10

Fairness

Sospital is a leading (fictional) health insurance company in the United States. Imagine that you are the lead data scientist collaborating with a problem owner in charge of transforming the company's *care management* programs. Care management is the set of services that help patients with chronic or complex conditions manage their health and have better clinical outcomes. Extra care management is administered by a dedicated team composed of physicians, other clinicians, and caregivers who come up with and execute a coordinated plan that emphasizes preventative health actions. The problem owner at Sospital has made a lot of progress in implementing software-based solutions for the care coordination piece and has changed the culture to support them, but is still struggling with the patient intake process. The main struggle is in identifying the members of health plans that need extra care management. This is a mostly manual process right now that the problem owner would like to automate.

You begin the machine learning lifecycle through an initial set of conversations with the problem owner and determine that it is not an exploitative use case that could immediately be an instrument of oppression. It is also a problem in which machine learning may be helpful. You next consult a paid panel of diverse voices that includes actual patients. You learn from them that black Americans have not been served well by the health care system historically and have a deep-seated mistrust of it. Therefore, you should ensure that the machine learning model does not propagate systematic disadvantage to the black community. The system should be *fair* and not contain unwanted biases.

Your task now is to develop a detailed problem specification for a fair machine learning system for allocating care management programs to Sospital members and proceed along the different phases of the machine learning lifecycle without taking shortcuts. In this chapter, you will:

- compare and contrast definitions of fairness in a machine learning context,
- select an appropriate notion of fairness for your task, and
- mitigate unwanted biases at various points in the modeling pipeline to achieve fairer systems.

10.1 The Different Definitions of Fairness

The topic of this chapter, algorithmic fairness, is the most contested topic in the book because it is intertwined with social justice and cannot be reduced to technical-only conceptions. Because of this broader conception of fairness, it may seem odd to you that this chapter is in a part of the book that also contains technical robustness. The reason for including it this way is due to the technical similarities with robustness which you, as a data scientist, can make use of and which are rarely recognized in other literature. This choice was not made to minimize the social importance of algorithmic fairness.

Fairness and justice are almost synonymous, and are political. There are several kinds of justice, including (1) *distributive justice*, (2) *procedural justice*, (3) *restorative justice*, and (4) *retributive justice*.

- Distributive justice is equality in what people receive—the outcomes.
- Procedural justice is sameness in the way it is decided what people receive.
- Restorative justice repairs a harm.
- Retributive justice seeks to punish wrongdoers.

All of the different forms of justice have important roles in society and sociotechnical systems. In the problem specification phase of a model that determines who receives Sospital's care management and who doesn't, you need to focus on distributive justice. This focus on distributive justice is generally true in designing machine learning systems because machine learning itself is focused on outcomes. The other kinds of justice are important in setting the context in which machine learning is and is not used. They are essential in promoting accountability and *holistically* tamping down racism, sexism, classism, ageism, ableism, and other unwanted discriminatory behaviors.

> "Don't conflate CS/AI/tech ethics and social justice issues. They're definitely related, but not interchangeable."
>
> —Brandeis Marshall, computer scientist at Spelman College

Why would different individuals and groups receive an unequal allocation of care management? Since it is a limited resource, not everyone can receive it.[1] The more chronically ill that patients are, the more likely they should be to receive care management. This sort of discrimination is generally acceptable, and is the sort of task machine learning systems are suited for. It becomes unacceptable and unfair when the allocation gives a *systematic* advantage to certain *privileged* groups and individuals and a systematic disadvantage to certain *unprivileged* groups and individuals. Privileged groups and individuals are defined to be those who have historically been more likely to receive the *favorable label* in a machine learning binary classification task. Receiving care management is a favorable label because patients are given extra services to keep them healthy. Other favorable labels include being hired, not being fired, being approved for a loan, not being arrested, and being granted bail. Privilege is a result of power imbalances, and the same groups may not be privileged in all contexts, even within the same society. In some narrow societal contexts, it may even be the elite who are without power.

[1] You can argue that this way of thinking is flawed and society should be doing whatever it takes so that care management is not a limited resource, but it is the reality today.

Privileged and unprivileged groups are delineated by *protected attributes* such as race, ethnicity, gender, religion, and age. There is no one universal set of protected attributes. They are determined from laws, regulations, or other policies governing a particular application domain in a particular jurisdiction. As a health insurer in the United States, Sospital is regulated under Section 1557 of the Patient Protection and Affordable Care Act with the specific protected attributes of race, color, national origin, sex, age, and disability. In health care in the United States, non-Hispanic whites are usually a privileged group due to multifaceted reasons of power. For ease of explanation and conciseness, the remainder of the chapter uses whites as the privileged group and blacks as the unprivileged group.

There are two main types of fairness you need to be concerned about: (1) *group fairness* and (2) *individual fairness*. Group fairness is the idea that the average classifier behavior should be the same across groups defined by protected attributes. Individual fairness is the idea that individuals similar in their features should receive similar model predictions. Individual fairness includes the special case of two individuals who are exactly the same in every respect except for the value of one protected attribute (this special case is known as *counterfactual fairness*). Given the regulations Sospital is operating under, group fairness is the more important notion to include in the care management problem specification, but you should not forget to consider individual fairness in your problem specification.

10.2 Where Does Unfairness Come From?

Unfairness in the narrow scope of allocation decisions (distributive justice) has a few different sources. The most obvious source of unfairness is unwanted bias, specifically social bias in the measurement process (going from the construct space to the observed space) and representation bias in the sampling process (going from the observed space to the raw data space) that you learned about in Chapter 4, shown in Figure 10.1. (This is a repetition of Figure 9.2 and an extension of Figure 4.3 where the concepts of construct space and observed space were first introduced.)

In the data understanding phase, you have figured out that you will use privacy-preserved historical medical claims from Sospital members along with their past selection for care management as the data source. Medical claims data is generated any time a patient sees a doctor, undergoes a procedure, or fills a pharmacy order. It is structured data that includes diagnosis codes, procedure codes, and drug codes, all standardized using the ICD-10, CPT, and NDC schemes, respectively.[2] It also includes the dollar amount billed and paid along with the date of service. It is administrative data used by the healthcare provider to get reimbursed by Sospital.

> "If humans didn't behave the way we do there would be no behavior data to correct. The training data is society."
>
> — M. C. Hammer, musician and technology consultant

[2]See https://www.cms.gov/files/document/blueprint-codes-code-systems-value-sets.pdf for details about these coding schemes.

