Ethical Challenges in Machine Learning: A Practical Guide addressing issues such as Bias, Fairness, and Self-Fulfilling Predictions

-Master Thesis-

SoSe 2025

Supervisor:

Dr. Karsten Tolle

Submitted by:

Lukas Welikat

Business Informatics M.Sc.

Submission date: 23.05.2025

Abstract

Machine learning systems increasingly influence decision-making processes. Their deployment can raise ethical concerns that are especially important in critical areas such as education, healthcare or criminal justice. Fundamental concepts that are discussed in this thesis are fairness, bias, and self-fulfilling predictions. They will be examined through a structured analysis of three practical studies, each highlighting specific aspects of fairness in machine learning.

The theoretical part of the thesis introduces definitions of fairness and discrimination, key approaches to fairness, types of bias that are relevant for machine learning applications, and the implications of self-fulfilling predictions. These concepts are linked to practical studies illustrating ethical considerations as well as the challenges in achieving fairness in machine learning. The studies reveal that ensuring fairness is not straightforward, as conflicting metrics, feedback loops, and different contextual biases complicate equitable outcomes. Building on these insights, the broader challenges identified in the studies are discussed and a critical analysis of their implications is provided. It highlights, for example, the incompatibility of some fairness definitions, contextual dependencies, and the sociotechnical nature of algorithmic decision-making. Based on these findings, the thesis provides practical recommendations for addressing fairness issues. It emphasizes the need for procedural, context-sensitive approaches. Transparency, diversity, interdisciplinary collaboration and a consequentialist view are additional factors to successfully assess fairness in machine learning systems.

To translate theoretical insights into practice, the thesis is accompanied by an interactive notebook, designed as a teaching tool for computer science students. The notebook represents a condensed and interactive version of this thesis and is divided into seven parts. It includes theoretical sections, but also executable code, exercises, and reflective questions for better understanding. Its structured layout and typographic emphasis support cognitive processing and enhance the learning experience in comparison to purely theoretical input.

This work underlines the complexity of achieving fairness in machine learning. It demonstrates that ethical assessment of automated systems requires not only technical solutions but also a deep understanding of the sociotechnical system that surrounds the machine learning application.

Table of Content

List of Tables

List of Figures

List of Abbreviations

COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
AI	Artificial Intelligence
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
PPV	Positive Predictive Value
FDR	False Discovery Rate
FOR	False Omission Rate
NPV	Negative Predictive Value
TPR	True Positive Rate
TNR	True Negative Rate
FPR	False Positive Rate
FNR	False Negative Rate
FRAIA	Fundamental Rights and Algorithms Impact Assessment

1 Introduction

1.1 Motivation

Machine learning is becoming more and more important in modern-day decision-making. It influences many aspects in people's lives such as lending, hiring, law enforcement, healthcare, and everyday interactions like search rankings or personalized recommendations. On the one hand these systems come with many benefits, such as efficiency, scalability or the ability to uncover complex relationships in data that humans might miss. On the other hand, they can also introduce serious risks. If machine learning models are not carefully designed and monitored, they can reinforce existing biases, contribute to discrimination and produce unfair outcomes that affect some societal groups more than others (Mehrabi et al., 2021).

One of the main challenges is that these models primarily learn patterns from historical data, often with little or no human oversight. When this data reflects societal inequalities, the algorithm can not only replicate but also amplify these biases or even lead to new forms of bias. Especially errors in domains with big impact on individuals, like healthcare, education or law enforcement, need to be recognized and mitigated. This is evident, for example, in predictive policing systems like PredPol, where historical crime data is used to forecast crime patterns. This can lead to increased law enforcement presence in already heavily policed neighborhoods, and with that reinforcing existing patterns of surveillance and suspicion (Osoba & Welser, 2017).

In a similar way, risk assessment tools like COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) have shown biased behavior. COMPAS is used in criminal justice systems to predict the likelihood of defendants recommitting a crime. A systematic analysis of the system by ProPublica revealed that COMPAS misclassifies black defendants as high-risk disproportionately more often than white defendants. This happened even though the system was intended to reduce judicial bias by explicitly avoiding race as an input (Angwin et al., 2016). Using biased systems like COMPAS or PredPol can create cycles of discrimination and reinforce existing societal inequalities (Barocas & Selbst, 2016).

Such biases are often unintentional but can have far-reaching consequences. Algorithmic decision-making can lead to allocative harms, where opportunities or resources such as loans or jobs are withheld from a certain population. These are often easier to detect than representational harms, where stereotypes are reinforced in automated systems (Barocas et al., 2023). An example of representational harm can be found in facial recognition technology. A

study by Buolamwini and Gebru (2018) demonstrated that many gender classification tools performed better on male faces than female faces and even worse on dark-skinned women. This is because minority groups are often underrepresented in training datasets, leading to poor model performance for these groups.

Underrepresentation of marginalized communities in datasets used for machine learning can also lead to the exclusion of such communities from data-driven decision making. The pothole detection app Street Bump, used in Boston, illustrates this issue. The purpose of the app was to help the city with the detection of potholes by collecting reports from smartphone users. The data collection was biased, because residents from wealthier areas were more likely to own a smartphone. As a result, poorer neighborhoods, where less people had access to the app, were underrepresented in the data and less roads got repaired in those areas (Barocas & Selbst, 2016).

Moreover, machine learning systems work with statistical guarantees rather than moral or ethical reasoning (Osoba & Welser, 2017). At its core, machine learning reflects a system of probabilities that come from data with a feedback loop that enables learning, so the model can modify its approach on whether an outcome is correct or not (Yapo & Weiss, 2018). However, this feedback loop can also amplify existing biases. In hiring systems, companies only receive feedback on the candidates they hire, not on those who were rejected. Predictive policing systems target certain areas, often due to already biased data, and with more officers in the area more crime gets reported. This confirms the model's prediction influenced by prior biases and results in a self-reinforcing system (Ensign et al., 2017).

Machine learning makes decisions based on statistical inference, which can lead to statistical discrimination. Instead of assessing individuals on their unique characteristics, algorithmic models generalize based on group attributes. This raises ethical concerns about fairness and individual merit. Algorithmic decisions fail to treat people as individuals by design. While such generalizations may sometimes be statistically sound and necessary, they can only be morally acceptable if they are sufficiently accurate and do not create systematic disadvantages (Binns, 2018).

Another challenge is the lack of transparency in machine learning systems. Many models, especially those developed by private companies, are highly opaque. This makes it difficult to understand how decisions are made or to contest them when they seem unfair. Missing explainability does have a negative impact on trust in machine learning systems and makes it harder to detect and address biases. Individuals that do not have insight into how and why a

decision was made lose the ability to advocate for themselves. Machine learning systems operate at speeds and levels of complexity beyond human intuition, sometimes even their creators cannot fully predict their outcomes (Yapo & Weiss, 2018). The effects of machine learning bias can be subtle and intangible, so much that individuals which are affected by biased decisions may not even realize they were disadvantaged. Unlike explicit discrimination, which is easier to detect, algorithmic bias is often manifested through proxies – correlations in data that unintentionally replicate societal inequalities. Because these proxies are harder to identify, well-intentioned machine learning applications can still reinforce structural biases unnoticed and unchallenged (Wachter et al., 2021).

At the same time, the opacity of these systems also leads to a lack of accountability. When decisions are made by humans, in most cases it is easy to assign responsibility. But with algorithmic decision-making, it is often unclear who should be held accountable when outcomes are unfair or harmful. Is it the developers who designed the model, the organizations that deployed it, or the data sources that shaped it? This uncertain responsibility makes it hard to challenge decisions or implement corrective measures (Osoba & Welser, 2017). On top of that, ethically problematic effects of algorithmic decisions and their impact are often recognized only retrospectively (Howard & Borenstein, 2018).

As machine learning becomes more important in decision-making, power shifts away from decision subjects and domain experts towards machine learning specialists and policymakers (Barocas et al., 2023). In the past, people could challenge decisions by human decision-makers by providing explanations or considering contextual factors. Less accountability, together with missing transparency, leads to less causal understanding behind algorithmic decisions, making it harder to understand why certain outcomes occur and how they can be corrected. This results in hindered public discourse (Osoba & Welser, 2017).

