

Introducing the FairnessLab

A better tool to audit your algorithms for bias

October 10, 2022

Authors: Joachim Baumann & Corinna Hertweck

TL;DR: In today's digitized world, biased algorithms take decisions in many areas of our lives. Auditing these systems is imperative and yet it is not an easy task given that fairness is not a well-defined mathematical equation, but a highly contextual and debated concept. In this post, we introduce the [FairnessLab](#): an implementation of a new theoretical approach for defining context-specific fairness metrics, consisting of a series of questions that need to be answered to dynamically derive a morally appropriate definition of fairness for the audited system. The theoretical approach and thus the FairnessLab alleviate important shortcomings of existing fairness metric. We showcase the tool's capabilities using an algorithm that is used in some regions of the US in the criminal justice system. We will show that previous analyses of this same tool that have relied on fairness metrics from the existing literature fall short: Minority groups have to be favored more than suggested by previous analyses in order to lessen the bias of the algorithm.

[Why we need a new bias audit tool](#)

[Theoretical foundations of the FairnessLab](#)

[Running an audit](#)

[Recidivism prediction with COMPAS](#)

[Dataset](#)

[The old and the new way of auditing COMPAS](#)

[Auditing COMPAS – the old way](#)

[Decision maker utility](#)

[Fairness score part I: Whose utility should be compared?](#)

[Fairness score part II: What is the utility of the decisions for the decision subjects?](#)

[Fairness score part 3: How should the utility be distributed between Black and white defendants?](#)

[Result](#)

[Auditing COMPAS – the new way 🎉](#)

[Conclusions 🍷🍷🍷](#)

[Open questions](#)

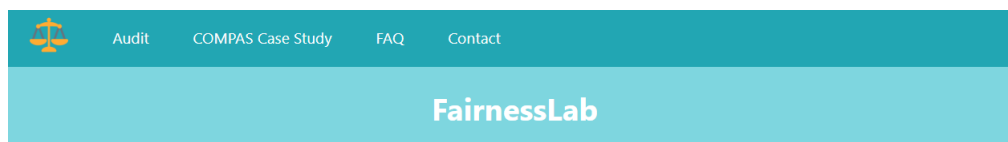
[Ethics statement](#)

[References](#)

[Resources](#)

Why we need a new bias audit tool

With the widespread use of algorithmic decision making systems, we have seen many instances in the last years where these systems were systematically biased against minority groups. Decision making systems are usually built to achieve a certain goal, which at the same time often happens to produce unfairness. As a result, people and institutions have called for audits of these systems to avoid unfair outcomes. Audit tools can support such audits. However, many of these audit tools are based on just a small set of fairness definitions — despite fairness being a highly debated and contextual concept. The fairness definitions that are often implemented in these tools are the so-called group fairness criteria. These criteria are fundamentally flawed as they are based on several assumptions, which are hardly ever met in practice. Fortunately, this does not mean that building a less biased system is a lost cause as we introduce a new audit tool, which no longer requires those assumptions to be true: the **FairnessLab!** 🎉



Don't just audit your decisions but their consequences

Welcome to the FairnessLab! This tool is intended to help you audit machine learning models. It takes a group fairness approach, meaning it audits a tool across socio-demographic groups. However, it is built to avoid some of the shortcomings of standard group fairness metrics. Specifically, it aims to help domain experts translate their knowledge into fairness metrics that uniquely fit the context in which the machine learning model is deployed. This sets it apart from existing tools for group fairness audits such as AIF360 or FairLearn. The FairnessLab is fit for both internal audits as well as external audits, but it is currently limited to auditing binary classifiers, so to models that are used to make binary decisions (e.g. whether an applicant will receive a loan or not). If your model is not a binary classifier, you might still be able to use this model if converting your data to a binary setting makes sense for your application or at least for its audit. [You can read more about that in the FAQ.](#)

Who is this for?

The tool is meant to be accessible to people with limited technical knowledge. We highly encourage everyone to use this tool to audit publicly accessible models. What you need for your own audit is mainly domain knowledge: You have to have a good understanding of the context in which the machine learning model is deployed to be able to define an appropriate way of measuring fairness in this context. The FairnessLab will guide you through this process of defining an appropriate fairness metric.

Want to see the FairnessLab in action?

You can find an audit of the well-known COMPAS dataset in the section [COMPAS Case Study](#). The dataset was published along with the ProPublica story "[Machine Bias](#)", that started a debate about how to evaluate the fairness of machine learning models. When you follow our audit of COMPAS in the FairnessLab, you will see that there are more ways to evaluate the fairness of the predictions than discussed by ProPublica (which audited COMPAS) and Northpointe (the company that developed COMPAS). Importantly, institutional racism might imply that the consequences of the predictions should be evaluated differently than standard group fairness metrics do. Going through this case study should thus both give you new insights into the COMPAS case and a better understanding of what the FairnessLab is capable of.

Ready to get started?

In the navigation bar you can find the section [audit](#). Choose a dataset to explore or upload your own dataset and then follow the instructions to derive a fairness metric that fits the context of the dataset. You can then evaluate the model based on this metric.

Introducing the FairnessLab

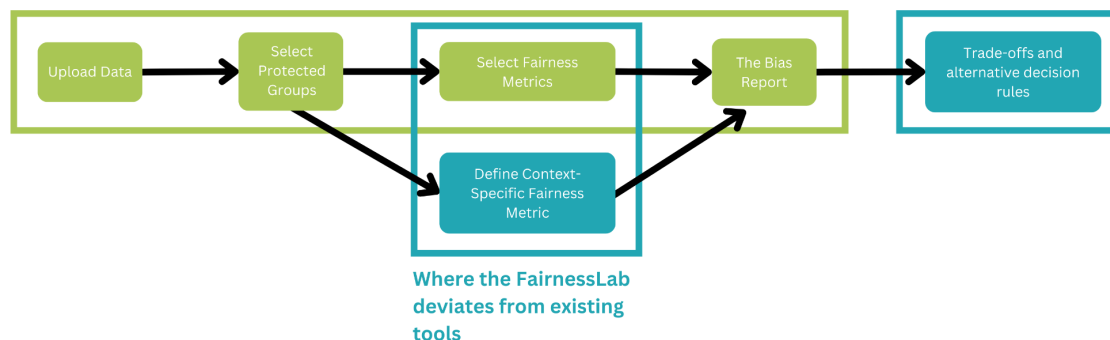
The FairnessLab is a tool developed for bias audits of binary decision-making systems. The current version particularly focuses on prediction-based decision-making systems, i.e., systems that first make a probabilistic prediction, on which they then base a decision. Compared to existing fairness audit tools (such as [Fairlearn](#), [AIF365](#), [Aequitas](#)), the overall structure of our tool is quite similar: It takes in a dataset, which represents previously taken decisions of the audited system. The tool evaluates the audited system's fairness with respect to metrics which have been chosen by the user. The FairnessLab deviates from existing tools in the fairness metrics it offers to its users: We believe that fairness is highly contextual and so evaluating fairness is a context-specific task and there is no one size fits all solution. We developed a theoretical approach, which allows for the definition of context-specific fairness metrics [[Hertweck et al. \(2022\)](#), [Baumann et al. \(2022a\)](#)]. The