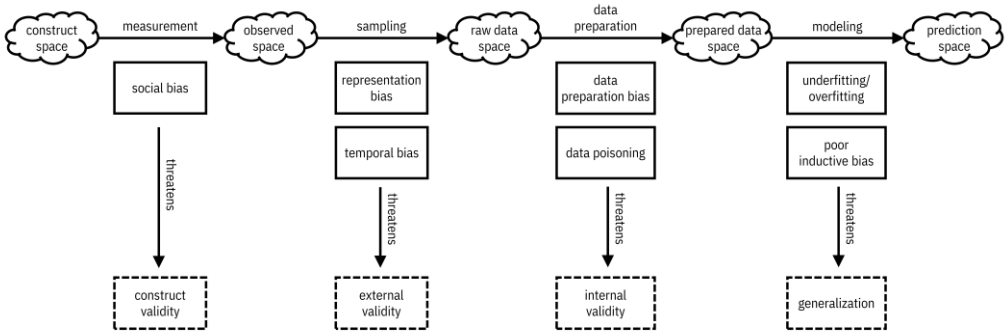

Figure 10.1. *Bias in measurement and sampling are the most obvious sources of unfairness in machine learning, but not the only ones.* Accessible caption. A sequence of five spaces, each represented as a cloud. The construct space leads to the observed space via the measurement process. The observed space leads to the raw data space via the sampling process. The raw data space leads to the prepared data space via the data preparation process. The prepared data space leads to the prediction space via the modeling process. The measurement process contains social bias, which threatens construct validity. The sampling process contains representation bias and temporal bias, which threatens external validity. The data preparation process contains data preparation bias and data poisoning, which threaten internal validity. The modeling process contains underfitting/overfitting and poor inductive bias, which threaten generalization.

Social bias enters claims data in a few ways. First, you might think that patients who visit doctors a lot and get many prescriptions filled, i.e. utilize the health care system a lot, are sicker and thus more appropriate candidates for care management. While it is directionally true that greater health care utilization implies a sicker patient, it is not true when comparing patients across populations such as whites and blacks. Blacks tend to be sicker for an equal level of utilization due to structural issues in the health care system.[3] The same is true when looking at health care cost instead of utilization. Another social bias can be in the codes. For example, black people are less-often treated for pain than white people in the United States due to false beliefs among clinicians that black people feel less pain.[4] Moreover, there can be social bias in the human-determined labels of selection for care management in the past due to implicit cognitive biases or prejudice on the part of the decision maker. Representation bias enters claims data because it is only from Sospital's own members. This population may, for example, undersample blacks if Sospital offers its commercial plans primarily in counties with larger white populations.

Besides the social and representation biases given above that are already present in raw data, you need to be careful that you don't introduce other forms of unfairness in the problem specification and data preparation phases. For example, suppose you don't have the labels from human decision makers in the past. In that case, you might decide to use a threshold on utilization or cost as a proxy outcome

[3]Moninder Singh and Karthikeyan Natesan Ramamurthy. "Understanding Racial Bias in Health Using the Medical Expenditure Panel Survey Data." In: *Proceedings of the NeurIPS Workshop on Fair ML for Health*. Vancouver, Canada, Dec. 2019.
[4]Oluwafunmilayo Akinlade. "Taking Black Pain Seriously." In: *New England Journal of Medicine* 383.e68 (Sep. 2020).

variable, but that would make blacks less likely to be selected for care management at equal levels of infirmity for the reasons described above. Also, as part of feature engineering, you might think to combine individual cost or utilization events into more comprehensive categories, but if you aren't careful you could make racial bias worse. It turns out that combining all kinds of health system utilization into a single feature yields unwanted racial bias, but keeping inpatient hospital nights and frequent emergency room utilization as separate kinds of utilization keeps the bias down in nationally-representative data.[5]

> "As AI is embedded into our day to day lives it's critical that we ensure our models don't inadvertently incorporate latent stereotypes and prejudices."
>
> —Richard Zemel, computer scientist at University of Toronto

You might be thinking that you already know how to measure and mitigate biases in measurement, sampling, and data preparation from Chapter 9, distribution shift. What's different about fairness? Although there is plenty to share between distribution shift and fairness,[6] there are two main technical differences between the two topics. First is access to the construct space. You can get data from the construct space in distribution shift scenarios. Maybe not immediately, but if you wait, collect, and label data from the deployment environment, you will have data reflecting the construct space. However, you never have access to the construct space in fairness settings. The construct space reflects a perfect egalitarian world that does not exist in real life, so you can't get data from it. (Recall that in Chapter 4, we said that *hakuna matata* reigns in the construct space (it means no worries).) Second is the specification of what is sought. In distribution shift, there is no further specification beyond just trying to match the shifted distribution. In fairness, there are precise policy-driven notions and quantitative criteria that define the desired state of data and/or models that are not dependent on the data distribution you have. You'll learn about these precise notions and how to choose among them in the next chapter.

Related to causal and anticausal learning covered in Chapter 9, the protected attribute is like the environment variable. Fairness and distributive justice are usually conceived in a causal (rather than anticausal) learning framework in which the outcome label is extrinsic: the protected attribute may cause the other features, which in turn cause the selection for care management. However, this setup is not always the case.

10.3 Defining Group Fairness

You've gone back to the problem specification phase after some amount of data understanding because you and the problem owner have realized that there is a strong possibility of unfairness if left unchecked. Given the Section 1557 regulations Sospital is working under as a health insurer, you start by looking

[5]Moninder Singh. "Algorithmic Selection of Patients for Case Management: Alternative Proxies to Healthcare Costs." In: *Proceedings of the AAAI Workshop on Trustworthy AI for Healthcare*. Feb. 2021.
[6]Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. "Exchanging Lessons Between Algorithmic Fairness and Domain Generalization." arXiv:2010.07249, 2020.

deeper into group fairness. Group fairness is about comparing members of the privileged group and members of the unprivileged group on average.

10.3.1 Statistical Parity Difference and Disparate Impact Ratio

One key concept in unwanted discrimination is *disparate impact*: privileged and unprivileged groups receiving different outcomes irrespective of the decision maker's intent and irrespective of the decision-making procedure. Statistical parity difference is a group fairness metric that you can consider in the care management problem specification that quantifies disparate impact by computing the difference in selection rates of the favorable label $P(\hat{y}(X) = \text{fav})$ (rate of receiving extra care) between the privileged ($Z = \text{priv}$; whites) and unprivileged groups ($Z = \text{unpr}$; blacks):

$$\text{statistical parity difference} = P(\hat{y}(X) = \text{fav} \mid Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Z = \text{priv}).$$

Equation 10.1

A value of 0 means that members of the unprivileged group (blacks) and the privileged group (whites) are getting selected for extra care management at equal rates, which is considered a fair situation. A negative value of statistical parity difference indicates that the unprivileged group is at a disadvantage and a positive value indicates that the privileged group is at a disadvantage. A requirement in a problem specification may be that the learned model must have a statistical parity difference close to 0. An example calculation of statistical parity difference is shown in Figure 10.2.