It is important to recognize that human decision-making is also full of biases. Cognitive biases are a heuristic that helps humans to make efficient decisions. These biases can be useful, allowing people to filter information quickly and make judgements in complex situations (Alelyani, 2021, Kahneman, 2012). When machine learning models adopt these biases from training data, they can automate and scale harmful biases in ways that are difficult to detect and correct. Unlike human decision-makers, machine learning systems do not have the ability to distinguish between useful correlations and harmful generalizations. They fail to recognize when predictions reinforce societal inequalities. Machines do not have a moral compass to

detect stereotypical treatment, and they prioritize predictive accuracy over ethical considerations. While humans can consider which attributes are morally relevant and won't make absurd decisions, machines easily could (Barocas et al., 2023).

Despite the presented risks, machine learning is often perceived by society as objective, infallible and bias free (Osoba & Welser, 2017). Such overconfidence can lead to uncritical acceptance of algorithmic decisions and therefore can strengthen and legitimize biases as global truth (Howard & Borenstein, 2018). In reality, machine learning models are only as unbiased as the data they are trained on and the assumptions used in their design. Addressing fairness in machine learning is more than just a technical challenge, it also needs an ethical perspective. These systems need to be more transparent, accountable, and context-aware to prevent them from strengthening existing social inequalities.

Assessing machine learning fairness might produce systems that are not only optimized for accuracy and efficiency but also acknowledge ethical principles. With the increasing role of machine learning in society, dealing with fairness is important to ensure that machine learning technologies benefit everyone, rather than reinforcing historical disadvantages.

1.2 Objective and Structure

Objective

The objective of this thesis is to give an overview of ethical challenges in machine learning. It aims to provide a deeper understanding of key issues such as fairness and bias, as well as specific phenomena such as self-fulfilling predictions and feedback loops. An interactive Jupyter notebook has been developed as a practical outcome that summarizes the key concepts of this thesis in a compact and interactive format. It is available on GitHub at: https://github.com/LukasWel/ethical-challenges-in-ml. Images of the notebook can be found in Appendix C-I. The goal of the notebook is to raise awareness among computer science students and encourage critical reflection on these topics. Students should get a feeling for the complexity and societal implications of the problems described. The notebook is intended as a learning tool for university courses and mirrors the structure of this thesis. By using real-world data for the case studies and hands-on exercises, the notebook supports theoretical understanding but also allows students to critically question the implications of machine learning in practice.

The focus of this work lies on the broader issues of fairness and bias in machine learning. It aims to raise awareness of the challenges, rather than to provide concrete solutions for individual cases. The thesis can be seen as a starting point to get into the topic of ethical issues in machine learning. Specific mitigation strategies are therefore not covered in depth. Legal aspects, such as compliance with AI regulations, are also excluded, as this area is in a state of continuous development. Furthermore, this work concentrates on unfairness that affects humans in a direct way. Broader ethical concerns beyond human fairness, such as the environmental impact of machine learning in general, are not addressed.

Structure

The relevance of the topic was highlighted in the motivation section. In the following chapters, the thesis is divided into two major parts. The first part introduces the fundamental concepts necessary to understand ethical challenges in machine learning (Chapter 2). It begins with a discussion about definitions of fairness and relevant forms of bias. Then it introduces feedback loops and self-fulfilling predictions. Each of these thematic blocks is illustrated with case studies, which are also part of the Jupyter notebook.

The second part (Chapter 3) of the thesis explores why achieving fairness in machine learning is difficult. It discusses conceptual, technical, and contextual challenges and provides some

general recommendations. These recommendations are intentionally kept at a higher level, as the topic of fairness interventions is still evolving and far from settled. In the final Chapter 4 the design considerations and didactic goals of the Jupyter notebook are presented.

2 Basic Concepts of Fair Machine Learning

This chapter introduces the foundational ideas of fair machine learning. It begins by discussing key concepts such as fairness and discrimination. Following this, various fairness metrics are introduced and assessed using the first case study. The next section focuses on different types of bias that are relevant in machine learning, which is further explored in the second study. The concepts of self-fulfilling predictions and feedback loops conclude this chapter. They are illustrated in the third and final case study.

2.1 Fairness & Discrimination

Fairness and discrimination are often used interchangeably. They are related but still distinct concepts. Fairness is the goal that wants to make sure individuals and groups are treated equitably. Discrimination, on the other hand, is a violation of this goal. This violation often results in unfair outcomes. In the context of machine learning, fairness is used as a concept to evaluate the impact of algorithmic decision-making and discrimination is the core concern (Barocas et al., 2023). Several factors influence fairness, such as transparency, accountability, explainability, and bias. All these aspects impact how fair a machine learning system is perceived, but bias is most important in contributing to discrimination.

2.1.1 Concept of fairness

The concept of fairness has been debated in philosophy, psychology and law long before computer science. Still there is no universal definition of fairness, showing how complex it is to solve this problem. Different cultures, legal and ethical perspectives influence how fairness is understood and applied, which complicates defining a single universal fairness definition that everyone accepts even more. In computer science fairness is often translated into mathematical constraints for algorithms, but there is no consensus on which fairness notions are best suited for specific problems (Mehrabi et al., 2021).

The appropriate notion of fairness depends on the domain and context. What is considered fair in hiring decisions might be different from what is fair in criminal justice or credit scoring (Binns, 2018). Although universal agreement is missing, fairness in decision-making can be broadly understood as the “absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making” (Mehrabi et al., 2021, p. 11). This principle is difficult to implement in practice, due to conflicting notions of fairness and trade-offs between different fairness goals. Statistical fairness criteria, such as demographic parity, equalized odds, or equal opportunity are often used to assess fairness in

machine learning. However, these definitions do not include the broader moral and ethical considerations that would be necessary for fair decision-making (Barocas et al., 2023).

One influential framework for thinking about fairness in decision-making is equality of opportunity, which displays different views on how fairness should be interpreted and applied in practice.

The narrow view focuses on individual fairness. It demands that similar individuals should be treated similarly based on relevant factors for a decision, like job performance. Unequal group outcomes are not seen as unfair when the decision was based on relevant skills or qualifications. This view connects to the concept of meritocracy, where people are judged according to their abilities and ambitions and not on social factors like race or gender. Though defining merit and similarity is subjective and depends on the goals of the decision-maker (Barocas et al., 2023).

The broad view is not concerned with single decision-making processes, it focuses on structural equality. This view wants to make sure that all individuals, regardless of their background, have equal opportunities from the beginning, by redistributing resources. This perspective aligns with egalitarian principles, demanding policies like universal access to education and healthcare to prevent the arise of inequalities in the first place. The broad view represents an idealistic approach to fairness, its scope is mostly too extensive for direct implementation in decision-making systems. Nevertheless, the idea behind this perspective supports real-world efforts such as reparation payments (Barocas et al., 2023).

The middle view balances the two previous views. It concentrates on the single decision level but recognizes that fairness also involves addressing historical disadvantages. In this perspective, decision-making should account for structural barriers that may have influenced an individual’s opportunities. The middle view is particularly valuable when designing new decision-making processes, because it considers historical inequalities and structural barriers from the beginning. For example, in university admissions, considering the socio-economic background of applicants can help increase diversity and equality of opportunity (Barocas et al., 2023).

Just as different statistical fairness metrics are appropriate in different contexts, the right interpretation of fairness depends on the specific situation. There is no one-size-fits-all approach, what constitutes fairness in one setting may not be applicable in another.

2.1.2 Concept of discrimination

While fairness is a relatively universal principle, algorithmic discrimination is significantly different from discrimination in a classical way. Traditional discrimination theories often focus on the intentions, beliefs, and moral character of human decision-makers. The mental state is used as a condition for discrimination. Algorithmic systems on the other hand do not directly have these intentions or beliefs, neither do they understand morality. Since they rely on statistical correlations and machine consciousness is not there yet, it is difficult to assess algorithmic discrimination with traditional moral frameworks. The outcome of these systems can still be discriminatory. Because of these differences, algorithmic fairness has been connected to egalitarian principles, which focus on addressing inequalities regardless of the intent (Binns, 2018).

Discrimination occurs when one group is systematically disadvantaged compared to others. In machine learning this arises when algorithms implicitly or explicitly use sensitive attributes, in ways that lead to unfair outcomes. Sensitive attributes are legally protected categories, like race, gender, age, which should ideally not influence decision-making. While traditional discrimination is often more direct and easier to detect, statistical discrimination in machine learning is mostly unintentional and harder to identify (Barocas et al., 2023).

