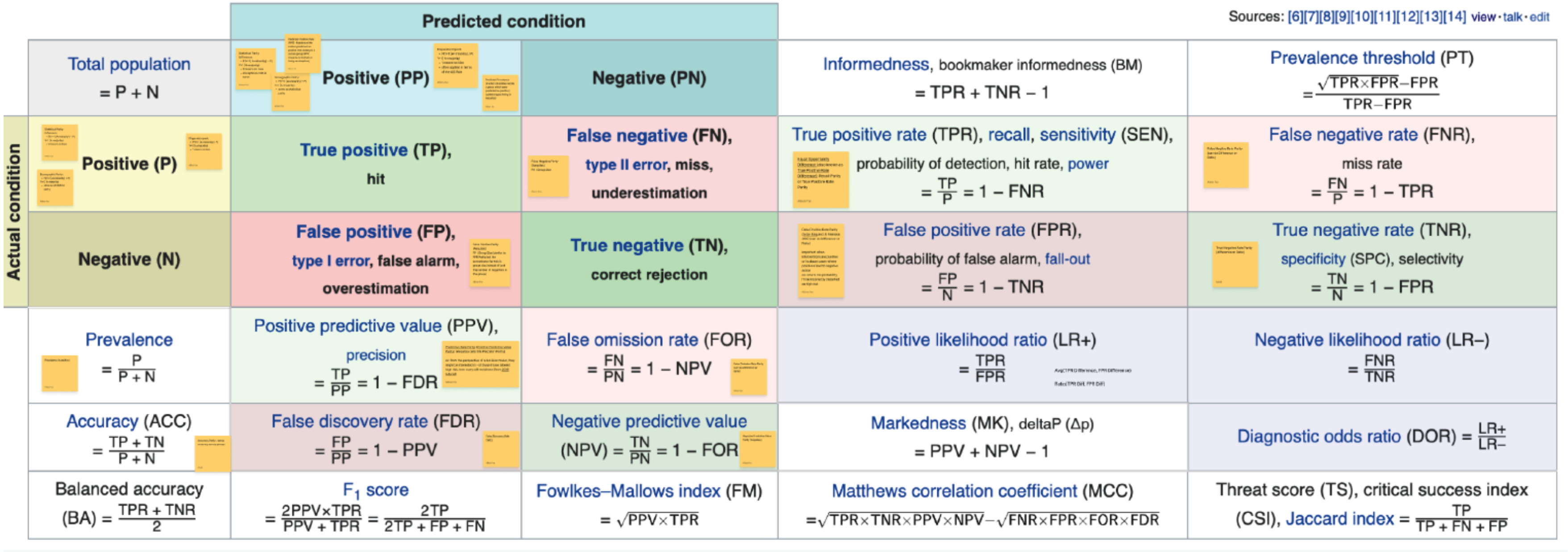
Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics, 29*(1), e100457. 10.1136/bmjhci-2021-100457

Ullah, R., Khan, S., Chaudhary, I. I., Shahzad, S., Ali, H., & Bilal, M. (2020). Cost effective and efficient screening of tuberculosis disease with raman spectroscopy and machine learning algorithms. *Photodiagnosis and Photodynamic Therapy, 32*, 101963. 10.1016/j.pdpdt.2020.101963

Woldaregay, A. Z., Årsand, E., Walderhaug, S., Albers, D., Mamykina, L., Botsis, T., & Hartvigsen, G. (2019). Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artificial Intelligence in Medicine, 98*, 109-134. 10.1016/j.artmed.2019.07.007

Appendix A

Confusion Matrix for Group Metrics



Sources: [6][7][8][9][10][11][12][13][14] view · talk · edit

		Predicted condition			
	Total population $= P + N$	Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$

	Prevalence
	$= \frac{P}{P + N}$
		Positive predictive value (PPV),
		precision
		$= \frac{TP}{PP} = 1 - FDR$
			False omission rate (FOR)
			$= \frac{FN}{PN} = 1 - NPV$
				Positive likelihood ratio (LR+)
				$= \frac{TPR}{FPR}$
				Avg(TPR Difference, FPR Difference)
				Ratio(TPR Diff, FPR Diff)
					Negative likelihood ratio (LR−)
					$= \frac{FNR}{TNR}$