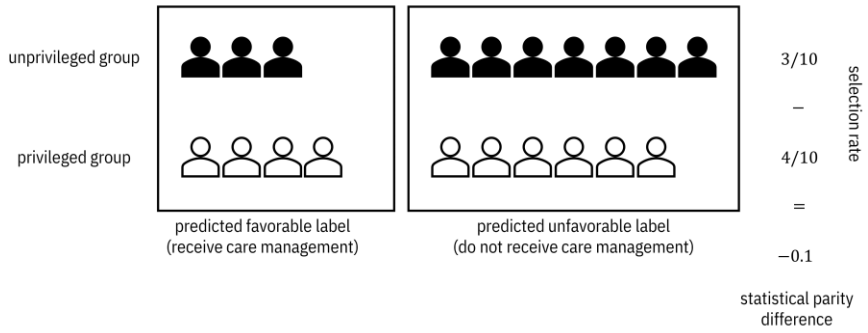

Figure 10.2. *An example calculation of statistical parity difference.* Accessible caption. 3 members of the unprivileged group are predicted with the favorable label (receive care management) and 7 are predicted with the unfavorable label (don't receive care management). 4 members of the privileged group are predicted with the favorable label and 6 are predicted with the unfavorable label. The selection rate for the unprivileged group is 3/10 and for the privileged group is 4/10. The difference, the statistical parity difference is −0.1.

Disparate impact can also be quantified as a ratio:

$$\text{disparate impact ratio} = P(\hat{y}(X) = \text{fav} \mid Z = \text{unpr})/P(\hat{y}(X) = \text{fav} \mid Z = \text{priv}).$$

Equation 10.2

Here, a value of 1 indicates fairness, values less than 1 indicate disadvantage faced by the unprivileged group, and values greater than 1 indicate disadvantage faced by the privileged group. The *disparate impact ratio* is also sometimes known as the *relative risk ratio* or the *adverse impact ratio*. In some application domains such as employment, a value of the disparate impact ratio less than 0.8 is considered unfair and values greater than 0.8 are considered fair. This so-called *four-fifths rule* problem specification is asymmetric because it does not speak to disadvantage experienced by the privileged group. It can be symmetrized by considering disparate impact ratios between 0.8 and 1.25 to be fair. Statistical parity difference and disparate impact ratio can be understood as measuring a form of *independence* between the prediction $\hat{y}(X)$ and the protected attribute Z.[7] Besides statistical parity difference and disparate impact ratio, another way to quantify the independence between $\hat{y}(X)$ and Z is their mutual information.

Both statistical parity difference and disparate impact ratio can also be defined on the training data instead of the model predictions by replacing $\hat{y}(X)$ with Y. Thus, they can be measured and tested (1) on the dataset before model training, as a *dataset fairness metric*, as well as (2) on the learned classifier after model training as a *classifier fairness metric*, shown in Figure 10.3.

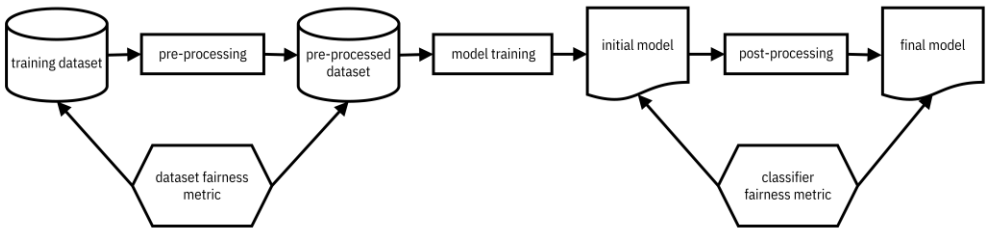

Figure 10.3. *Two types of fairness metrics in different parts of the machine learning pipeline.* Accessible caption. A block diagram with a training dataset as input to a pre-processing block with a pre-processed dataset as output. The pre-processed dataset is input to a model training block with an initial model as output. The initial model is input to a post-processing block with a final model as output. A dataset fairness metric block is applied to the training dataset and pre-processed dataset. A classifier fairness metric block is applied to the initial model and final model.

10.3.2 Average Odds Difference

You've examined disparate impact-based group fairness metrics so far, but want to learn another one before you start comparing and contrasting them as you figure out the problem specification for the care management model. A different group fairness metric is *average odds difference*, which is based on model performance metrics rather than simply the selection rate. (It can thus only be used as a classifier fairness metric, not a dataset fairness metric as shown in Figure 10.3.) The average odds difference involves the two metrics in the ROC: the true favorable label rate (true positive rate) and the false favorable label rate (false positive rate). You take the difference of true favorable rates between the

[7]Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. URL: https://fairmlbook.org, 2020.

unprivileged and privileged groups and the difference of the false favorable rates between the unprivileged and privileged groups, and average them:

$$\begin{aligned}&\text{average odds difference}\\&= \tfrac{1}{2}\big(P(\hat{y}(X) = \text{fav} \mid Y = \text{fav},\ Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{fav},\ Z = \text{priv})\big)\\&+ \tfrac{1}{2}\big(P(\hat{y}(X) = \text{fav} \mid Y = \text{unf},\ Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{unf},\ Z = \text{priv})\big).\end{aligned}$$

Equation 10.3

An example calculation of average odds difference is shown in Figure 10.4.

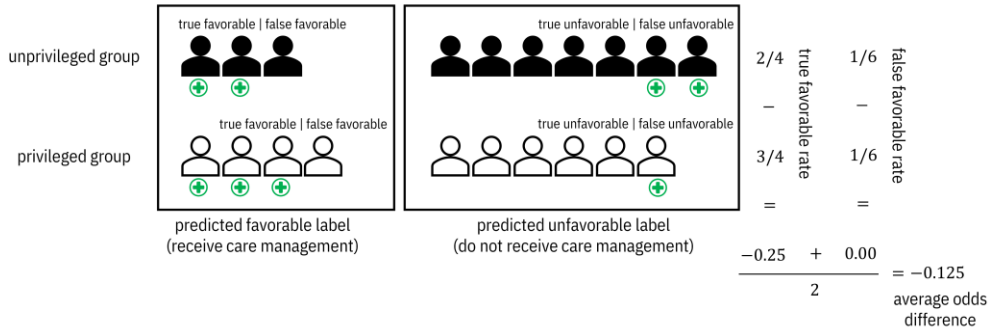

Figure 10.4. *An example calculation of average odds difference. The crosses below the members indicate a true need for care management.* Accessible caption. In the unprivileged group, 2 members receive true favorable outcomes and 2 receive false unfavorable outcomes, giving a 2/4 true favorable rate. In the privileged group, 3 members receive true favorable outcomes and 1 receives a false unfavorable outcome, giving a 3/4 true favorable rate. The true favorable rate difference is −0.25. In the unprivileged group, 1 member receives a false favorable outcome and 5 receive a true unfavorable outcome, giving a 1/6 false favorable rate. In the privileged group, 1 member receives a false favorable outcome and 5 receive a true unfavorable outcome, giving a 1/6 false favorable rate. The false favorable rate difference is 0. Averaging the two differences gives a −0.125 average odds difference.