There are two forms in which discrimination in machine learning can occur. Direct discrimination leads to unequal outcomes by explicitly using protected attributes. This type of discrimination is often illegal, as decisions based on race, gender or other sensitive characteristics are forbidden by law. With indirect discrimination unequal treatment arises even though it seems like decisions are made using only neutral attributes. For example, ZIP codes in loan applications are seemingly neutral attributes, but as they are correlated with race in the US, they can indirectly lead to racial disparities. Even when a decision-making algorithm does not explicitly use sensitive attributes, discrimination can still occur when it relies on proxy attributes, such as ZIP codes in the example above. Indirect discrimination often happens without intention and is harder to detect, but as previously mentioned, intention is secondary when assessing machine learning systems (Mehrabi et al., 2021).

Machine learning in general involves a form of statistical discrimination, because it relies on generalizations to make predictions. This does not imply that machine learning mainly produces unfair outcomes, it depends on whether these generalizations are accurate and justified. Some argue that all decisions involve some form of generalization, whereas others focus on the ethical

risks when failing to recognize people as individuals and make decisions with the help of group averages (Binns, 2018). In the end it is important to notice that some form of statistical discrimination is inevitable, the moral evaluation depends on the consequences. There is explainable discrimination that can be justified, for example, higher insurance costs for younger drivers (Mehrabi et al., 2021).

Discrimination can also be divided into different levels, and every level needs to be approached with different mitigation strategies. Structural discrimination arises from societal factors, such as laws, cultural norms, and institutions. Different treatment in the past can lead to inequalities today and historical disadvantages can persist for centuries, as shown by the Jesuits missions in South America. Areas with former mission activities have higher literacy rate and income even 250 years later. Structural discrimination can also result from seemingly neutral laws that disproportionately harm certain groups. For example, U.S. drug laws affect minorities more. Organizational discrimination occurs within companies, institutions or decision-making units. Hiring policies that prefer certain social groups are a common example of organizational discrimination. When individuals act based on stereotypes it is a case of interpersonal discrimination. This happens mostly indirectly, for example, bias against women having lower brilliance than men and are therefore not suited for certain jobs (Barocas et al., 2023).

Bias in machine learning can reinforce discrimination across all three levels. Among all factors influencing fairness, bias plays the most significant role in shaping discrimination. To understand how bias leads to discrimination is essential for designing fair machine learning algorithms. Before we dive deeper into the different sources and forms of bias in machine learning, the next chapter gives an overview of the most discussed notions of fairness and how they relate to each other.

2.1.3 A collection of fairness metrics

As already stated in Chapter 2.1.1 there is no universal definition of fairness. In recent years scientific literature on fairness in machine learning proposed many different fairness metrics. However, despite this variety of statistical and causal measures, the problem of fairness in machine learning is far from solved.

The large number of fairness definitions can be overwhelming, partly because there is no clear guidance on when to use which metric. Simply satisfying as many notions as possible is not an option, as some definitions are mathematically incompatible and conflict with each other (Chouldechova, 2017; Kleinberg et al., 2016). The suitability of a fairness metric is highly

context-dependent, therefore choosing appropriate notions requires trade-offs, domain expertise and moral reasoning. Fairness metrics should not be seen as a solution for directly improving fairness in machine learning systems, they are more diagnostic tools to reveal unfairness or discrimination (Barocas et al., 2023).

A more detailed discussion about the challenges with fairness definitions will follow in Chapter 3. This section gives an overview of a selection of the most cited metrics and how they are connected to each other. The selected metrics represent the main ideas behind measures of fairness, as many of them are similar in their approach of uncovering unfairness. For a more extensive list of fairness metrics see the paper by Verma and Rubin (2018). Note that as most proposed solutions are built around classification algorithms, the following sections will also focus on classifiers. Some of the definitions are also applicable to other machine learning tasks like clustering, or there are some specific approaches. Still, the field of fairness metrics outside of classification is relatively empty and needs further research.

In general, we can divide approaches into group and individual fairness. Group Fairness is concerned with outcome disparities aggregated by individuals from the same sensitive category. It demands equal treatment of different groups, whereas individual fairness requests similar people receive similar predictions regardless of group membership. Individual fairness compares single outcomes without aggregation. Further we can divide the fairness metrics into observational notions, similarity-based measures and causal fairness approaches (Calegari et al., 2023).

2.1.3.1 Observational fairness

Observational fairness includes metrics that are computed only from observable data. It excludes assumptions about the impact of decisions, the internal logic of a classifier, or correlations between features and outcomes. Observational fairness metrics are the most common form of fairness criteria in machine learning, as they are relatively simple and straightforward to implement (Calegari et al., 2023).

Almost all observational fairness definitions are based on statistical measures derived from the confusion matrix, in particular, true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Each of the eight measures in Table 1 provides valuable insight into the classifier’s performance and is used in fairness definitions (Verma & Rubin, 2018).

Table 1

Core statistical measures

Measure	Formula
Positive Predictive Value / Precision (PPV)	$\frac{TP}{TP + FP}$
False Discovery Rate (FDR)	$\frac{FP}{TP + FP}$
False Omission Rate (FOR)	$\frac{FN}{TN + FN}$
Negative Predictive Value (NPV)	$\frac{TN}{TN + FN}$
True Positive Rate / Sensitivity / Recall (TPR)	$\frac{TP}{TP + FN}$
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$
False Negative Rate (FNR)	$\frac{FN}{TP + FN}$
True Negative Rate (TNR)	$\frac{TN}{TN + FP}$

Note. Adapted from Verma & Rubin (2018).

Independence

We can divide observational fairness into three categories: independence, separation and sufficiency. Independence requires that the predicted outcome $\hat{Y}$ does not statistically depend on the sensitive attribute A. This means that all demographic groups should receive positive predictions at equal rates, regardless of their true label Y (Barocas et al., 2023). Since this category only considers the predicted outcome, it is the easiest and most intuitive form of fairness definitions. The main fairness notion in literature to check for independence is statistical parity (often also called demographic parity). Statistical parity ensures that the probability of being assigned to the positive predicted class is equal for individuals from protected and unprotected groups (Verma & Rubin, 2018). Although intuitive, independence as a fairness criterion has several limitations. It is related to the broad view of equality of opportunity, aiming for an equal distribution of resources. However, enforcing statistical parity ensures short-term equality which does not guarantee long-term equality of outcomes. Equality of outcome ignores individual differences in ability, ambition or context. It also fails to account for how benefits or harms might vary across groups. Being the most basic definition of observational fairness, independence does not account for explainable or justified discrimination. For example, if men have higher reoffending rates, enforcing equal positive prediction rates across genders in a recidivism system could result in disproportionately longer

sentences for women (Binns, 2018). While it is easy to work with mathematically, independence can lead to ethically problematic decisions that overlook deeper fairness considerations, which is also true for all other statistical criteria (Barocas et al., 2023).

Separation

Separation requires that the predicted outcome $\hat{Y}$ is independent of the sensitive attribute A given the true label Y. This ensures that individuals with the same qualification (same Y) have equal chances of receiving a correct prediction across groups. This is often expressed as error rate parity across groups. Prominent examples in this category include predictive equality, equal opportunity and equalized odds. Equalized odds is the most demanding criterion. It requires that both the false positive rate (FPR) and false negative rate (FNR) are equal across groups. Predictive equality relaxes this condition by requiring only equal FPR across groups. This is advisable in contexts where false accusations or restrictions (e.g. wrongful detention) are the primary concern. A classifier with equal FPR will mathematically also have equal TNR (FPR = 1- TNR). Equal opportunity, on the other hand, focuses only on equal FNR between protected and unprotected groups. This is useful when missed opportunities (e.g. denying a loan or access to education – false negatives) are considered more ethically problematic than false positives. A classifier with equal FNR will mathematically also have equal TPR (Verma & Rubin, 2018). While separation ensures that individuals with the same true label are treated equally across groups, enforcing error rate parity does not come without trade-offs. It can lead to reduced predictive performance, especially if groups differ in data quality, sample size, or base rates. Yet, equalizing error rates can be crucial in contexts where historically disadvantaged groups have faced systematically higher misclassification risk. It makes sure that no group is unfairly harmed by the predictions. These trade-offs can be visualized with ROC curves. The intersection of curves for each group marks the region where error rate parity is possible without losing too much performance (Barocas et al., 2023).