FairnessLab is an implementation of this approach and thus offers three main conceptual novelties compared to existing tools:

- **Allows for evaluation of consequences instead of just decisions:**
 - Built on the idea that fairness should be about the consequences that decisions have on people's lives, the FairnessLab compares the consequences of the decisions across groups instead of just the decisions.
- **Offers different notions of distributive justice:**
 - Distributive justice is a branch of philosophy that is concerned with how goods or services should be distributed. Acknowledging that there is no universally accepted notion of distributive justice, the FairnessLab offers different philosophical concepts of what constitutes a fair distribution of the previously mentioned consequences of the audited system's decisions (what we will refer to as "pattern of justice").
- **Visualization of tradeoffs and opportunities:**
 - The logical first step of a fairness audit tool is to assess the fairness of a specific decision system. However, the FairnessLab goes a step further in that it visualizes the existing tradeoffs and also provides possible alternatives to answer questions such as:
 - What is the tradeoff between efficiency and fairness?
 - What are Pareto-optimal solutions?
 - How to make a decision system more fair without giving up efficiency?
 - How could maximum fairness be achieved? And what would it "cost"?

Compared to existing bias audit tools, the set-up of the FairnessLab is very similar, as it also analyzes a given dataset for bias w.r.t. specified groups. However, the way fairness can be defined using the FairnessLab is conceptually different and, in addition to this, it outputs not only a bias report but also offers insights into existing tradeoffs and alternatives:

Existing bias audit tools (based on Aequitas)



Comparison of the FairnessLab to existing bias audit tools

We showcase the FairnessLab by auditing an algorithm that is used in parts of the US criminal justice system. Using the FairnessLab, we replicate existing analyses of this algorithm [Angwin et al. (2016), Hao & Stray (2019)] and provide new insights. Surprisingly, we find a way to make “better” decisions from the predictions given by the audited tool both with respect to *bias* and *efficiency*. This shows that previous audits may have been flawed – meaning that they have not gone far enough in their call for favoring minority groups.¹

¹ We just look at one particular novelty the FairnessLab introduces: differences in consequences. At the end, we list further limitations of the existing approaches and how our tool solves them.

Theoretical foundations of the FairnessLab

Following the idea that fairness is about the consequences of decisions on people's lives, the FairnessLab compares two different perspectives:

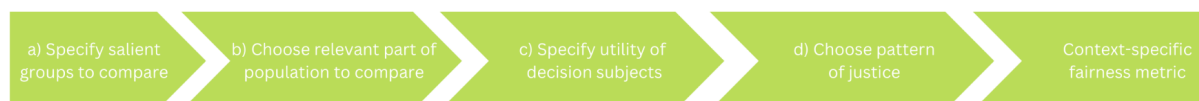
- **Decision maker:** The people or organization designing the algorithm, deciding on its design and thereby ultimately taking the decisions in question.
- **Decision subjects:** The people subjected to the decisions of the algorithm. They may or may not be aware that this algorithm is being deployed and used to make decisions about them.

Depending on the assumptions about the consequences a specific decision has, the fairness property is different, and, as a result, the optimal decision rule that achieves fairness is also different. We use the well-known concept of *utility* (as it is common in computer science and economics) to make these assumptions explicit and to quantify the consequences.

Running an audit

The fairness audit is currently prepared by following these steps:

1. Upload a dataset in the specified format with the specified column names
2. Describe the different outcomes: what is a positive/negative decision/outcome?
3. Specify the decision maker's utility (and the unit in which it is measured)
4. Define fairness for the given context (see diagram below)
 - a. First, specify the salient groups to compare. These could be groups defined by race, gender, disability status, sexual orientation, etc.
 - b. Then, choose the relevant part of the population to compare across those groups: What is a group of people where people in that group deserve the same utility while people outside of this group do not deserve this same utility?
 - c. Specify the (group-specific) utilities (and the corresponding unit) of the decision subjects (i.e., those affected by the algorithmic decision system).
 - d. Choose a so-called "pattern of justice", which specifies what constitutes a fair distribution of those utilities across groups.
5. Based on these configurations:
 - a. The decisions specified in the input dataset are audited, i.e., their fairness is quantified
 - b. A menu of options is presented to evaluate and derive optimal decision rules² for a certain level of fairness.



Process of defining a context-specific fairness metric (step 4 above)

² A "decision rule" specifies how the predictions of a predictive algorithm are turned into a final decision for each individual. Apart from the prediction, this rule might also use other available information such as group membership.

Recidivism prediction with COMPAS



The publication of an analysis of the investigative journalism group ProPublica on the COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool in 2016 kicked off a broader public and scientific debate on the fairness of algorithmic decision making systems [[Angwin et al. \(2016\)](#)]. The COMPAS algorithm is used by U.S. courts to assess the likelihood of an incarcerated person being rearrested within the next two years based on more than 100 factors. ProPublica's analysis focused on pretrial detention, so when the tool is used in the decision whether a person should be detained or released before their trial. They concluded that the tool is unfair. In particular, it reports that *"blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes."*

Note that broader criticism of tools like COMPAS questions whether such tools should be used at all. The main criticisms raised are that the US criminal justice system is deeply flawed and is in need of deep reforms while tools like COMPAS could even reinforce the current system and its injustices, which has led to the US having the highest per capita prison population worldwide by which Black people are particularly affected [[Alexander \(2011\)](#), [Green \(2018\)](#), [Green \(2022\)](#), [Bao et al. \(2022\)](#)]. Other scholars have pointed out the difficulty of predicting human behavior and pointed out that this is demonstrated by COMPAS's low success in correctly predicting recidivism [[Arvind Narayanan \(2019\)](#)].

COMPAS has been developed by Northpointe, who claim that it is fair [[Dieterich et al. \(2016\)](#)]. The reason why ProPublica and Northpointe came to these opposite conclusions is that they defined fairness in very different ways:

- Northpointe defined fairness as **equal positive predictive values (PPV) and equal false omission rates (FOR) across groups**: In a group of Black and white individuals with the same decision (released or detained before trial), Black and white individuals should be arrested at the same rate.
- ProPublica, on the other hand, defined fairness as **equal false negative rates (FNR) and equal false positive rates (FPR) across groups**: In a group of Black and white individuals with the same outcome (reoffended or not), the decision (released or detained before trial) should be false at the same rate for Black and white individuals.

As it turns out, these definitions of fairness are mathematically incompatible, so we are left with the choice of selecting one over the other [[Chouldechova \(2017\)](#), [Kleinberg et al. \(2016\)](#)]. This choice is not a technical one but a moral one: How do we define fairness in this context?

Dataset

We use the COMPAS dataset provided by ProPublica (<https://github.com/propublica/compas-analysis>). Central to this dataset is the column that is 1 if the individual was rearrested within 2 years and 0 if they were not. This is what COMPAS is trying to predict.³ For this, the COMPAS algorithm outputs risk scores ranging from 1 to 10 for each defendant⁴, with 10 being the highest risk. As the FairnessLab can only audit *binary* decision-making systems and COMPAS only returns likelihood scores, we have to decide how to turn these scores into binary decisions that we can then audit. We call a function that transforms a score into a binary decision a “decision rule.” For this audit, we use a simple threshold rule: Scores above a certain threshold lead to a positive decision and scores below or equal to said threshold lead to a negative decision. This is also what ProPublica did in their audit, so we use the same thresholds as them: People with a score of 5 or higher are seen as high risk.