In the average odds difference, the true favorable rate difference and the false favorable rate difference can cancel out and hide unfairness, so it is better to take the absolute value before averaging:

$$\begin{aligned}&\text{average absolute odds difference}\\&= \tfrac{1}{2}\big|P(\hat{y}(X) = \text{fav} \mid Y = \text{fav},\ Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{fav},\ Z = \text{priv})\big|\\&+ \tfrac{1}{2}\big|P(\hat{y}(X) = \text{fav} \mid Y = \text{unf},\ Z = \text{unpr}) - P(\hat{y}(X) = \text{fav} \mid Y = \text{unf},\ Z = \text{priv})\big|.\end{aligned}$$

Equation 10.3

The average odds difference is a way to measure the *separation* of the prediction $\hat{y}(X)$ and the protected attribute Z by the true label Y in any of the three Bayesian networks shown in Figure 10.5. A value of 0 average absolute odds difference indicates independence of $\hat{y}(X)$ and Z conditioned on Y. This is deemed a fair situation and termed *equality of odds*.

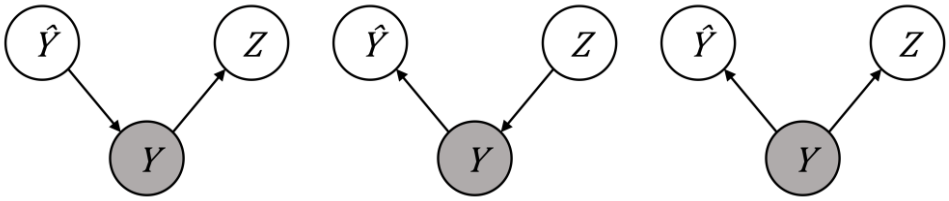

Figure 10.5. *Illustration of the true label Y separating the prediction and the protected attribute in various Bayesian networks.* Accessible caption. Three networks that show separation: $\hat{Y} \rightarrow \underline{Y} \rightarrow Z$, $\hat{Y} \leftarrow \underline{Y} \leftarrow Z$, and $\hat{Y} \leftarrow \underline{Y} \rightarrow Z$.

10.3.3 Choosing Between Statistical Parity and Average Odds Difference

What's the point of these two different group fairness metrics? They don't appear to be radically different. But they actually *are* radically different in an important conceptual way: either you believe there is social bias during measurement or not. These two worldviews have been named (1) "we're all equal" (the privileged group and unprivileged group have the same inherent distribution of health in the construct space, but there is bias during measurement that makes it appear this is not the case) and (2) "what you see is what you get" (there are inherent differences between the two groups in the construct space and this shows up in the observed space without a need for any bias during measurement).[8] Since under the "we're all equal" worldview, there is already structural bias in the observed space (blacks have lower health utilization and cost for the same level of health as whites), it does not really make sense to look at model accuracy rates computed in an already-biased space. Therefore, independence or disparate impact fairness definitions make sense and your problem specification should be based on them. However, if you believe that "what you see is what you get"—the observed space is a true representation of the inherent distributions of the groups and the only bias is sampling bias—then the accuracy-related equality of odds fairness metrics make sense. In this case, your problem specification should be based on equality of odds.

10.3.4 Average Predictive Value Difference

And if it wasn't complicated enough, let's throw one more group fairness definition into the mix: *calibration by group* or *sufficiency*. Recall from Chapter 6 that for continuous score outputs, the predicted score corresponds to the proportion of positive true labels in a *calibrated* classifier, or $P(Y = 1 \mid S = s) = s$. For fairness, you'd like the calibration to be true across the groups defined by protected attributes, so $P(Y = 1 \mid S = s, Z = z) = s$ for all groups z. If a classifier is calibrated by group, it is also *sufficient*, which means that Y and Z conditioned on S (or $\hat{y}(X)$) are independent. The graphical models for sufficiency are shown in Figure 10.6. To allow for better comparison to Figure 10.5 (the graphical models of separation), the predicted score is indicated by $\hat{Y}$ rather than S.

[8]Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making." In: *Communications of the ACM* 64.4 (Apr. 2021), pp. 136–143.

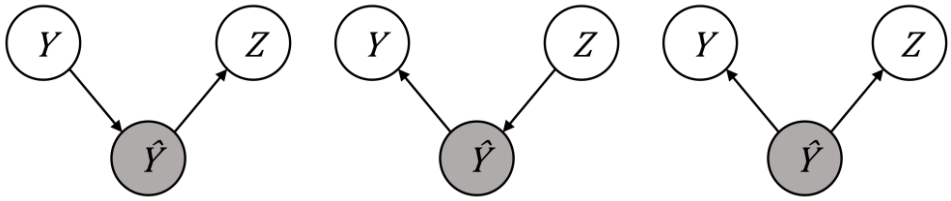

Figure 10.6. *Illustration of the predicted label $\hat{Y}$ separating the true label and the protected attribute in various Bayesian networks, which is known as sufficiency.* Accessible caption. Three networks that show sufficiency: $Y \rightarrow \underline{\hat{Y}} \rightarrow Z$, $Y \leftarrow \underline{\hat{Y}} \leftarrow Z$, and $Y \leftarrow \underline{\hat{Y}} \rightarrow Z$.

Since sufficiency and separation are somewhat opposites of each other with Y and $\hat{Y}$ reversed, their quantifications are also opposites with Y and $\hat{Y}$ reversed. Recall from Chapter 6 that the positive predictive value is the reverse of the true positive rate: $P(Y = \text{fav} \mid \hat{y}(X) = \text{fav})$ and that the false omission rate is the reverse of the false positive rate: $P(Y = \text{fav} \mid \hat{y}(X) = \text{unf})$. To quantify sufficiency unfairness, compute the average difference of the positive predictive value and false omission rate across the unprivileged (black) and privileged (white) groups:

$$\begin{aligned}&\text{average predictive value difference}\\ &= \tfrac{1}{2}\big(P(Y = \text{fav} \mid \hat{y}(X) = \text{fav},\ Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{fav},\ Z = \text{priv})\big)\\ &+ \tfrac{1}{2}\big(P(Y = \text{fav} \mid \hat{y}(X) = \text{unf},\ Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{unf},\ Z = \text{priv})\big).\end{aligned}$$

Equation 10.4

An example calculation for average predictive value difference is shown in Figure 10.7. The example illustrates a case in which the two halves of the metric cancel out because they have opposite sign, so a version with absolute values before averaging makes sense:

$$\begin{aligned}&\text{average absolute predictive value difference}\\ &= \tfrac{1}{2}\left|P(Y = \text{fav} \mid \hat{y}(X) = \text{fav},\ Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{fav},\ Z = \text{priv})\right|\\ &+ \tfrac{1}{2}\left|P(Y = \text{fav} \mid \hat{y}(X) = \text{unf},\ Z = \text{unpr}) - P(Y = \text{fav} \mid \hat{y}(X) = \text{unf},\ Z = \text{priv})\right|.\end{aligned}$$