	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Original image sourced from Wikipedia with additions (yellow sticky notes) made in Figma by the Egaleco team

Appendix B

Egaleco Fairness Metrics Glossary

Demographic Parity

- **Synonyms: Disparate Impact, Statistical Parity**
- **Definition**: Equal Selection Rate (Predicted Positives/Total Predictions) between two groups
- **When to Use**: When you care about the percentage of data points that are classified as positive, independent of ground truth. Don't use it when you know that positive rates justifiably differ among groups. E.g. when screening for Glaucoma (primarily affects the elderly), or when predicting likelihood of breast cancer (more likely in women).
- **Example Application:** *Does each group have equal opportunity of achieving a favorable outcome given their gender? An AI tool that screens for the flu and predicts that 2 of the 8 men (25%) will test positive and 3 of the 12 women (25%) will test positive is achieving demographic parity.*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want the probability of your predicted outcome to be equal across groups.

True Positive Rate (TP/P)

- **Synonyms: Sensitivity, Recall**
- **Definition**: Equal probabilities by subgroup for subjects in the positive class (P) to have positive predictions.
- **When to Use**: When the resources are limited, it's important to ensure that people who truly need the resource (P) are getting it (TP) through the model prediction, and that this rate is the same across groups.
- **Example Application:** Is your likelihood of being given a limited resource like a ventilator, given that you actually need the ventilator, dependent on your race? Black individuals who need a ventilator should be equally as likely to receive a ventilator as white individuals who need a ventilator.
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors give opportunity to one group over another, you have an assistive intervention, and resources are limited so you're only able to provide the intervention to a small fraction of the people who need the intervention.

False Positive Rate (FP/N)

- **Synonyms: False Alarm Rate**
- **Definition**: Equal probabilities by subgroup for subjects in the negative class (N) to have positive predictions (FP).
- **When to Use**: When people who don't need a punitive intervention are being subjected to such an intervention, it can cause harm to such individuals, so it's important to make

sure that this rate is not higher for one group than it is for another group. Also, when people who don't need an assistive intervention are mistakenly given some assistive resources, it can lead to wastage of those resources which can be undesirable in some contexts, so it's important to make sure that this rate is not higher for one group than it is for another group.

- **Example Application:** *Among the people who shouldn't have been charged higher premiums (N), what are the chances that they were incorrectly charged higher premiums (FP) given their age?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors adversely harm people, you have a punitive intervention, and you're most concerned about treating those who should not receive the intervention fairly by group.

False Negative Rate (FN/P)

- **Synonyms: Miss Rate**
- **Definition**: Equal probabilities by subgroup for subjects in the positive class (P) to have negative predictions (FN).
- **When to Use**: When someone who is in need of an assistive intervention but does not receive the intervention because of the model's prediction (FN), this can have severe consequences in the healthcare context, and so it's important to ensure that this rate is not different by subgroup.
- **Example Application:** *What are your chances of being wrongly left out of assistive resources like enrollment in a healthcare management program given your race?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors give opportunity to one group over another, you have an assistive intervention, you're able to intervene with most of the people with need, and you're most concerned about treating those with need fairly by group.

Equalized Odds (TPR and FPR)

- **Synonyms: Equality of Odds**
- **Definition**: Is satisfied if both True Positive Rate Parity and False Positive Rate Parity are satisfied.
- **When to Use**: When both allocating resources fairly by group (TPR) and preserving resources fairly by group (FPR) is important.
- **Example Application:** *Are the chances that an individual is moved into the intensive care unit when they are in need of it independent of their race, and are the chances that an individual is moved into the intensive care unit when they are not in need of it also independent of their race?*
- **When it's suggested:** Egaleco suggests this metric whenever FPR Parity or TPR Parity is suggested - equalized odds takes each of these metrics a step further and enforces a stricter definition of fairness because it requires parity for both FPR and TPR.

False Omission Rate (FN/PN)

- **Definition**: Equal percentage of data points by subgroup that are incorrectly classified as negative (FN) out of all data points classified as negative (PN).
- **When to Use**: Among people who are not being given the resource (PN), what fraction are incorrectly predicted to not need that resource (FN)? Focuses on people not receiving assistance.
- **Example Application:** *Among people who don't receive additional health services to prevent a stroke, what are the chances that they actually had a stroke and were in need of those additional health services? Is this rate the same for men as it is for women?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors give opportunity to one group over another, you have an assistive intervention, you're able to intervene with most of the people with need, and you're most concerned about treating those who don't receive assistance fairly by group.