As we want to audit the tool for fairness, we also have to specify which groups we’re concerned about. Among other attributes, the dataset contains the age, gender, and race of the individuals. COMPAS found unfairness between racial groups, so that is what we will focus on.

Next, we audit COMPAS, focusing on non-recidivists for the fairness assessment. Thus, we investigate whether the benefit or harm experienced by non-recidivists is distributed fairly across races.

The old and the new way of auditing COMPAS

First, we replicate ProPublica's report (“the old way”). When we say “the old way,” we mean that we will use the statistical fairness criterion that has been used in previous audits. We will thus not use the FairnessLab to create a context-specific fairness criterion, but rather recreate the previously used statistical fairness criterion through the FairnessLab. This means that we have to view this statistical fairness metric through the lens of utility distribution. Through this, we use our proposed framework to bring implicit moral assumptions to light. Second, we reanalyze the dataset as we disagree with at least one of the assumptions of the statistical fairness metric used in previous audits (“the new way”). What we change is that we will no longer assume that the consequences of decisions are the same for Black and white people. Previous work has explicitly pointed out just this issue with ProPublica’s audit: “Black defendants as well as black communities may be systematically harmed more by false positives; false positives may serve to entrench and worsen patterns of racial inequality; and considerations of prioritarianism, desert, or redress might all favor differential thresholds” [Long (2021), p.64] and “the effects of pretrial detention in an unjust basic structure are disproportionately more harmful to Black communities’ access to basic capabilities” [Zhang (2022), p. 504]. This is the reason why Long has argued against false positive rate equality. We pick up at this point and propose a different way to audit COMPAS based on these changed assumptions.

³ COMPAS is actually trying to predict whether or not a person commits a crime within the next two years and is using arrests as a proxy for this. It should, however, be noted that arrests are a biased sample of crimes as not all crimes are equally likely to lead to arrests.

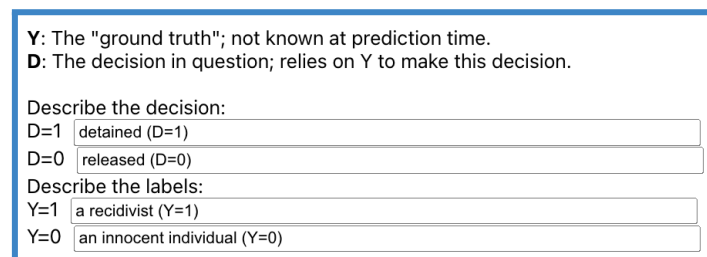
⁴ We preprocessed the dataset in the same way as ProPublica [Larson et al. (2016)].

Auditing COMPAS – the old way

[Follow our analysis by opening the FairnessLab prefilled with the configuration we'll describe in the following. For this, go to "COMPAS Case Study" and click on the button 'AUDIT COMPAS "THE OLD WAY."'](#)

Before diving into the conceptual novelties offered by the FairnessLab, let us use the tool to replicate existing results from [Angwin et al. \(2016\)](#) and [Hao & Stray \(2019\)](#).

Following the report of ProPublica, we compare false positive rates (FPR) across the groups of Black and white people. The FairnessLab first asks us to label the two possible decisions and the two possible ground truth labels (representing the outcome for a specific individual). This will then be used in the utility specification to make it easier to follow the questions. Here's how we labeled the decisions and ground truth labels for COMPAS:⁵



The screenshot shows a form with the following content:

Y: The "ground truth"; not known at prediction time.
D: The decision in question; relies on Y to make this decision.

Describe the decision:
D=1
D=0

Describe the labels:
Y=1
Y=0

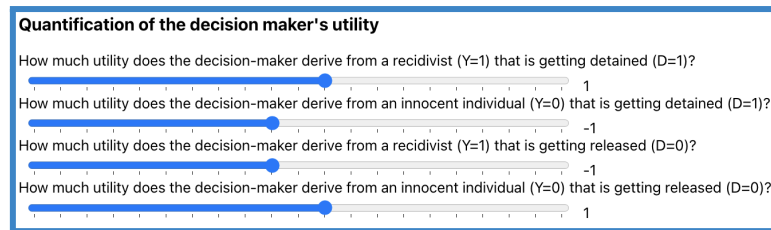
The FairnessLab asks for labels for the decision D and the outcome Y

Next, we audit the outcomes of the decisions taken based on the COMPAS algorithm from two different perspectives: the *decision maker* represents the efficiency of the system (here, the efficiency is measured through the goal of public safety) and that of the *decision subjects*, which represent the fairness desideratum. For this, we have to specify the utility values for all 4 possible outcomes from both of these perspectives. Note that ProPublica did not consider the decision maker's perspective, so this part will be up to us to define.

Decision maker utility

In their quest to transfer a risk score into a decision, decision makers have many decision rules to choose from. In order to compare them, decision makers can assign utilities to detaining and releasing individuals who would be rearrested and individuals who would not be rearrested. In the case of COMPAS, we assume the judge to be this decision maker. Judges have to weigh the severity of detaining someone who would not reoffend and releasing someone who would reoffend. This weighting is also needed to construct a utility function. For simplicity, we here assume that both are equally bad (-1) and that both correct decisions, releasing someone who is not rearrested and detaining someone who would have been rearrested, are equally good (+1). Again, these are our own assumptions for the decision-maker's utility as ProPublica did not consider the decision maker's utility.

⁵ Notice that if we inverted the decisions D and the labels Y in the preprocessing, we would need to look at true positive rates (TPR) instead, which is equivalent to the fairness notion **equality of opportunity** [\[Hardt et al. \(2016\)\]](#). To stay in line with ProPublica's audit, we chose not to invert the labels even though a "positive" decision is not actually positive in this context as it implies getting detained.



Specification of the decision maker's utility

Fairness score part I: Whose utility should be compared?

Then, the FairnessLab asks us to specify whether all Black and white defendants deserve the same utility or whether there are differences: For example, it might be morally right that people who do not commit a crime have a higher utility than people who do commit a crime. In that case, we would only compare Black defendants who do not commit another crime with white defendants who also do not commit another crime to see if these equally deserving groups can in fact expect the same utility from COMPAS. Likewise, we only compare reoffending Black individuals with reoffending white individuals. ProPublica did just this, implying that they assume that not everyone has the same claim to utility. To replicate this, we set the claims differentiator to $Y=0$. This means that we only look at the utilities of non-recidivists.⁶

Claims differentiator

Do the socio-demographic groups have the same moral claims to utility or is it only a subgroup of them? For example, one could argue that the subgroup of people with $Y=1$ deserves a higher (or lower) utility than people with $Y=0$.