A key limitation of separation comes from its reliance on the target variable Y. When Y itself reflects structural inequalities, such as biased arrest records, it can reinforce injustice rather than correct it. In such cases, the fairness metrics must be questioned, but also the appropriateness of using machine learning at all. From a normative perspective, error rate parity is often hard to justify on its own. Like other statistical fairness measures, it focuses on statistical balance rather than moral reasoning and it overlooks the fact that not all errors are equally severe. The COMPAS case shows how disparities in error rate, in this instance higher false positive rates

for black defendants, can translate into real-world harm with huge consequences on individuals and entire communities. Separation is particularly relevant in perception tasks, like face or language recognition, where the true labels are unambiguous and differences in qualification are not expected. In such cases, disparities in error rates are especially unjustifiable. In general, disparities in error rates not only reflect existing inequalities but can also perpetuate them, by feeding biased outcomes back into decision-making systems. For example, if black defendants are systematically rated as high risk, this can lead to stricter sentencing, which in turn reinforces the data patterns used to justify future predictions. We will return to these dynamics when discussing feedback loops and self-fulfilling predictions in Chapter 2.3.

Separation offers an approach that acknowledges the historical and social conditions that led to observed group differences. Once again, statistical measures alone do not account for context or moral reasoning. However, separation metrics can still be a valuable starting point for investigating structural discrimination in machine learning systems (Barocas et al., 2023).

Sufficiency

Sufficiency requires that the true outcome Y is independent of the sensitive attribute A given the predicted outcome $\hat{Y}$. Once a prediction has been made, the probability that it is correct should be the same for protected and unprotected groups. Predictions should be equally reliable across groups. This is often evaluated using group calibration and predictive parity. Calibration, in general, means that for any predicted probability score (e.g. 0.7), the actual proportion of individuals that belong to the positive class should be approximately equal to that score. For example, among all individuals with a predicted risk of 70%, around 70% should truly belong to the positive class. This ensures that predicted scores are meaningful and interpretable as probabilities. In the context of fairness, we are not only concerned with overall calibration, but more often with group calibration. This requires that individuals from different groups who receive the same predicted score must have the same likelihood of truly belonging to the positive class. The predicted score must have the same meaning for protected and unprotected groups. Predictive parity is a related criterion and demands that the positive predictive value (PPV) is the same across groups. A model that satisfies predictive parity will mathematically also have equal false discovery rates (FDR) across groups (Verma & Rubin, 2018).

Sufficiency ensures predictive consistency. If two individuals from different groups receive the same prediction or score, they should have the same likelihood of a correct outcome. It aligns with the narrow view of equality of opportunity, where decisions are based only on task-relevant

factors and not on protected attributes. Moreover, sufficiency can often be achieved without explicit fairness constraints and can serve as a sanity check for evaluating whether predicted scores behave consistently across groups. However, there are also downsides, as calibration does not guarantee overall accuracy or fairness. A model can perform poorly and still satisfy calibration. Similar to separation, sufficiency also assumes that the target variable is valid and unbiased. If Y reflects historical inequities, sufficiency will not correct for this and may even legitimize unfair predictions (Barocas et al., 2023).

The main challenge with observational fairness definitions is their mathematical incompatibility, leading to context-dependent choices of metrics. As Barocas et al. (2023) show, sufficiency and separation cannot both hold if the sensitive attribute A and the true label Y are statistically dependent, which is often the case. This is because of their conflicting conditional independence assumptions. Sufficiency is conditioned on the predicted outcome ($\hat{Y}$), while separation is conditioned on the actual outcome (Y). Similar incompatibilities exist between sufficiency and independence, and between independence and separation, because enforcing multiple fairness constraints often requires mutually incompatible assumptions about the statistical relationship among A, $\hat{Y}$, and Y. This tension became evident in the COMPAS case, where the model was calibrated (satisfied sufficiency) but still showed disparities in error rates (violated separation). Following work by Kleinberg et al. (2016) and Chouldechova (2017) formally proved that calibration and equalized odds cannot be satisfied simultaneously, except under strict conditions like perfect predictions or independence between A and Y.

As mentioned earlier, while observational fairness definitions provide valuable diagnostic tools to detect statistical disparities, they are fundamentally limited in scope. They only rely on observed relationships between the prediction, outcome and sensitive attributes, and ignore all other features that could also contribute to unfairness. As a result, they can miss the actual driving mechanisms of discrimination. Similarity-based definitions, discussed in the following section, try to solve this issue by including attributes that might seem irrelevant to fairness. Observational metrics can confirm that certain groups are treated unequally, but they cannot explain why. These definitions are blind to context, they do not include how unfairness may have its roots in structural inequalities, decision-making processes, or causal factors outside the model. The 1973 UC Berkeley admission case illustrates this problem. Aggregated data suggested gender bias against women, but a disaggregated view on department-level revealed that there was no systematic discrimination within departments. This example of a Simpson's Paradox – trends for subpopulation must not hold on population level and vice versa – shows

how observational fairness can be misleading and why understanding causality, if possible, is crucial (Barocas et al., 2023; Loftus et al., 2018). Section 2.1.3.3 gives a brief overview of causal fairness definitions.

2.1.3.2 Similarity-based fairness

The observational metrics discussed so far are focused on observable statistical differences at the group level. Similarity-based fairness concentrates on how individual cases are treated in relation to one another. Instead of asking whether outcomes are balanced across groups, these definitions ask if similar individuals receive similar outcomes, regardless of group membership. This reflects the intuitive idea that fairness means treating similar cases consistently and is often used to evaluate whether a decision process is fair based on how it works, not just on its aggregated outcomes (Mehrabi et al., 2021). Many observational definitions ignore relevant attributes other than the sensitive feature, which can potentially hide unfairness. Similarity-based fairness in contrast explicitly takes individual features into account and avoids this limitation (Verma & Rubin, 2018). The following section gives an overview of three similarity-based approaches, causal discrimination, fairness through unawareness and fairness through awareness.

Causal discrimination is a fundamental form of similarity-based fairness. It demands that two individuals who are identical in all attributes except for a sensitive feature (e.g. race or gender) should receive the same outcome. This captures the intuition that sensitive attributes should not causally influence decisions when everything else is equal. In theory this is a straightforward approach, but in practice it is very rare that two individuals differ in only one dimension, because sensitive features are often correlated with other variables. Although the term causal is used, this definition does not use formal causal modeling. Instead, it uses pairwise comparisons based on feature similarity, which is why it is considered part of similarity-based fairness rather than causal fairness (Verma & Rubin, 2018).

Fairness through unawareness enforces fairness by ensuring that a model does not explicitly use sensitive attributes (Kusner et al., 2017). The idea is simple, if the algorithm does not use protected attributes, it cannot discriminate based on them. This principle is often implemented through blinding, which means removing protected attributes from training data. However, research has found out that fairness through unawareness is not sufficient to guarantee fairness. Since non-sensitive features often act as proxies for protected ones, the model may still learn biased patterns. Blinding can even result in miscalibration when predictions are systematically

less accurate for certain groups. For example, Corbett-Davies et al. (2023) demonstrate that race-blind models for diabetes risk tend to underestimate the risk for Asian patients while overestimating it for White patients, leading to suboptimal medical recommendations. In a similar way, gender-blind recidivism models overpredict risk for women, who statistically reoffend less frequently than men. This loss of calibration occurs because the model is forced to ignore meaningful predictive information. This can lead to reduced decision quality, accuracy and fairness (Corbett-Davies et al., 2023). Nonetheless, at times blinding can still be necessary for legal, social, or political reasons, even if it reduces utility. Fairness through unawareness requires a trade-off between statistical accuracy and explicit equality of treatment (Barocas et al., 2023).

In contrast, fairness through awareness requires that the sensitive attribute must be included to ensure fair decisions. Like causal discrimination, it builds on the principles of individual fairness, ensuring that similar individuals receive similar outcomes. Fairness through awareness uses a more elaborate approach than causal discrimination though. It defines similarity using a distance metric over the input features and evaluates fairness based on how consistently the model treats individuals with a small feature distance. This approach is more flexible and nuanced than blinding because the model can account for relevant group-specific information when justified. However, it also introduces a new challenge in determining a suitable distance metric. Different definitions of similarity can lead to different fairness outcomes. Since fairness through awareness requires both a definition of similarity and the inclusion of sensitive features, it focuses on how decisions are made. That is why it is often seen as a way to ensure fairness within the decision-making process itself, rather than only evaluating the final outcomes. It forms the conceptual basis for more advanced causal approaches like counterfactual and path-specific fairness, where fairness is defined through explicitly modeled causal relationships (Dwork et al., 2012; Verma & Rubin, 2018).