False Discovery Rate (FP/PP)

- **Definition**: Equal percentage of data points by subgroup that are incorrectly classified as positive (FP) out of all data points classified as positive (PP).
- **When to Use**: When the intervention is punitive, it's important to ensure that the likelihood of incorrectly receiving the punitive intervention (FP) is the same for different groups.
- **Example Application:** *Among the people who are charged higher health insurance premiums, what are the chances they actually had higher health costs and should have been charged higher premiums, given their race?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors adversely harm people, you have a punitive intervention, and you're most concerned about treating those who receive the intervention fairly by group.

True Negative Rate (TN/N)

- **Synonyms: Specificity, Selectivity**
- **Definition**: Equal probabilities by subgroup for subjects in the negative class (N) to have negative predictions (PN).
- **When to Use**: When the resources are limited, it's important to ensure that people who truly don't need the resource (N) are not receiving the resource (TN) through the model prediction, and that this rate is the same across groups.
- **Example Application:** *Among healthy people, is the likelihood that they are correctly identified as not having a condition the same given their race?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but it can still be useful to explore if the benefit of a true negative is higher than the cost of a false positive.

Accuracy

- **Synonyms: Error Rate**
- **Definition**: Percent of correctly predicted data points out of all data points.
- **When to Use**: When you want to check if a model performance metric that accounts for all types of errors is equal between groups.
- **Example Application:** *If we have a model that predicts the likelihood of not arriving at a doctor's appointment, is the accuracy the same for older individuals as it is for younger individuals?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but it can be useful to explore alongside the other metrics Egaleco recommended for your use case, given its ease of interpretability.

Positive Predictive Value Parity (TP/PP)

- **Synonyms: Predictive Rate Parity, Precision**
- **Definition**: Checks if the Positive Predictive Value (True Positives divided by Predicted Positives) is equal between subgroups.
- **When to Use**: When you want to equalize the chance of success, given a positive prediction (success in this case is defined as correctly predicting someone as positive when they are indeed positive).
- **Example Application:** *Of the individuals the model predicted as being at high risk for heart disease (Predicted Positive), what percent of them actually had heart disease (True Positive)? Is this percent the same for white individuals as it is for black individuals?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but given that precision (which is the same as PPV Parity) is frequently used as a traditional model performance metric this metric may be useful to explore alongside the other metrics Egaleco recommended for your use case, given its ease of interpretability.

Negative Predictive Value Parity (TN/PN)

- **Definition**: Checks if the Negative Predictive Value (True Negatives divided by Predicted Negatives) is equal between subgroups.
- **When to Use**: When you want to equalize the chance of success, given a negative prediction (success in this case is defined as correctly predicting someone as negative when they are indeed negative).
- **Example Application:** *Of the individuals that the model predicted as not needing to be moved into an intensive care unit (Predicted Negative), what percent of them actually didn't need to be moved into an intensive care unit (True Negative)? Is this percentage the same for males as it is for females?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but if you are particularly concerned with ensuring that when the model makes a negative prediction the model is correct, this metric may be useful to explore alongside the other metrics Egaleco recommended for your use case.

Predictive Value Parity

- **Definition**: Is satisfied if both Positive Predictive Value Parity and Negative Predictive Value Parity are satisfied.
- **When to Use**: When equalizing the chance of success is important for both positive and negative predictions (correctly predicting someone as positive when they are indeed positive, and correctly predicting someone as negative when they are indeed negative).
- **Example Application:** *Say we have a model that predicts the severity of someone's condition in the emergency room. Is the positive predictive value (True Positives divided by Predicted Positives) the same for men as it is for women, and is the negative predictive value (True Negatives divided by Predicted Negatives) the same for men as it is for women?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but if you choose to explore PPV Parity or NPV Parity, then Predictive Value Parity can also be useful to explore because it combines both of these metrics to enforce a stricter definition of fairness.

References: Fairlearn, Aequitas; FairMLHealth; Verma 2018