Define the subgroup in which people are deserving of the same amount of utility:

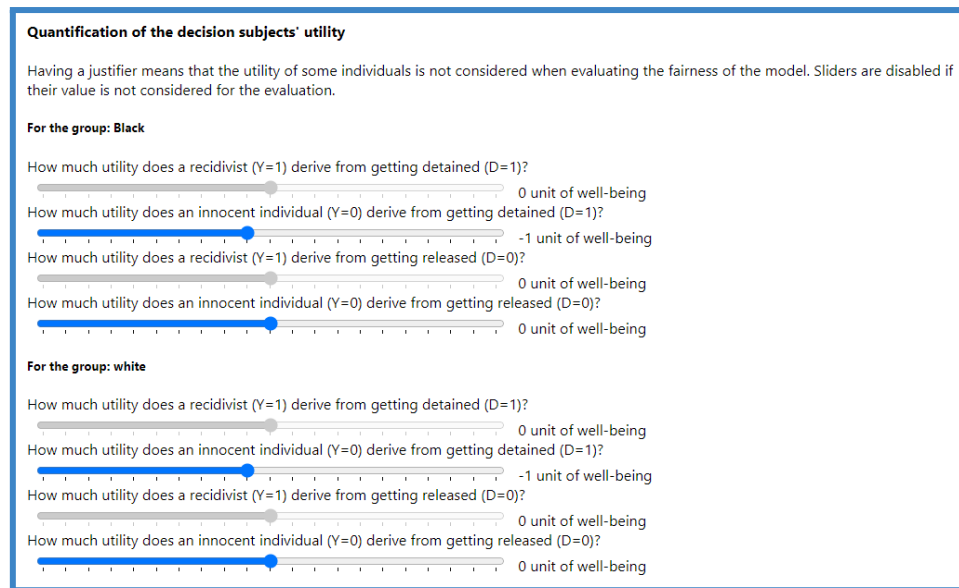
- ☐ Everyone deserves the same utility
- ☒ People with $Y=0$ deserve the same utility
- ☐ People with $Y=1$ deserve the same utility
- ☐ People with $D=0$ deserve the same utility
- ☐ People with $D=1$ deserve the same utility

Specification of the claims differentiator for the fairness score

Fairness score part II: What is the utility of the decisions for the decision subjects?

Next, we need to assign a utility value to each possible outcome. Since we are only looking at non-recidivists, there are only two possible outcomes. As is common in fairness audits, we here **assume that the utility for each outcome is constant across groups**. From the perspective of an affected individual, being detained is worse than being released, so the utility for this outcome is lower. Here, we specify the utility of a detained individual as -1 and the one of a released individual as 0. Only the utilities of non-recidivating individuals are adjustable as $Y=0$ was chosen as the claims differentiator, so only these individuals are considered for the fairness metric.

⁶ Non-recidivists who are released ($D=0$) represent a true negative (TN), and those who are detained represent a false positive (FP). The share of non-recidivists that are detained constitutes the false positive rate (FPR).



Specification of the decision subjects' utility for the fairness score

Fairness score part 3: How should the utility be distributed between Black and white defendants?

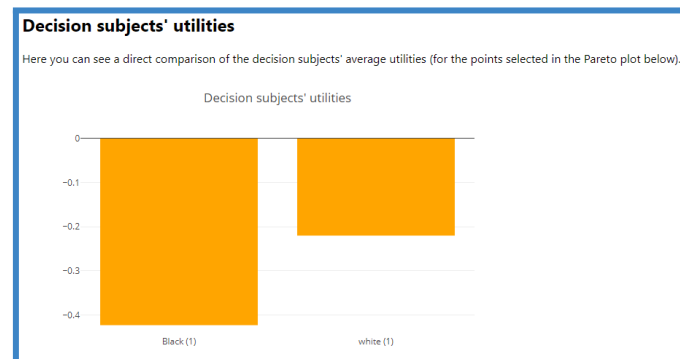
Now that the utilities are specified for each possible outcome, we must decide what we consider to be a fair distribution of this utility across Black and white people. More specifically, we must choose a pattern of justice. Here, we choose an egalitarian distribution, which is what most existing group fairness metrics follow. Egalitarianism requires that the expected utilities are equal across groups, which is what ProPublica implicitly expected when they stated: “*blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend*”.

With the above specification of the decision subject utility values, requiring equal expected utilities across groups is equivalent to requiring the existing fairness metric FPR parity.

Result

Now that we have defined how the decision-maker sees utility and what fairness means in this application context, it is time for the bias report. In this audit, we check the fairness of the decisions in our dataset, i.e., how well do these decisions align with the fairness metric we defined. When scrolling down to the audit, we can see that the average utility of Black non-reoffending people is -0.42 and the utility of white non-reoffending people is -0.22. The average utility of a Black defendant is 91% lower than that of an equally deserving white defendant. As we picked egalitarianism as the pattern of justice, we require the utilities to be the same and thus measure their difference for the fairness metric. Their difference is 0.20 units of well-being (the FairnessLab shows the *negative* absolute difference, so -0.20, to ensure that higher values imply a better alignment with the chosen fairness metric). Given

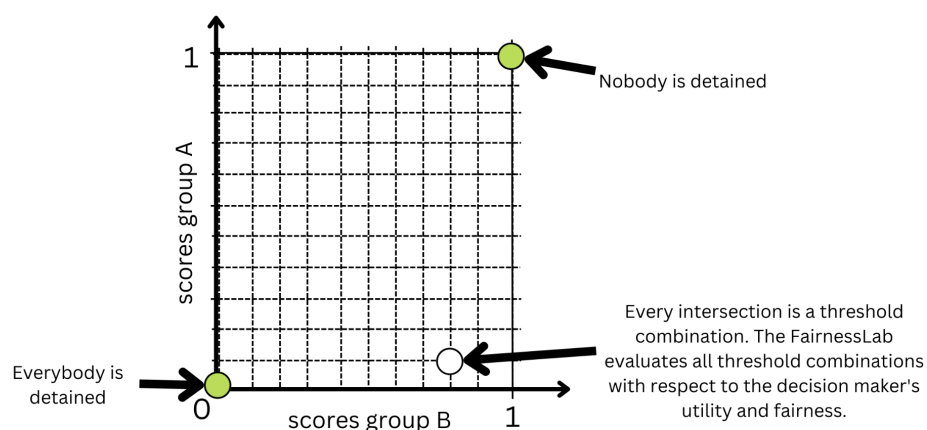
that we assume that the utility of defendants in this context varies between -1 and 0 (see assignment of decision subjects' utilities), this seems like a large gap.



Average utility for the two decision subject groups. The utility for the average non-recidivating Black defendant is almost two times lower than that of the average non-recidivating white defendant.