While similarity-based fairness wants to ensure consistent treatment based on individual attributes, it relies on assumptions about which features are relevant and how similarity should be defined. Causal fairness methods explicitly model these assumptions by describing how variables influence each other. This allows a deeper understanding of how sensitive attributes influence decisions.

2.1.3.3 Causal fairness

Causal fairness relies on causal models, typically represented as graphs, to illustrate the interaction between variables such as race, gender, education, or income. These models make it possible to distinguish between legitimate and illegitimate influences. For example, it can help to clarify whether the low chances of a person being hired are due to their qualifications or unfairly influenced by their demographic background (Mehrabi et al., 2021). Compared to the previous two fairness approaches, causal models offer a richer and more precise framework. They make it possible to use hypothetical scenarios and allow for more nuanced fairness assessments. However, causal reasoning also brings some challenges. Building a causal model requires expert knowledge and a clear understanding of how different variables interact. This can be particularly difficult for social attributes like race or gender. Unlike biological phenomena such as rain, these categories do not have a stable, clearly defined meaning. Although categories like race and gender may appear stable, they are socially constructed and have changed over time, across cultures and contexts. This makes them hard to model causally, since their meaning is not fixed. Another challenge is the “looping effect”, a form of a self-fulfilling prediction, which will be discussed in Chapter 2.3. This occurs when people change their behavior in response to being assigned to a category, such as being labeled dangerous, talented, or disadvantaged. This categorization can influence reality, which makes it harder to treat such attributes as fixed causes in a model. Consequently, it is difficult to treat gender and race as stable input variables in causal models. The meanings of these categories can change over time, and their influence depends strongly on how society defines and responds to them. This makes causal reasoning about these attributes conceptually and practically challenging (Barocas et al., 2023). Despite these difficulties, causal approaches are promising to address fairness concerns in machine learning. In particular, when fairness does not just depend on treating similar individuals the same, but to understand why differences in treatment arise in the first place. Two important concepts in this area are counterfactual fairness and path-specific fairness.

Counterfactual fairness asks whether a decision would have been different if a person had belonged to another demographic group, while keeping everything else about them the same. If the decision stays the same in this hypothetical scenario, it is considered fair. This approach focuses on individual fairness and uses causal models to evaluate how sensitive attributes directly and indirectly influence the outcome. The concept of counterfactual fairness is powerful, but hard to apply in practice. It requires detailed models of how variables influence

each other, and the construction of realistic counterfactuals is ethically and technically challenging. Still, this method provides deeper insights into how and where unfairness comes from in a system (Kusner et al., 2017; Loftus et al., 2018).

Path-specific fairness builds on the idea of counterfactual reasoning and extends it by enabling the distinction between acceptable and unacceptable paths through which sensitive attributes influence decisions. For example, it might be considered fair in college admission that race affects access to educational opportunities, which then shapes a student's qualification, because the admissions decision is based on actual achievements. On the other hand, it may be seen as unfair if race influences standardized test scores through unequal access to preparation resources, since the scores then reflect social disadvantage and not true ability. Path-specific fairness allows decision-makers to explicitly define which causal paths are considered fair and makes sure that only these paths are active in the model. It offers a practical and context-sensitive balance between fairness and utility. However, like all causal models, it depends on an accurate understanding of the domain (Corbett-Davies et al., 2023; Loftus et al., 2018).

Causal fairness offers a deeper and more structured way to address fairness in machine learning, by focusing on how and why decisions are made. Unlike observational or similarity-based approaches, causal methods enable targeted interventions and the identification of even indirect unjustified influences on fairness. The downside is the high amount of knowledge and modeling effort needed. Nevertheless, causal approaches offer important tools for promoting fairness in complex systems.

Table 2 gives an overview of the fairness metrics discussed in this section. Each of these metrics captures a different aspect of fairness, and there is no single, universally applicable solution. Instead, the choice of a fairness definition must take the specific context, data characteristics, and societal goals of the machine learning application into account. It is often necessary to apply more than one metric to assess fairness from multiple perspectives. Simply choosing a couple of fairness definitions and satisfying them is not what fair machine learning is about. Fairness metrics are typically used as a diagnostic tool to uncover potential sources of unfairness, but they do not provide a complete picture. The broader assessment of fairness in machine learning involves many additional aspects and perspectives that will be discussed in Chapter 3. The next section presents a practical example that shows how different statistical fairness measures contradict each other. This highlights the need to carefully balance between fairness definitions, depending on the context.

Table 2

Fairness metrics discussed in Chapter 2.1.3

Metric	Fairness Category	Short Description
Statistical Parity	Observational	Ensures equal rates of positive outcomes across groups
Equalized Odds	Observational	Requires equal FPR and FNR across groups
Predictive Equality	Observational	Requires equal FPR across groups (or TNR)
Equal Opportunity	Observational	Requires equal FNR across groups (or TPR)
Group Calibration	Observational	Ensures predicted probabilities reflect actual outcomes equally across groups
Predictive Parity	Observational	Requires equal PPV across groups
Causal Discrimination	Similarity-Based	Individuals differing only in a sensitive attribute should receive the same outcome
Fairness through Unawareness	Similarity-Based	Achieved by excluding sensitive attributes from the decision-making process
Fairness through Awareness	Similarity-Based	Similar individuals (according to a distance metric) should receive similar outcomes
Counterfactual Fairness	Causal	An individual's outcome should remain unchanged in a counterfactual world where their sensitive attribute is different
Path-Specific Fairness	Causal	Allows only specific causal paths from sensitive attributes to influence decisions

Note. Adapted from Verma & Rubin (2018).

2.1.4 Practical Study: Fairness notions in predictive model to support students

Objective

This practical study is the third part of the interactive notebook developed for computer science students (full notebook with code in Appendix E). The overall goal of the notebook is to raise awareness about fairness challenges in machine learning and to illustrate why achieving fair outcomes is not straightforward. This study translates the theoretical foundations about fairness and discrimination into a concrete application. Students are given the opportunity to explore how different fairness metrics behave in a practical scenario, how they are computed, and how they can be interpreted. A main lesson from this study is that fairness is not a universal property that can be fulfilled by satisfying a handful of criteria. It depends on the context in which the model is applied and the fairness definitions used to evaluate it. Each metric captures different aspects of potential unfairness between groups. Students should learn that fairness metrics, especially observational metrics, are more diagnostic tools that help to uncover patterns of bias or differential treatment and not direct solutions themselves. The incompatibility of some metrics is also demonstrated in the study, further proving the point that informed, and context-sensitive metric selection is necessary.

The model developed in this study predicts whether a student will graduate or drop out. Such a model could be used in a system that provides targeted support for individual students that are predicted to drop out. While the potential benefit of such an application is clear, it can also raise ethical concerns. If a model incorrectly predicts that a student is likely to graduate, when they are in fact at high risk of dropping out, the system could ignore the student even though assistance would be crucial. The study therefore shows that the severity of false positives and false negatives can differ in certain contexts. By setting the study in this relatable context, students are encouraged to move past abstract definitions and to critically reflect on the real-world implications of model decisions. After evaluating the model's overall performance, differences in performance between genders are analyzed showing how group disparities can persist even when overall accuracy is high. The study is not intended to only show how to calculate fairness metrics, but also to provide deeper understanding of the limitations, trade-offs, and moral questions that come with the use of predictive models in sensitive domains.

Dataset and Preprocessing

The dataset used in this study is the Student Performance Dataset from the UCI Machine Learning Repository (Realinho et al., 2022). It contains information about students enrolled in

higher education institutions in Portugal. The dataset includes demographic data (e.g. gender, age, marital status), academic performance (e.g. grades, failures), and institutional factors (e.g. scholarship status, application preferences). For this study, the dataset was filtered to exclude students with the label ‘Enrolled’, so only the outcomes ‘Graduate’ and ‘Dropout’ remained. The target variable was then binarized, with 1 indicating graduation and 0 indicating dropout. The categorical feature ‘Nationality’ was removed as it was highly imbalanced and offered little informative value for the analysis. The sensitive attribute gender was more balanced and suitable to demonstrate fairness-related disparities. It allowed for more meaningful group comparisons and clearer insights into how fairness metrics respond to uneven model performance. Gender was encoded as 0 for female and 1 for male.

Preprocessing followed a standard machine learning pipeline using scikit-learn. Categorical features were transformed using a ‘OneHotEncoder’. Numerical features were standardized with ‘StandardScaler’. The data was split into training and test sets using stratified sampling to preserve the outcome distribution.