Equation 10.5

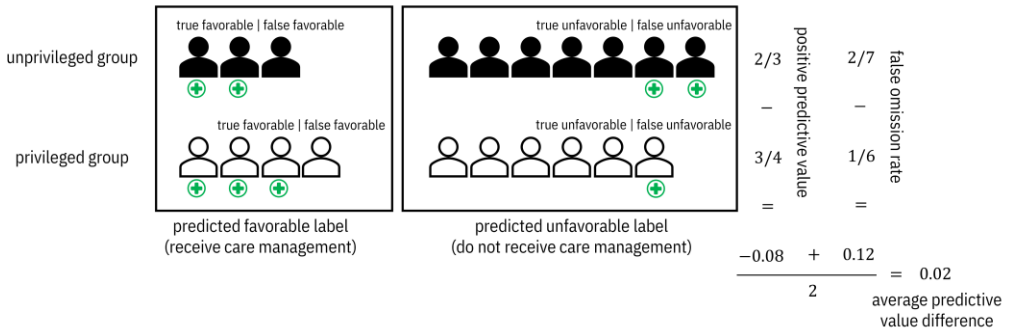

Figure 10.7. *An example calculation of average predictive value difference. The crosses below the members indicate a true need for care management.* Accessible caption. In the unprivileged group, 2 members receive true favorable outcomes and 1 receives a false unfavorable outcome, giving a 2/3 positive predictive value. In the privileged group, 3 members receive true favorable outcomes and 1 receives a false unfavorable outcome, giving a 3/4 positive predictive value. The positive predictive value difference is −0.08. In the unprivileged group, 2 members receive a false unfavorable outcome and 5 receive a true unfavorable outcome, giving a 2/7 false omission rate. In the privileged group, 1 member receives a false unfavorable outcome and 5 receive a true unfavorable outcome, giving a 1/6 false omission rate. The false omission rate difference is 0.12. Averaging the two differences gives a 0.02 average predictive value difference.

10.3.5 Choosing Between Average Odds and Average Predictive Value Difference

What's the difference between separation and sufficiency? Which one makes more sense for the Sospital care management model? This is not a decision based on politics and worldviews like the decision between independence and separation. It is a decision based on what the favorable label grants the affected user: is it assistive or simply non-punitive?[9] Getting a loan is assistive, but not getting arrested is non-punitive. Receiving care management is assistive. In assistive cases like receiving extra care, separation (equalized odds) is the preferred fairness metric because it relates to recall (true positive rate), which is of primary concern in these settings. If receiving care management had been a non-punitive act, then sufficiency (calibration) would have been the preferred fairness metric because precision is of primary concern in non-punitive settings. (Precision is equivalent to positive predictive value, which is one of the two components of the average predictive value difference.).

10.3.6 Conclusion

You can construct all sorts of different group fairness metrics by computing differences or ratios of the various confusion matrix entries and other classifier performance metrics detailed in Chapter 6, but independence, separation, and sufficiency are the three main ones. They are summarized in Table 10.1.

[9]Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. "On the Applicability of ML Fairness Notions." arXiv:2006.16745, 2020. Boris Ruf and Marcin Detyniecki. "Towards the Right Kind of Fairness in AI." arXiv:2102.08453, 2021.

Table 10.1. *The three main types of group fairness metrics.*

Type	Statistical Relationship	Fairness Metric	Can Be A Dataset Metric?	Social Bias in Measurement	Favorable Label
independence	$\hat{Y} \perp\!\!\!\perp Z$	statistical parity difference	yes	yes	assistive or non-punitive
separation	$\hat{Y} \perp\!\!\!\perp Z \mid Y$	average odds difference	no	no	assistive
sufficiency (calibration)	$Y \perp\!\!\!\perp Z \mid \hat{Y}$	average predictive value difference	no	no	non-punitive

Based on the different properties of the three group fairness metrics, and the likely social biases in the data you're using to create the Sospital care management model, you should focus on independence and statistical parity difference.

10.4 Defining Individual and Counterfactual Fairness

An important concept in fairness is *intersectionality*. Things might look fair when you look at different protected attributes separately, but when you define unprivileged groups as the intersection of multiple protected attributes, such as black women, group fairness metrics show unfairness. You can imagine making smaller and smaller groups by including more and more attributes, all the way to a logical end of groups that are just individuals that share all of their feature values. At this extreme, the group fairness metrics described in the previous section are no longer meaningful and a different notion of sameness is needed. That notion is *individual fairness* or *consistency*: that all individuals with the same feature values should receive the same predicted label and that individuals with similar features should receive similar predicted labels.

10.4.1 Consistency

The consistency metric is quantified as follows:

$$\text{consistency} = 1 - \frac{1}{n}\sum_{j=1}^{n}\left|\hat{y}_j - \frac{1}{k}\sum_{j' \in \mathcal{N}_k(x_j)} \hat{y}_{j'}\right|.$$

Equation 10.6

For each of the n Sospital members, the prediction $\hat{y}_j$ is compared to the average prediction of the k nearest neighbors. When the predicted labels of all of the k nearest neighbors match the predicted label of the person themselves, you get 0. If all of the nearest neighbor predicted labels are different from the predicted label of the person, the absolute value is 1. Overall, because of the 'one minus' at the beginning of Equation 10.6, the consistency metric is 1 if all similar points have similar labels and less than 1 if similar points have different labels.

The biggest question in individual fairness is deciding the distance metric by which the nearest neighbors are determined. Which kind of distance makes sense? Should all features be used in the distance computation? Should protected attributes be excluded? Should some feature dimensions be corrected for in the distance computation? These choices are where politics and worldviews come into play.[10] Typically, protected attributes are excluded, but they don't have to be. If you believe there is no bias during measurement (the "what you see is what you get" worldview), then you should simply use the features as is. In contrast, suppose you believe that there are structural social biases in measurement (the "we're all equal" worldview). In that case, you should attempt to undo those biases by correcting the features as they're fed into a distance computation. For example, if you believe that blacks with three outpatient doctor visits are equal in health to whites with five outpatient doctor visits, then your distance metric can add two outpatient visits to the black members as a correction.

10.4.2 Counterfactual Fairness

One special case of individual fairness is when two patients have exactly the same feature values and only differ in one protected attribute. Think of two patients, one black and one white who have an identical history of interaction with the health care system. The situation is deemed fair if both receive the same predicted label—either both are given extra care management or both are not given extra care management—and unfair otherwise. Now take this special case a step further. As a thought experiment, imagine an intervention $do(Z)$ that changes the protected attribute of a Sospital member from black to white or vice versa. If the predicted label remains the same for all members, the classifier is *counterfactually fair*.[11] (Actually intervening to change a member's protected attribute is usually not possible immediately, but this is just a thought experiment.) Counterfactual fairness can be tested using treatment effect estimation methods from Chapter 8.

Protected attributes *causing* different outcomes across groups is an important consideration in many laws and regulations.[12] Suppose you have a full-blown causal graph of all the variables given to you or you discover one from data using the methods of Chapter 8. In that case, you can see which variables have causal paths to the label nodes, either directly or passing through other variables. If any of the variables with causal paths to the label are considered protected attributes, you have a fairness problem to investigate and mitigate.