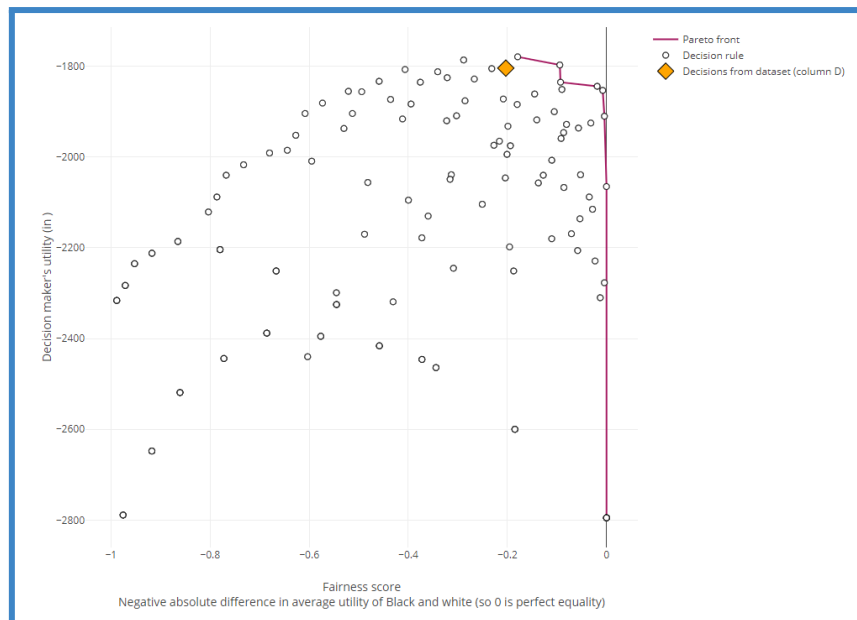
One might already assume that alternative decision rules exist that reduce this gap. The FairnessLab can help us find these decision rules. It also shows us how such decision rules affect the utility of the decision maker as the decision maker is presumably likelier to implement a fairer decision rule if it is not too costly.

It is known that decision rules that achieve the maximum possible utility for decision makers while leading to equal expected utilities across groups take the form of group-specific thresholds [see [Hardt et al. \(2016\)](#), [Corbett-Davies et al. \(2017\)](#), [Baumann et al. \(2022b\)](#)]. Therefore, the FairnessLab tests different thresholds on the scores. Specifically, it goes over all possible 11 thresholds (every step between the scores from 1 to 10) for Black defendants and over all possible 11 thresholds for white defendants. The combination of these thresholds leads to 121 threshold combinations. For each such threshold combination, the FairnessLab calculates the resulting fairness score and the utility of the decision maker. This is visualized in the following diagram:



Visualization of how the Pareto plot is created: For every group, n thresholds are tested. This yields $n \times n$ threshold combinations, for which the FairnessLab calculates the resulting decision maker's utility and the fairness score. The top right point shows the threshold combination (group A: 1, group B: 1), the point on the bottom left shows the threshold combination (group A: 0, group B: 0).

Every threshold combination is then plotted as a dot in a plot where the y-axis represents the decision-maker's utility. The x-axis represents the fairness score, where fairness scores are adjusted in a way such that a *higher* scores mean the audited system is *better* aligned with the configured fairness metric.



The Pareto plot resulting from all tested threshold combinations (y-axis: decision maker wants to maximize the utility, i.e., more utility is better / x-axis: from the perspective of the decision subjects larger values are more fair)

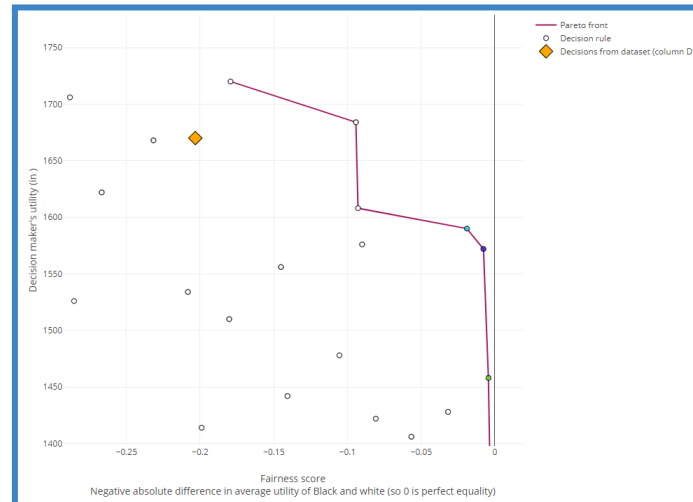
We then plot the Pareto front (red line) of all possible threshold combinations. This line represents all threshold combinations where (1) the fairness cannot be improved without worsening the utility of the decision maker or keeping it constant and where (2) the utility of the decision maker cannot be improved without worsening the utility of the decision maker or keeping it constant — among the presented decision rules.

Besides the dots (○), the figure also shows an orange diamond (◆). This diamond represents ProPublica's decision rule. It is thus directly comparable to other decision rules.

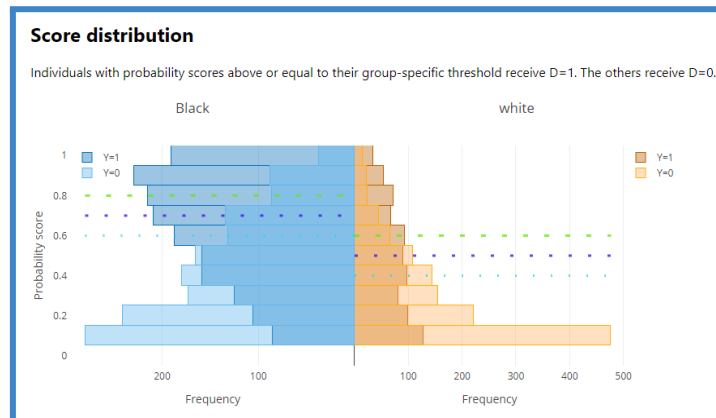
In particular, we can see that the decision rule that ProPublica audited (◆) is not Pareto-optimal. This means that more fairness and/or a higher utility for the decision-maker can be achieved without compromising on the other dimension.

One possible improvement, for example, is the point at the top left of the Pareto front. It achieves a slightly higher utility for the decision maker and a slightly better fairness score with thresholds of 6 for both Black and white defendants. More interestingly for our bias audit, however, are the threshold combinations that achieve a notably higher fairness score, so the points further on the right of the Pareto front. There are three points that are all very close to equal average utilities and do not compromise much on the decision-maker's utility. When hovering over those points, we can see the threshold combinations that they represent. When we click on them, we can compare the thresholds visually in the score distribution plot below. As we can see, the thresholds have to be “pushed apart” in order to

align with the configured fairness metric: The blue point represents the threshold combination (Black: 6, white: 4), the purple point represents (Black: 7, white: 5) and the green point represents (Black: 8, white: 6). Notice that the threshold for Black defendants is consistently higher as that of white defendants – remember that a higher threshold corresponds to fewer detained individuals. This is necessary to keep the false positive rates equal as Black defendants have a higher false positive rate than white defendants if the same threshold is applied.

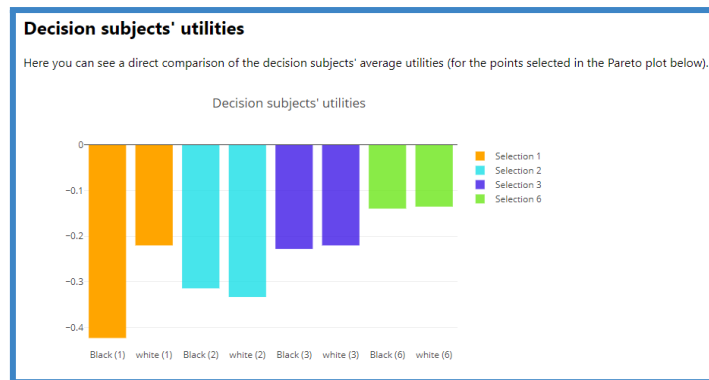


Zoomed in version of the Pareto plot that shows which points were selected. The colors of the points are used in the following plots to refer to these points.