Methodology

A ‘RandomForestClassifier’ was trained within a pipeline and used to predict graduation outcomes on the test set. In addition to standard performance measures such as overall accuracy and confusion matrix, the model was evaluated with a focus on observational fairness metrics. Through a combination of provided results and student input, all eight core statistical measures (Table 1) and all six discussed observational fairness definitions were computed.

In a second step, post-processing mitigation was applied using the Fairlearn ‘ThresholdOptimizer’ (Fairlearn Organization, n.d.). The goal was to show that even when separation-based metrics like equalized odds are satisfied with the help of the ‘ThresholdOptimizer’, other metrics can still indicate unfairness (e.g. sufficiency metrics).

Results and Interpretation

The model demonstrated strong overall predictive performance with 91% accuracy on the test set. It was slightly more accurate for female students (91.9%) than for male students (89.5%). While this seems reasonably balanced, deeper fairness analysis revealed disparities in how the model behaved across groups.

The clearest violation of the six observational fairness definitions appeared in terms of statistical parity. Female students were significantly more likely to receive a positive prediction

(classified as ‘Graduate’) than male students – 75% versus 51%. This means that the model distributed positive outcomes unequally across groups. In the context of this study, a positive prediction leads to the exclusion from student support programs. This imbalance could result in male students receiving more support. While at first glance such skewed predictions could seem unfair, the true fairness implications must be assessed carefully. Looking at the distribution of actual graduation outcomes in the dataset, the number of females graduating is way higher than male graduates. Dropout rates are similar across genders, but the number of female graduates (1661) is more than double the number of male graduates (548). Such unequal base rates have a big influence on fairness metrics. Statistical parity, as an independence-based criterion, compares the rate of positive predictions across groups, without considering the actual outcomes or whether those predictions are justified. It assumes that equal treatment means equal distribution of outcomes, regardless of actual differences in qualification. This can be problematic in settings where base rates differ for valid reasons (Binns, 2018). To satisfy statistical parity in such cases, a model would need to assign positive predictions that do not reflect the actual distribution of the target variable. This would lead to a direct violation of calibration, which demands the distribution of model predictions to reflect the actual target distribution. Independence metrics like statistical parity are often criticized for their lack of sensitivity to context and for being normatively weak, especially when they compromise overall model performance or predictive reliability (Barocas et al., 2023).

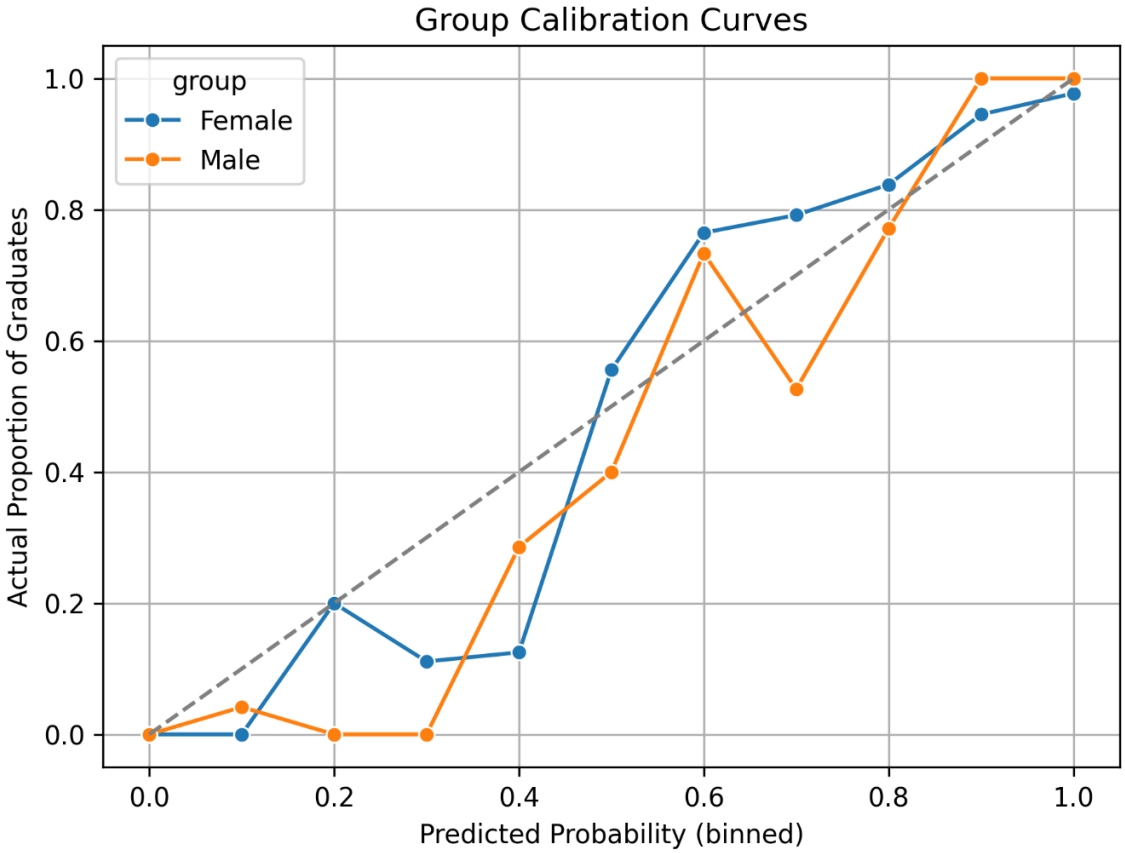
In contrast, equal opportunity was approximately satisfied. The false negative rate for male students was 0.036 and 0.021 for female students. The absolute difference of 0.015 is considered acceptable in many contexts. There exists no universal threshold, but in several toolkits, such as Fairlearn (Fairlearn Organization, n.d.) and AIF360 (LF AI & Data Foundation, n.d.), metric differences below 0.05 are considered negligible and differences below 0.1 are often treated as acceptable. Verma and Rubin (2018) also accepted differences below 0.05 to be minor. Still these thresholds must always be interpreted depending on the context and potential consequences (Barocas et al., 2023). The small difference in false negative rates between genders suggests that the model performs fairly in identifying those who need help.

A bigger gap appeared between the false positive rates. Female students were more likely to be misclassified as graduates (FPR = 0.223) than male students (FPR = 0.159). Predictive equality was not satisfied with a difference of 0.064. This led to a violation of equalized odds, which demands FPR and FNR to be equal across groups.

Predictive parity was also violated. The positive predictive value was higher for female students (0.912) than for male students (0.824), which indicates that the positive predictions from the model were more reliable for women. From the perspective of sufficiency-based fairness this is problematic, because the same predictions should have the same meaning for all individuals. This was further supported by looking at group calibration. A mean absolute calibration difference of 0.101 together with a graphical analysis (Figure 1) showed that the model's predicted scores correspond to different actual graduation rates depending on gender, meaning that group-level calibration was violated.

Figure 1

Group calibration



Note. Based on data from the Student Performance Dataset (Realinho et al., 2022).

To reduce disparities in error rates, Fairlearn's post-processing method 'ThresholdOptimizer' was applied with the constraint of equalized odds. After applying the threshold adjustment, the overall accuracy stayed almost the same with 90.5%. The accuracy for female students increased from 91.9% to 92.5%, while the accuracy for male students dropped from 89.5% to 86.8%. The performance difference between genders increased.

Group calibration was not recomputed after the threshold adjustment. This is because the 'ThresholdOptimizer' does not modify the model's predicted probabilities but only changes the classification threshold for each group. This approach makes calibration assessment unreliable, since the same predicted score can lead to different outcomes depending on group membership. Because group calibration needs a consistent mapping from predicted probabilities to outcomes, it can no longer be meaningfully evaluated under these conditions (Pleiss et al., 2017). Table 3 gives an overview of the observational fairness metrics after the intervention.

Table 3

Results student prediction study

Metric	Before Threshold Adjustment	After Threshold Adjustment
Statistical Parity Difference	0.245	0.201
Equal Opportunity Difference	0.015	0.003
Predictive Equality Difference	0.064	0.005
Equalized Odds Difference	0.079	0.008
Predictive Parity Difference	0.088	0.142
Group Calibration Difference	0.101	-

Note. Based on data from the Student Performance Dataset (Realinho et al., 2022).