10.4.3 Theil Index

If you don't want to decide between group and individual fairness metrics as you're figuring out the Sospital care management problem specification, do you have any other options? Yes you do. You can use the Theil index, which was first introduced in Chapter 3 as a summary statistic for uncertainty. It naturally combines both individual and group fairness considerations. Remember from that chapter that the Theil index was originally developed to measure the distribution of wealth in a society. A value

[10]Reuben Binns. "On the Apparent Conflict Between Individual and Group Fairness." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain, Jan. 2020, pp. 514–524.
[11]Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. "Causal Reasoning for Algorithmic Fairness." arXiv:1805.05859, 2018.
[12]Alice Xiang. "Reconciling Legal and Technical Approaches to Algorithmic Bias." In: *Tennessee Law Review* 88.3 (2021).

of 1 indicates a totally unfair society where one person holds all the wealth and a value of 0 indicates an egalitarian society where all people have the same amount of wealth.

What is the equivalent of wealth in the context of machine learning and distributive justice in health care management? It has to be some sort of non-negative benefit value b_j that you want to be equal for different Sospital members. Once you've defined the benefit b_j, plug it into the Theil index expression and use it as a combined group and individual fairness metric:

$$\text{Theil index } = \frac{1}{n}\sum_{j=1}^{n}\frac{b_j}{\bar{b}}\log\frac{b_j}{\bar{b}}.$$

Equation 10.7

The equation averages the benefit divided by the mean benefit $\bar{b}$, multiplied by its natural log, across all people.

That's all well and good, but benefit to who and under which worldview? The research group that proposed using the Theil index in algorithmic fairness suggested that b_j be 2 for false favorable labels (false positives), 1 for true favorable labels (true positives), 1 for true unfavorable labels (true negatives), and 0 for false unfavorable labels (false negatives).[13] This recommendation is seemingly consistent with the "what you see is what you get" worldview because it is examining model performance, assumes the costs of false positives and false negatives are the same, and takes the perspective of affected members who want to get care management even if they are not truly suitable candidates. More appropriate benefit functions for the problem specification of the Sospital model may be b_j that are (1) 1 for true favorable and true unfavorable labels and 0 for false favorable and false unfavorable labels ("what you see is what you get" while balancing societal needs), or (2) 1 for true favorable and false favorable labels and 0 for true unfavorable and false unfavorable labels ("we're all equal").

10.4.4 Conclusion

Individual fairness consistency and Theil index are both excellent ways to capture various nuances of fairness in different contexts. Just like group fairness metrics, they require you to clarify your worldview and aim for the same goals in a bottom-up way. Since the Sospital care management setting is regulated using group fairness language, it behooves you to use group fairness metrics in your problem specification and modeling. Counterfactual or causal fairness is a strong requirement from the perspective of the philosophy and science of law, but the regulations are only just catching up. So you might need to utilize causal fairness in problem specifications in the future, but not just yet. As you've learned so far, the problem specification and data phases are critical for fairness. But that makes the modeling phase no less important. The next section focuses on bias mitigation to improve fairness as part of the modeling pipeline.

---

[13]Till Speicher, Hoda Heidari, Nina Grgić-Hlača, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. "A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. London, England, UK, Jul. 2018, pp. 2239–2248.

10.5 Mitigating Unwanted Bias

From the earlier phases of the lifecycle of the Sospital care management model, you know that you must address unwanted biases during the modeling phase. Given the quantitative definitions of fairness and unfairness you've worked through, you know that mitigating bias entails introducing some sort of statistical independence between protected attributes like race and true or predicted labels of needing care management. That sounds easy enough, so what's the challenge? What makes bias mitigation difficult is that other regular predictive features X have statistical dependencies with the protected attributes and the labels (a node for X was omitted from Figure 10.5 and Figure 10.6, but out-of-sight does not mean out-of-mind). The regular features can reconstruct the information contained in the protected attributes and introduce dependencies, even if you do the most obvious thing of dropping the protected attributes from the data. For example, race can be strongly associated both with certain health care providers (some doctors have predominantly black patients and other doctors have predominantly white patients) and with historical selection for extra care management.

Bias mitigation methods *must* be more clever than simply dropping protected attributes. Don't take a shortcut: dropping protected attributes is never the right answer. Remember the two main ways of mitigating the ills of distribution shift in Chapter 9: *adaptation* and *min-max robustness*. When applied to bias mitigation, adaptation-based techniques are much more common than robustness-based ones, but rely on having protected attributes in the training dataset.[14] They are the subject of the remainder of this section. If the protected attributes are not available in the training data, min-max robustness techniques for fairness that mirror those for distribution shift can be used.[15]

Figure 10.8 (a subset of Figure 10.3) shows three different points of intervention for bias mitigation: (1) *pre-processing* which alters the statistics of the training data, (2) *in-processing* which adds extra constraints or regularization terms to the learning process, and (3) *post-processing* which alters the output predictions to make them more fair. Pre-processing can only be done when you have the ability to touch and modify the training data. Since in-processing requires you to mess with the learning algorithm, it is the most involved and least flexible. Post-processing is almost always possible and the easiest to pull off. However, the earlier in the pipeline you are, the more effective you can be.

There are several specific methods within each of the three categories of bias mitigation techniques (pre-processing, in-processing, post-processing). Just like for accuracy, no one best algorithm outperforms all other algorithms on all datasets and fairness metrics (remember the no free lunch theorem). Just like there are differing domains of competence for classifiers covered in Chapter 7, there are differing domains of competence for bias mitigation algorithms. However, fairness is a new field that has not yet been studied extensively enough to have good characterizations of those domains of competence yet. In Chapter 7, it was important to go down into the details of machine learning methods

[14]The assumption that training datasets contain protected attributes can be violated for regulatory or privacy reasons. The situation is known as *fairness under unawareness*. See: Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. "Fairness Under Unawareness." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Atlanta, Georgia, USA, Jan. 2019, pp. 339–348.

[15]Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. "Fairness Without Demographics in Repeated Loss Minimization." In: *Proceedings of the International Conference on Machine Learning*. Stockholm, Sweden, Jul. 2018, pp. 1929–1938. Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. "Fairness Without Demographics through Adversarially Reweighted Learning." In: *Advances in Neural Information Processing Systems* 33 (Dec. 2020), pp. 728–740.

because that understanding is used in this and other later chapters. The reason to dive into the details of bias mitigation algorithms is different. In choosing a bias mitigation algorithm, you have to (1) know where in the pipeline you can intervene, (2) consider your worldview, and (3) understand whether protected attributes are allowed as features and will be available in the deployment data when you are scoring new Sospital members.

Figure 10.8. *Three types of bias mitigation algorithms in different parts of the machine learning pipeline.* Accessible caption. A block diagram with a training dataset as input to a bias mitigation pre-processing block with a pre-processed dataset as output. The pre-processed dataset is input to a bias mitigation in-processing block with an initial model as output. The initial model is input to a bias mitigation post-processing block with a final model as output.

10.5.1 Pre-Processing

At the pre-processing stage of the modeling pipeline, you don't have the trained model yet. So pre-processing methods cannot explicitly include fairness metrics that involve model predictions. Therefore, most pre-processing methods are focused on the "we're all equal" worldview, but not exclusively so. There are several ways for pre-processing a training data set: (1) augmenting the dataset with additional data points, (2) applying instance weights to the data points, and (3) altering the labels.