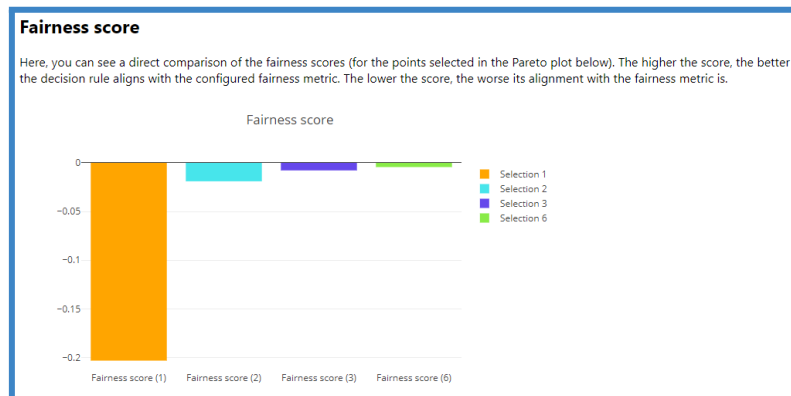
The score distribution for Black and white defendants with the thresholds of the points selected in the Pareto plot.

The threshold combinations that better fulfill the fairness metrics also lead to higher expected utilities as can be seen in the following plot:



Utilities of average non-recidivating Black defendant and average non-recidivating white defendant under different threshold combinations selected in the Pareto plot

As expected, this leads to higher (i.e., closer to 0) and thus better fairness scores. As the purple point (Black: 7, white: 5) achieves almost perfect equality, one could choose this threshold combination to improve the fairness score at the cost of some decision maker's utility – compared to the threshold combination ProPublica audited.



The fairness scores of the points selected in the Pareto plot

Auditing COMPAS – the new way 🎉

[Follow our analysis by opening the FairnessLab prefilled with the configuration we'll describe in the following. For this, go to "COMPAS Case Study" and click on the button 'AUDIT COMPAS "THE NEW WAY."](#)

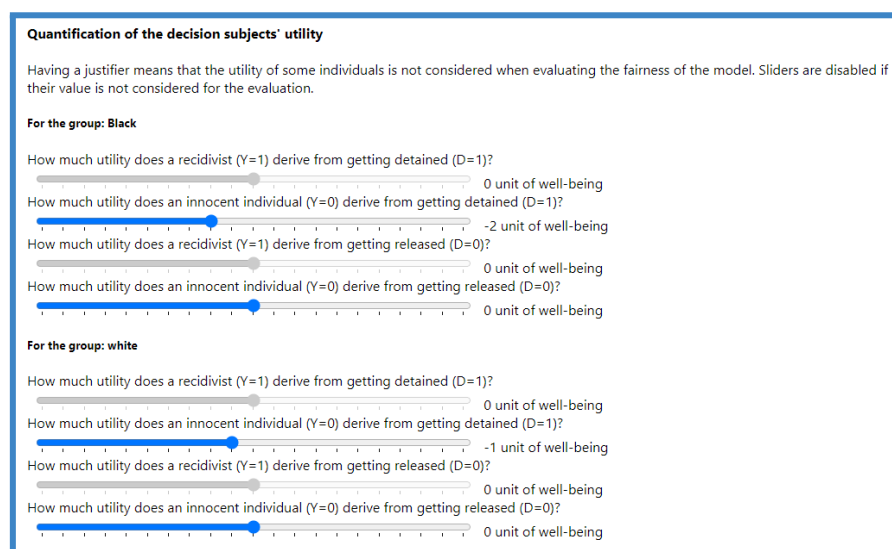
After we have replicated ProPublica's COMPAS audit, let us now loosen the inherent assumptions by not just looking at decisions but rather at the consequences these decisions have on people's lives.

Compared to the audit above, we only change the decision subjects' utility values, so all the other configurations remain the same:

- The utility values for the decision maker stay the same.
- The relevant population is still $Y=0$, i.e., non-recidivists.
- We still choose egalitarianism as the pattern of justice – requiring equal expected utilities across groups.

For this “new” audit, we no longer assume that the harm experienced by being detained before trial is constant across groups. As [Zhang (2022), p. 499] argues, equal false positive rates are therefore insufficient: “A 20% false positive rate likely leads to more injustices for Black communities than it does for white communities. Why? False positives translate to pretrial detention, and [...] detaining a Black defendant harms Black communities more than detaining a white defendant harms white communities.” Being imprisoned even for a short period can easily result in unemployment [Dobbie et al. (2018)]. This leads to financial hardship by which Black communities are more severely affected due to economic inequalities that stem from the history of slavery, segregation, housing discrimination, redlining and more [Baradaran (2019)]. Black communities are also more likely to be overpoliced [Scrivener et al. (2020), Chalfin et al. (2022)]. Black people are therefore more likely to be a target of police searches and arrests, which reinforces this inequality.

In line with claims by Long (2021) and Zhang (2022), we argue that Black non-recidivists lose more utility if detained than white non-recidivists who are detained and adjust the utilities accordingly.⁷ We will continue with a utility value of -2 for Black non-recidivists who are detained. This implies that, when being detained, Black non-recidivists lose twice as much utility as white non-recidivists.



Specification of the decision subjects' utility for the fairness score in the new audit

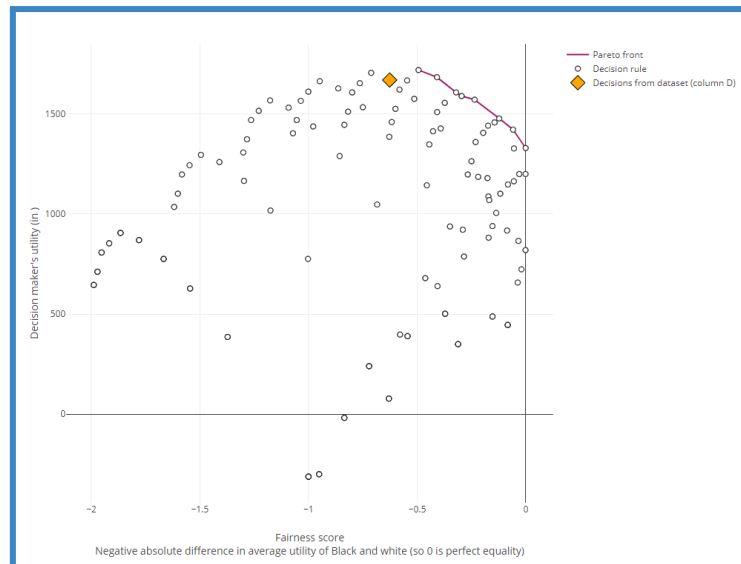
This change in the utility function for Black defendants results in a new fairness metric that is no longer equivalent to false positive rate equality. Instead, we demand equality in the expected utility of Black and white defendants, where we acknowledge that among non-recidivists, imprisonment is more harmful to Black defendants as it reinforces existing inequalities.