Before threshold optimization, equal opportunity was approximately fulfilled, but the model violated all the other fairness criteria. The post-processing intervention improved separation-based fairness metrics. On the other hand, it led to greater violation of sufficiency-based criteria. This outcome illustrates the trade-off, that sufficiency and separation cannot be satisfied simultaneously when base rates differ, unless the model is perfectly accurate or the sensitive attribute unrelated to the target variable (Chouldechova, 2017; Kleinberg et al., 2016). Fairness metrics are not interchangeable, and no model can satisfy all fairness definitions at once. The choice of which fairness goals to prioritize must depend on the context, ethical considerations, and the consequences of the prediction.

Discussion

The primary goal of this practical study was not only to demonstrate how fairness metrics can be computed, but to help students develop a deeper, contextual understanding of what these metrics mean and what they do not. Students should get a feeling for the difficulties in assessing fairness. A practical example is more concrete than theoretical discussions.

The notebook was intentionally designed without using prebuilt fairness evaluation toolkits such as Fairlearn. Instead, all fairness metrics were implemented manually, using basic model

outputs and confusion matrices. This was done to increase transparency, mathematical understanding and interpretability. The direct engagement with the structure and origin of metrics like true positive rate, positive predictive value, or statistical parity, enables students not only to observe the results but also to comprehend their implications more deeply.

The study highlighted the potential as well as the complexity of fairness evaluations. The model performed well overall, but it violated several fairness metrics to varying degrees. The results showed that fairness cannot be captured by a single definition and not all fairness metrics can be satisfied simultaneously, especially when base rates differ. Another takeaway for the students is that fairness metrics are different in what they want to achieve. For example, statistical parity only compares prediction rates across groups, without considering whether predictions are justified or accurate. As already discussed in the previous section, with an unequally distributed target variable, enforcing statistical parity can be misleading. This illustrates that fairness definitions are dependent on the context and the importance of looking at the consequences of prediction errors. In the case study, the goal of the model is to identify students at risk of dropping out. As resources to support students are limited, targeted aid is needed. In this scenario, a false negative (missing a student at risk) is more severe than a false positive (offering support unnecessarily). Because of that, equal opportunity (equal FNR) can be seen as the most relevant criterion in this study.

From a technical perspective it could be argued that no threshold adjustment was necessary, because the model already showed very low and balanced false negative rates between groups before any post-processing. The ‘ThresholdOptimizer’ was applied more as an educational tool, to demonstrate exemplarily that such interventions exist and show students the unintended side effects that can occur. Although the intervention improved separation-based metrics, it also led to increased disparities in calibration and predictive parity. In addition to the changes in fairness metrics, the threshold adjustment also resulted in a slight redistribution of predictive performance. The overall accuracy did not change much, but the accuracy for female students increased, while the accuracy for male students decreased. This demonstrated that fairness interventions are not neutral, improvements in one area often come at the cost of another. The study should help students to understand that fairness cannot be fixed with technical tools alone but always requires normative and context-specific reflection.

After the introduction of different definitions of fairness in Part 2 of the notebook, this practical study encourages students to acknowledge the difficulties of measuring fairness in practice. The

study was designed not only to demonstrate how to apply fairness metrics, but to enable the students to question the metrics and realize that context and goals ultimately decide which metrics are relevant.

2.2 Bias in machine learning

As mentioned before, bias plays a central role in discussions around fairness in machine learning. Fairness is the goal in mind when evaluating algorithmic systems. It includes aspects like transparency, explainability and accountability, but bias is the factor that contributes the most to discrimination and thus unfairness. Before Section 2.2.1 introduces the most relevant types of bias in the context of machine learning, it is important to understand what bias means and why it matters.

Bias is a broad term that is used across many disciplines. The definition of bias is not always consistent. In this thesis we refer to biases that act as sources of discrimination in machine learning systems (Suresh & Guttag, 2021). Bias can take many forms. It can be expressed negatively, such as unjust discrimination, or positively in the form of favoritism. Both forms can be unfair. Bias can show up everywhere from everyday judgements, like trusting someone because of an outfit, to the design of large decision-making systems in areas such as loans, hiring or healthcare. That biases influence decisions explicitly and consciously is rarely the case. Often, we find implicit biases in machine learning applications. Implicit biases are unconscious, automatic attitudes or stereotypes that all people have in some form. They are often shaped through cultural norms, media or personal experiences. These biases can influence decisions even when we believe we are acting fairly and objectively. What makes implicit bias problematic is that it enters machine learning systems unintentionally. Developers or data scientists do not encode bias on purpose. Instead, it sneaks in through training data that reflects social inequalities, subjective labels, or design decisions that did not anticipate the full downstream effects (Howard & Borenstein, 2018).

For humans, cognitive biases can be very useful. They allow us to make quick decisions in complex situations with high uncertainty (Gendler, 2011; Kahneman, 2012). Using these heuristics in seemingly neutral and data-driven machine learning applications can be problematic. Machine learning bias is not just a faster way to come to a decision, it can introduce and reinforce unfairness that is hard to detect and even harder to fix (Alelyani, 2021).

Bias can arise at any stage of the machine learning lifecycle and addressing bias requires deep understanding of its sources. Listing every possible form of bias in machine learning would not be possible. Instead, the following section lists the most common biases and their location in the machine learning lifecycle. Biases can also interact and reinforce each other in feedback loops, where biased system outputs influence future data, user behavior and system decisions

(Mehrabi et al., 2021). The phenomenon of self-fulfilling predictions and feedback loops will be discussed in Chapter 2.3.

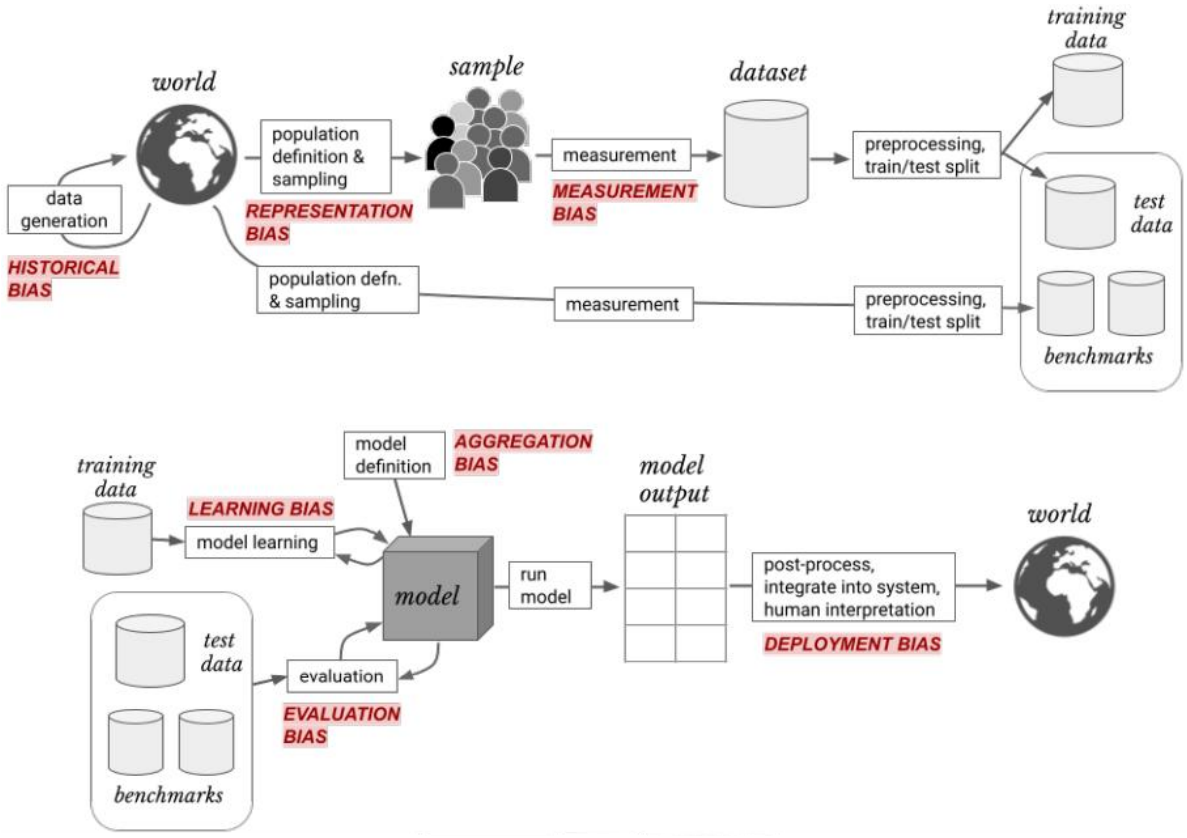
Biases are deeply embedded in our cognitive and societal structures. In general, not all biases are harmful, but those that influence decisions about people's lives need to be assessed. While it may not be possible to eliminate all bias, we can learn to identify it, understand where it comes from, and then try to reduce negative impacts (Alelyani, 2021). Increasing awareness and understanding about biases will help to build fairer machine learning systems.