One of the simplest algorithms for pre-processing the training dataset is to append additional rows of made-up members that do not really exist. These imaginary members are constructed by taking existing member rows and flipping their protected attribute values (like counterfactual fairness).[16] The augmented rows are added sequentially based on a distance metric so that 'realistic' data points close to modes of the underlying dataset are added first. This ordering maintains the fidelity of the data distribution for the learning task. A plain uncorrected distance metric takes the "what you see is what you get" worldview and only overcomes sampling bias, not measurement bias. A corrected distance metric like the example described in the previous section (adding two outpatient visits to the black members) takes the "we're all equal" worldview and can overcome both measurement and sampling bias (threats to both construct and external validity). This data augmentation approach needs to have protected attributes as features of the model and they must be available in deployment data.

Another way to pre-process the training data set is through sample weights, similar to inverse probability weighting and importance weighting seen in Chapter 8 and Chapter 9, respectively. The *reweighing* method is geared toward improving statistical parity ("we're all equal" worldview), which can be assessed before the care management model is trained and is a dataset fairness metric.[17] The goal of

[16]Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. "Data Augmentation for Discrimination Prevention and Bias Disambiguation." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, New York, USA, Feb. 2020, pp. 358–364.

[17]Faisal Kamiran and Toon Calders. "Data Preprocessing Techniques for Classification without Discrimination." In: *Knowledge and Information Systems* 33.1 (Oct. 2012), pp. 1–33.

independence between the label and protected attribute corresponds to their joint probability being the product of their marginal probabilities. This product probability appears in the numerator and the actual observed joint probability appears in the denominator of the weight:

$$w_j = \frac{p_Y(y_j)p_Z(z_j)}{p_{Y,Z}(y_j, z_j)}.$$

Equation 10.8

Protected attributes are required in the training data to learn the model, but they don't have to be part of the model or the deployment data.

Whereas data augmentation and reweighing do not change the training data you have from historical care management decisions, other methods do. One simple method, only for statistical parity and the "we're all equal" worldview, known as *massaging* flips unfavorable labels of unprivileged group members to favorable labels and favorable labels of privileged group members to unfavorable labels.[18] The chosen data points are those closest to the decision boundary that have low confidence. Massaging does not need to have protected attributes in the deployment data.

A different approach, the *fair score transformer*, works on (calibrated) continuous score labels $S = p_{Y|X}(Y = fav \mid x)$ rather than binary labels.[19] It is posed as an optimization in which you find transformed scores S' that have small cross-entropy with the original scores S, i.e. $H(S \parallel S')$, while constraining the statistical parity difference, average odds difference, or other group fairness metrics of your choice to be of small absolute value. You convert the pre-processed scores back into binary labels with weights to feed into a standard training algorithm. You can take the "what you see is what you get" worldview with the fair score transformer because it assumes that the classifier later trained on the pre-processed dataset is competent, so that the pre-processed score it produces is a good approximation to the score predicted by the trained model. Although there are pre-processing methods that alter both the labels and (structured or semi-structured) features,[20] the fair score transformer proves that you only need to alter the labels. It can deal with deployment data that does not come with protected attributes.

Data augmentation, reweighing, massaging, and fair score transformer all have their own domains of competence. Some perform better than others on different fairness metrics and dataset characteristics. You'll have to try different ones to see what happens on the Sospital data.

---

[18]Faisal Kamiran and Toon Calders. "Data Preprocessing Techniques for Classification without Discrimination." In: *Knowledge and Information Systems* 33.1 (Oct. 2012), pp. 1–33.

[19]Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P. Calmon. "Optimized Score Transformation for Fair Classification." In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. Aug. 2020, pp. 1673–1683.

[20]Some examples are the methods described in the following three papers. Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and Removing Disparate Impact." In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney, Australia, Aug. 2015, pp. 259–268. Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. "Optimized Pre-Processing for Discrimination Prevention." In: *Advances in Neural Information Processing Systems* 30 (Dec. 2017), pp. 3992–4001. Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. "Fairness GAN: Generating Datasets with Fairness Properties Using a Generative Adversarial Network." In: *IBM Journal of Research and Development* 63.4/5 (Jul./Sep. 2019), p. 3.

10.5.2 In-Processing

In-processing bias mitigation algorithms are straightforward to state, but often more difficult to actually optimize. The statement is as follows: take an existing risk minimization supervised learning algorithm, such as (a repetition of Equation 7.4):

$$\hat{y}(\cdot) = arg\ \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} L\left(y_j, f\left(x_j\right)\right) + \lambda J(f)$$

Equation 10.9

and regularize or constrain it using a fairness metric. The algorithm can be logistic regression and the regularizer can be statistical parity difference, in which case you have the *prejudice remover*.[21] More recent fair learning algorithms are broader and allow for any standard risk minimization algorithm along with a broad set of group fairness metrics as constraints that cover the different types of fairness.[22] A recent in-processing algorithm regularizes the objective function using a causal fairness term. Under strong ignorability assumptions (remember from Chapter 8 that these are no unmeasured confounders and overlap), the regularizer is an average treatment effect-like term $J = E[Y \mid do(z = 1), X] - E[Y \mid do(z = 0), X]$.[23]

Once trained, the resulting models can be used on new unseen Sospital members. These in-processing algorithms do not require the deployment data to contain the protected attribute. The trick with all of them is structuring the bias mitigating regularization term or constraint so that the objective function can tractably be minimized through an optimization algorithm.

10.5.3 Post-Processing

If you're in the situation that the Sospital care management model has already been trained and you cannot change it or touch the training data (for example if you are purchasing a pre-trained model from a vendor to include in your pipeline), then the only option you have is to mitigate unwanted biases using post-processing. You can only alter the output predictions $\hat{Y}$ to meet the group fairness metrics you desire based on your worldview (i.e. flipping the predicted labels from receiving care management to not receiving care management and vice versa). If you have some validation data with labels, you can post-process with the "what you see is what you get" worldview. You can always post-process with the "we're all equal" worldview, with or without validation data.

---

[21]Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. "Fairness-Aware Classifier with Prejudice Remover Regularizer." In: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Bristol, England, UK, Sep. 2012, pp. 35–50.

[22]Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. "A Reductions Approach to Fair Classification." In: *Proceedings of the International Conference on Machine Learning*. Stockholm, Sweden, Jul. 2018, pp. 60–69. L. Elisa Celis, Lingxiao Huang, Vijay Kesarwani, and Nisheeth K. Vishnoi. "Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Atlanta, Georgia, USA, Jan. 2019, pp. 319–328. Ching-Yao Chuang and Youssef Mroueh. "Fair Mixup: Fairness via Interpolation." In: *Proceedings of the International Conference on Learning Representations*. May 2021.

[23]Pietro G. Di Stefano, James M. Hickey, and Vlasios Vasileiou. "Counterfactual Fairness: Removing Direct Effects Through Regularization." arXiv:2002.10774, 2020.