Looking at the results of the audit, we can see that — as expected — the utility of the average Black defendant is now even lower than in the previous audit: -0.85. The inequality

⁷ It should be noted that the decision of how much these utilities should be adjusted is somewhat subjective. This is a well-known issue in philosophy [Sen (1985)] that we cannot settle here. What utility values are chosen will influence what threshold combinations are best according to the dimensions of interest (i.e., the decision maker utility and the specified notion of fairness). It is therefore advisable to test different reasonable utility value settings that can then be compared.

in the utility of Black and white defendants has thereby also increased, which is reflected in the fairness score: -0.63.

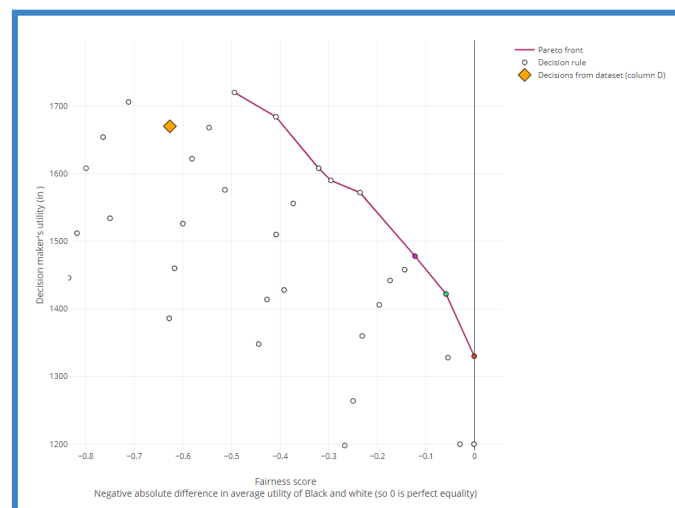
Again, we can plot all possible threshold combinations and draw the Pareto front:



The Pareto plot resulting of all tested threshold combinations in the new audit

Once again, the decision rule audited by ProPublica is not on the Pareto front. More importantly, we can see that the “fairer” threshold combinations (the points towards the bottom right of the Pareto front) have changed.

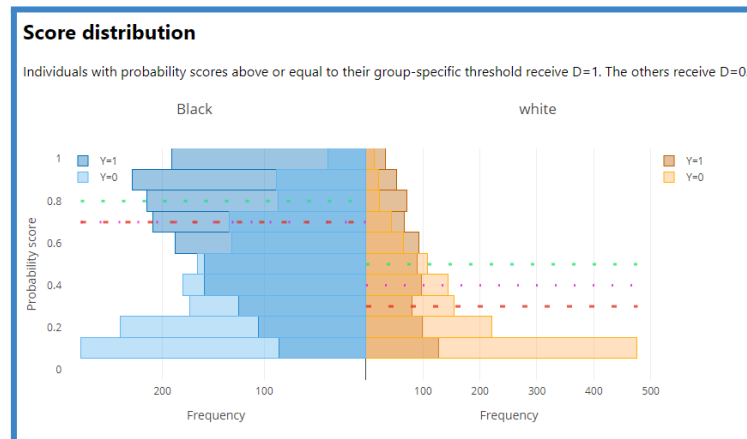
Again, we select three points that are fairer than the orange diamond (i.e., the decision rule audited by ProPublica).



Zoomed in Pareto front that shows the three selected threshold combinations besides the orange diamond, which represents ProPublica's decision rule.

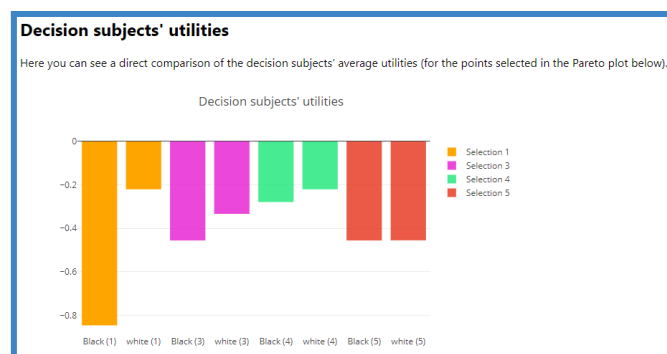
We again visualize thresholds: the “least fair” threshold combination is the pink one with (Black: 7, white: 4). The next one is the green point with (Black: 8, white: 5). The threshold combination that equalizes the expected utilities of on-recidivating Black and white

defendants is the dotted red one with (Black: 7, white: 3), which maximizes fairness with a fairness score of 0.



The score distribution for Black and white defendants with the thresholds of the points selected in the Pareto plot of the new audit.

The utilities resulting from these threshold combinations are shown in the next plot. Although the red threshold combination (Black: 7, white: 3) leads to more equal expected utilities, the expected utilities of the green threshold combination (Black: 8, white: 5) are actually higher for both groups while not causing much inequality between the groups. Therefore, one could argue that the green threshold combination should be preferred.



The fairness scores of the points selected in the Pareto plot of the new audit.

Either way, we see that the change in the utility value for non-recidivating Black defendants had a notable effect on what we consider to be a fairer threshold combination:

The thresholds for white and Black defendants have to be pushed further apart compared to the “old” audit in order to achieve a better level of equality in utilities. This means that the thresholds derived from the old audit is too conservative as it disregards the disparities in society — in particular, the fact that falsely being detained is much more harmful to Black communities than it is to white communities.

Conclusions 🍾🍾🍾

When ProPublica audited the COMPAS scoring model, they binarized the scores to “high” and “low” scores. Defendants with a score equal to or above 5 were assigned to the “high risk” category while defendants with scores below 5 were assigned to the “low risk” category. ProPublica’s audit then demonstrated unequal false positive rates for which COMPAS was heavily criticized. As our recreation of their audit showed, achieving false positive rates with COMPAS’s scores would require a threshold of 5 for white defendants and of 7 for Black defendants. Black defendants would thus have to have a higher score in order to be considered high risk. However, being a false positive, so being detained when one would not go on to be rearrested, has arguably different harms for Black and white defendants. Marginalized communities suffer much more from interactions with the criminal justice system and pretrial detentions than the majority population [[Long \(2021\)](#), [Zhang \(2022\)](#), [Baradaran \(2019\)](#), [Scrivener et al. \(2020\)](#), [Chalfin et al. \(2022\)](#)]. Considering these adverse effects, we performed a new audit. In this audit, we were able to show that when we do not just consider the *distribution* of pretrial detention decisions among non-reoffenders but rather consider the *harm* of pretrial detention for non-reoffenders, then ProPublica’s criticism does not go far enough: The inequality between Black and white defendants is even more pronounced when we consider the harms of pretrial detention. We showed that another decision rule is needed to equalize these harms: a threshold of 5 for white defendants and of 8 for Black defendants. Note that the exact thresholds will vary based on how harmful pretrial detention is defined in the FairnessLab’s utility function and based on what level of approximate equality one wants to achieve — we opted for threshold combinations that reach almost perfect equality while costing little in terms of the decision-maker’s utility. Independent of the exact choice of thresholds, we can see three important points:

1. Equal thresholds for Black and white defendants lead to inequality in utilities both in the “old” and “new” audit.
2. In order to achieve equality in the utilities of the “old” audit, the thresholds for Black and white defendants have to be pushed apart, with the threshold for Black defendants being *higher* than that for white defendants.
3. In order to achieve equality in the utilities of the “new” audit, the thresholds for Black and white defendants have to be pushed further apart than in (2), with the threshold for Black defendants being *much higher* than that for white defendants.

This difference in thresholds is necessary if we want to avoid the reinforcement of existing inequalities.

HOORAY!



YOU MADE IT TO THE END!

Now all that is left is to discuss some open questions and the ethics of risk assessment in general.

Open questions

In our “new” audit, we only changed one of the assumptions compared to the “old” audit. However, we could also question some other often implicit assumptions of existing audits, for example:

- For this report, we kept “egalitarianism” as the pattern of justice, but one could also argue that, for example, the utility of the worst-off group should be maximized or at least prioritized, which would imply “maximin” or “prioritarianism,” respectively.
- Further, one could argue that Black and white communities have different utilities to start with, meaning that even before a pretrial decision is made, the utility of minority groups is lower – this is slightly different compared to the “new” audit presented in this report, where we claimed that non-recidivists who are not detained have equal utilities across groups but Black non-recidivists lose more utility when being detained. Accounting for different starting utilities across groups would most likely result in different decision rules.
- Additionally, one could question whether the decision subjects’ utility really only depends on Y , D and the sensitive attribute. The utility of being released if one does not recidivate could, for example, also depend on one’s financial situation or support system. One could therefore try to quantify this through a more complicated utility function that takes other attributes into account.
- It would also be interesting to compare the COMPAS risk scores against other scoring algorithms. One could, for example, compare it against a simple rule-based decision algorithm based on only a handful of variables and see how this fares in terms of fairness as previous work has shown that it achieves a similar accuracy [[Rudin \(2019\)](#)] and might thus also achieve a similar utility for the decision maker.

Ethics statement

Note that group-specific thresholds cannot be said to make a tool like COMPAS “fair”: The systemic racism embedded in the US criminal justice system cannot be “fixed” by a risk assessment tool that has been audited for bias — deeper reforms are necessary [[Green \(2018\)](#), [Green \(2022\)](#)]. A tool used to decide who to detain may actually reinforce existing structures and get in the way of such deeper reforms. A better use of a predictive tool could be in rehabilitation efforts as highlighted by [[Bao et al. \(2022\)](#)]. Note that a change of how the tool is used would also change the audit as the decision to allow someone to participate in a rehabilitation program would result in different utilities for decision subjects than the decision to imprison them. More generally, tools like COMPAS do not only have a fairness issue — their low accuracy also raises questions about their deployment [[Dressel and Farid \(2018\)](#)]. As Arvind Narayanan points out, predicting social outcomes is extremely difficult or even impossible, so tools trying to do that are “fundamentally dubious” [[Arvind Narayanan \(2019\)](#), p. 9].

Our audit is thus in no way meant to legitimize the usage of risk assessment systems in the criminal justice system. Rather, it is meant to highlight one of the shortcomings of previous audits: The FairnessLab allows for a reevaluation of the assumptions hidden in existing audits and for new audits that make use of context-specific fairness criteria.

References

- [[Aequitas](#)] Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.
- [[AIF365](#)] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5), 4-1.
- [[Alexander \(2011\)](#)] Alexander, M. (2011). The new jim crow. *Ohio St. J. Crim. L.*, 9, 7.
- [[Angwin et al. \(2016\)](#)] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of Data and Analytics* (pp. 254-264). Auerbach Publications.
- [[Arvind Narayanan \(2019\)](#)] Narayanan, A. (2019). How to recognize AI snake oil. *Arthur Miller Lecture on Science and Ethics*.
- [[Bao et al. \(2022\)](#)] Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., ... & Venkatasubramanian, S. (2022). It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *arXiv preprint arXiv:2106.05498*.
- [[Baradaran \(2019\)](#)] Baradaran, M. (2019). *The color of money: Black banks and the racial wealth gap*. Harvard University Press.
- [[Baumann et al. \(2022a\)](#)] Baumann, J., Hertweck, C., Loi, M., & Heitz, C. (2022). Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. *arXiv preprint arXiv:2206.02897*.
- [[Baumann et al. \(2022b\)](#)] Baumann, J., Hannák, A., & Heitz, C. (2022). Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. *arXiv preprint arXiv:2206.02237*.
- [[Chalfin et al. \(2022\)](#)] Chalfin, A., Hansen, B., Weisburst, E. K., & Williams Jr, M. C. (2022). Police force size and civilian race. *American Economic Review: Insights*, 4(2), 139-58.
- [[Chouldechova \(2017\)](#)] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), 153-163.
- [[Corbett-Davies et al. \(2017\)](#)] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797-806).
- [[Dieterich et al. \(2016\)](#)] Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4).
- [[Dobbie et al. \(2018\)](#)] Dobbie, W., Goldin, J., & Yang, C. S. (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2), 201-40.
- [[Dressel and Farid \(2018\)](#)] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1), eaao5580.

- [[Fairlearn](#)] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., ... & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.
- [[Green \(2018\)](#)] Green, B. (2018). Fair” risk assessments: A precarious approach for criminal justice reform. In *5th Workshop on fairness, accountability, and transparency in machine learning* (pp. 1-5).
- [[Green \(2022\)](#)] Green, B. (2022). Escaping the "Impossibility of Fairness": From Formal to Substantive Algorithmic Fairness. *arXiv preprint arXiv:2107.04642*.
- [[Hertweck et al. \(2022\)](#)] Hertweck, C., Baumann, J., Loi, M., Viganò, E., & Heitz, C. (2022). A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs. *arXiv preprint arXiv:2206.02891*.
- [[Hao & Stray \(2019\)](#)] Hao, K., & Stray, J. (2019). Can you make AI fairer than a judge? Play our courtroom algorithm game. *MIT Technology Review*.
- [[Hardt et al. \(2016\)](#)] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- [[Kleinberg et al. \(2016\)](#)] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
- [[Larson et al. \(2016\)](#)] Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm.
- [[Long \(2021\)](#)] Long, R. (2021). Fairness in machine learning: against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy*, 19(1), 49-78.
- [[Rudin \(2019\)](#)] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.
- [[Scrivener et al. \(2020\)](#)] Scrivener, L., Meizlish, A., Bond, E., & Chauhan, P. (2020). Tracking enforcement trends in New York City: 2003-2018. *Data Collaborative for Justice*.
- [[Sen \(1985\)](#)] Sen, A. (1985). *The standard of living*. Cambridge University Press.
- [[Zhang \(2022\)](#)] Zhang, M. (2022). Affirmative Algorithms: Relational Equality as Algorithmic Fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 495-507).

Resources

- COMPAS dataset: <https://github.com/propublica/compas-analysis>
 - Our preprocessing Jupyter notebook: [COMPAS.ipynb](#)
- FairnessLab:
 - WebApp: <https://joebaumann.github.io/FairnessLab>
 - Code (MIT license): <https://github.com/joebaumann/FairnessLab>