Since group fairness metrics are computed on average, flipping any random member's label within a group is the same as flipping any other random member's.[24] A random selection of people, however, seems to be procedurally unfair. To overcome this issue, similar to massaging, you can prioritize flipping the labels of members whose data points are near the decision boundary and are thus low confidence samples.[25] You can also choose people within a group so that you reduce individual counterfactual unfairness.[26] All of these approaches require the protected attribute in the deployment data.

The fair score transformer described in the pre-processing section also has a post-processing version, which does not require the protected attribute and should be considered the first choice algorithm in the category of post-processing bias mitigation if the base classifier outputs continuous scores. It performs well empirically and is not computationally-intensive. Just like the pre-processing version, the idea is to find an optimal transformation of the predicted score output into a new score, which can then be thresholded to a binary prediction for the final care management decision that Sospital makes.

10.5.4 Conclusion

All of the different bias mitigation algorithms are options as you're deciding what to finally do in the care management modeling pipeline. The things you have to think about are:

1. where in the pipeline can you make alterations (this will determine the category pre-, in-, or post-processing)
2. which worldview you've decided with the problem owner (this will disallow some algorithms that don't work for the worldview you've decided)
3. whether the deployment data contains the protected attributes (if not, this will disallow some algorithms that require them).

These different decision points are summarized in Table 10.2. After that, you can just go with the algorithm that gives you the best quantitative results. But what is best? It is simply the pipeline with the best value for the fairness metric you've chosen in your problem specification.

But you might ask, shouldn't I consider a tradeoff of fairness and accuracy when I choose the pipeline? Balancing tradeoffs and relationships among different elements of trustworthy machine learning is more fully covered in Chapter 14, but before getting there, it is important to note one important point. Even though it is a convenient shortcut, measuring classification accuracy on data from the prepared data space, which already contains social bias, representation bias, and data preparation bias is not the right thing to do. Just like you should measure performance of distribution shift adaptation on data from the new environment—its construct space, you should measure accuracy after bias mitigation in its construct space where there is no unfairness. There is a tradeoff between fairness and accuracy measured in the prepared data space, but importantly there is no tradeoff between

---

[24]Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. "On Fairness and Calibration." In: *Advances in Neural Information Processing Systems* 31 (Dec. 2017), pp. 5684–5693.

[25]Faisal Kamiran, Asim Karim, and Xiangliang Zhang. "Decision Theory for Discrimination-Aware Classification." In: *Proceedings of the IEEE International Conference on Data Mining*. Brussels, Belgium, Dec. 2012, pp. 924–929.

[26]Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. "Bias Mitigation Post-Processing for Individual and Group Fairness." In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Brighton, England, UK, May 2019, pp. 2847–2851.

accuracy and fairness in the construct space.[27] You can approximate a construct space test set by using the data augmentation pre-processing method.

Table 10.2. *Characteristics of the main bias mitigation algorithms.*

Algorithm	Category	Fairness	Protected Attributes in Deployment Data
data augmentation	pre	counterfactual	yes
reweighing	pre	independence	no
massaging	pre	independence	no
fair score transformer	pre, post	independence, separation	no
prejudice remover	in	independence	no
recent in-processing algorithms	in	independence, separation, sufficiency	no
causal regularizer	in	counterfactual	no
group fairness post-processing	post	independence, separation	yes
individual and group fairness post-processing	post	counterfactual, independence, separation	yes

In your Sospital problem, you have almost complete flexibility because you do control the training data and model training, are focused on independence and the "we're all equal" worldview, and are able to include protected attributes for Sospital's members in the deployment data. Try everything, but start with the fair score transformer pre-processing.

10.6 Other Considerations

Before concluding the chapter, let's consider a couple other issues. The first did not come up in the Sospital care management use case, but can come up in other use cases. The Sospital problem lent itself to fairness in the context of direct allocation decisions, but that is not the only possibility. There are also harms in representation or quality-of-service, such as bias in search results. For example, image searches for professions might yield only white people, web search results for personal names overrepresented in the black community might be accompanied by advertisements for criminal defense attorneys, and natural language processing algorithms for language translation or query understanding might associate doctors with men and nurses with women automatically. Some of the bias mitigation algorithms for allocative fairness can be used in representational fairness, but different techniques may be more appropriate.

[27]Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. "Unlocking Fairness: A Trade-Off Revisited." In: *Advances in Neural Information Processing Systems* 32 (Dec. 2019), pp. 8783–8792. Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. "Empirical Observation of Negligible Trade-Offs in Machine Learning for Public Policy." In: *Nature Machine Intelligence* 3 (Oct. 2021), pp. 896–904.

> "Most of this work is narrow in scope, focusing on fine-tuning specific models, making datasets more inclusive/representative, and 'debiasing' datasets. Although such work can constitute part of the remedy, a fundamentally equitable path must examine the wider picture, such as unquestioned or intuitive assumptions in datasets, current and historical injustices, and power asymmetries."
>
> —Abeba Birhane, cognitive scientist at University College Dublin

> "I continue to worry that in CS (as in psychology), debates about bias have become a powerful distraction—drawing attention away from what's most important toward what's more easily measurable."
>
> —J. Nathan Matias, behavioral scientist at Cornell University

The second issue is as follows. Have we too easily swept the important considerations of algorithmic fairness under the rug of mathematics? Yes and no. If you have truly thought through the different sources of inequity arising throughout the machine learning lifecycle utilizing a panel of diverse voices, then applying the quantitative metrics and mitigation algorithms is actually pretty straightforward. It is straightforward because of the hard work you've done before getting to the modeling phase of the lifecycle and you should feel confident in going forward. If you have not done the hard work earlier in the lifecycle (including problem specification), blindly applying bias mitigation algorithms might not reduce harms and can even exacerbate them. So don't take shortcuts.

10.7 Summary

- Fairness has many forms, arising from different kinds of justice. Distributive justice is the most appropriate for allocation decisions made or supported by machine learning systems. It asks for some kind of sameness in the outcomes across individuals and groups.
- Unfairness can arise from problem misspecification (including inappropriate proxy labels), feature engineering, measurement of features from the construct space to the observed space, and sampling of data points from the observed space to the raw data space.
- There are two important worldviews in determining which kind of sameness is most appropriate for your problem.
- If you believe there are social biases in measurement (not only representation biases in sampling), then you have the "we're all equal" worldview; independence and statistical parity difference are appropriate notions of group fairness.
- If you believe there are no social biases in measurement, only representation biases in sampling, then you have the "what you see is what you get" worldview; separation, sufficiency, average odds difference, and average predictive value difference are appropriate notions of group fairness.
- If the favorable label is assistive, separation and average odds difference are appropriate notions of group fairness. If the favorable label is non-punitive, sufficiency and average predictive value difference are appropriate notions of group fairness.

- Individual fairness is a limiting version of group fairness with finer and finer groups. Worldviews play a role in determining distance metrics between individuals.
- Bias mitigation algorithms can be applied as pre-processing, in-processing, or post-processing within the machine learning pipeline. Different algorithms apply to different worldviews. The choice of algorithm should consider the worldview in addition to empirical performance.