2.2.1 Forms of bias

The framework by Suresh and Guttag (2021) identifies seven major categories and their location in the machine learning lifecycle. This section follows their framework and adds examples as well as relevant subtypes of these bias categories from other sources.

Figure 2

The machine learning lifecycle and different categories of bias



Note. From Suresh & Guttag, 2021.

Historical bias

Historical bias occurs when the current or past state of the world leads to unjust model outcomes, even if all the steps from data preparation to modeling are bias free. This bias often results in representational harm for certain groups, as structural inequalities are mirrored in the data. Examples of this form of bias are word embeddings that reflect stereotypes, like associating nurse with female and engineer with male (Suresh & Guttag, 2021). Various forms of bias can occur before actual sampling of the data. Oftentimes these are biases that come from the population into the data and are not directly connected to machine learning itself. In Figure 2 this is depicted by the data generation loop. Table 4 below lists some of the most relevant forms of bias for machine learning systems that come from the user and can be seen as specific manifestations of historical bias as defined by Suresh and Guttag (2021).

Table 4

Historical bias subtypes

Bias Type	Explanation
Temporal Bias	Data reflects outdated states of the world, failing to capture changes over time
Content Production Bias	Certain groups produce more or different data (e.g. online content), skewing representation
Behavioral Bias	Different behavior across contexts or plattforms
Social Bias	Actions from others affect own judgement (e.g. restaurant reviews)
Self-Selection Bias	Participation in data generation is non-random
User-Interaction Bias	Arises through feedback loops created by users interacting with the system

Note. Adapted from Mehrabi et al. (2021).

Representation bias

Representation bias arises when the dataset fails to represent all relevant parts of the population. This leads to poor generalization for underrepresented groups. Representation bias can occur because the defined target population might not reflect the true user population, for example because of outdated census data. This form is often called population bias. Representation bias can also happen because the target population includes underrepresented subgroups, like pregnant women within a group of adults. It can also arise because the sampling methods do not capture the full variation of the population. For example, medical data is often only available from patients. Literature often calls this sampling bias. An example for representational bias is the image dataset ImageNet, where 45% of images come from the U.S., and only 1% from

China. Models trained on ImageNet were shown to perform worse at image classification for groups from underrepresented regions (Suresh & Guttag, 2021). In Chapter 2.2.3 we will discuss such differences in accuracy with gender classification models.

Measurement bias

Measurement bias concerns how we choose, utilize and measure features and labels. It often arises when proxies are used to represent complex or unobservable concepts but fail to represent them accurately across groups. Measurement bias can occur when proxies oversimplify the targeted construct, the method of measurement differs across groups, or the measurement accuracy is different. The COMPAS risk assessment tool used arrests as a proxy for crime. Since minority communities are more policed, this led to biased outcomes, such as higher false positive rates for black defendants (Suresh & Guttag, 2021). Using labels that do not reflect reality accurately, is often called label bias (Corbett-Davies et al., 2023). Omitted variable bias and label bias are frequently mentioned in literature about machine learning bias and can be seen as subtypes of measurement bias. Omitted variable bias occurs when important variables with explanatory power are missing in the model (Mehrabi et al., 2021).

Aggregation bias

Aggregation bias occurs when a single model is applied to data that contains distinct subgroups with different characteristics. It assumes the same relationship between input and predictions across all groups, which often is not true. As a result, the model underperforms for all groups or favors the dominant one (Suresh & Guttag, 2021). For example, in diabetes prediction, risk levels differ across genders and ethnicities. When all groups are treated the same, this leads to misclassifications. The Simpson's Paradox is a form of aggregation bias, where trends in aggregated data reverse or disappear once the data is split into subgroups (Mehrabi et al., 2021).

Learning bias

Learning bias comes from modeling choices that lead to unequal performance across different groups. This is often caused by optimizing for a single objective, like overall accuracy, which can harm fairness-related goals. For example, training for model simplicity can reduce the model's ability to learn from underrepresented data, resulting in worse outcomes for those groups (Suresh & Guttag, 2021).

Evaluation bias

Evaluation biases arises when benchmark datasets do not reflect the population the model will be used on. This can influence model development because benchmarks are used for optimization that are not as diverse as the real world. Relying on standard metrics like overall accuracy can also hide poor performance on specific subgroups. The study by Buolamwini and Gebru (2018) showed that gender classification performed worse for dark-skinned women. This was not detected or mitigated because this group was underrepresented in the benchmark datasets (Suresh & Guttag, 2021).

Deployment bias

Deployment bias occurs when a model is used in ways that differ from the original design or evaluation context. This often happens when models are treated as standalone solutions, without paying attention to the real-world systems and human decisions that surround them. Even accurate models can cause harm when they are used in ways that were not intended. For example, risk assessment tools that were designed to estimate reoffending, are instead used to justify longer prison sentences (Suresh & Guttag, 2021).

2.2.2 Bias amplification

Existing biases in training data cannot only be replicated using machine learning models but can also be intensified. It occurs when the model makes biased predictions at a higher rate than what would be expected from the data alone, often leading to reinforced stereotypes or unjust performance differences between groups. A study by Hall et al. (2022) presents some first insights into how different factors, such as training set size or model capacity, influence bias amplification. The following section gives a short overview of their most important findings. The results were found in binary classification and image recognition. Therefore, the authors point out that more research about bias amplification is needed to see how generalizable the findings are.

First, bias amplification tends to increase with the level of bias in the training data. It reaches its peak before the data becomes fully biased and there is no bias amplification when the data is either completely unbiased or fully biased. The extent of amplification varies between datasets depending on how easy it is for the model to recognize group membership. Some datasets show bias dampening instead, when group signals are weak and noisy (Hall et al., 2022).

Another key element is model capacity, which describes the model's ability to learn complex relationships between inputs and outputs. It depends on factors such as hyperparameters, regularization, and model architecture. Models with low capacity tend to underfit the data. They cannot capture relevant patterns and instead rely on simple, easily detectable features – such as group membership - which can lead to bias amplification. In contrast, high-capacity models can learn complex patterns, but they risk overfitting by using patterns that do not generalize well, for instance irrelevant or biased features. Overfitting increases the generalization error on unseen data and can further amplify bias. The relationship between model capacity and bias amplification follows a V-shaped curve. Amplification is the highest on both ends with low and high capacity. A mid-range spot allows the model to generalize well with less bias amplification. Regularization techniques like weight decay limit overfitting and can reduce amplification, but they require a trade-off with accuracy during hyperparameter tuning (Hall et al., 2022).

The size of the training dataset also plays a role. With a larger training dataset more accurate modeling is possible and therefore bias amplification in general decreases with more training data. Very small datasets on the other hand can also lead to low bias amplification, just because models overfit to random noise rather than learning generalizable patterns. A model cannot amplify what it cannot capture (Hall et al., 2022).

Model calibration and confidence are also closely related to bias amplification. Calibration in the sense how well the model's confidence aligns with actual prediction accuracy and not group calibration which was discussed in Section 2.1.3.1. High-capacity models often become overconfident, especially on biased or underrepresented groups, which can lead to more bias amplification. Poor calibration means the model appears more certain than it should be, the accuracy of predictions is lower than its confidence. This masks its misclassifications and reinforces bias. Amplification and overconfidence do not always increase in parallel. In some low-capacity settings bias amplification can decrease even when overconfidence rises (Hall et al., 2022).

Bias amplification also varies in the different phases of training. In early stages, models tend to rely on easily detectable group signals, leading to high amplification. As training progresses, they start to learn class-specific features better, which can reduce bias amplification. However, near the end of training, when the learning rate is reduced, amplification may increase again a bit. Furthermore, if group membership is easier to detect than class labels, amplification is more likely. When the opposite is true, bias dampening can occur instead (Hall et al., 2022).

The mitigate bias amplification, the study recommends careful cross-validation, particularly hyperparameter tuning (e.g. model depth, width, learning rate, regularization). Access to sensitive attributes is necessary during development to monitor and control for bias amplification effects (Hall et al., 2022).

2.2.3 Practical Study: Bias in gender classification models

Objective

This study investigates representational bias in gender classification models (full notebook with code in Appendix G). To increase understanding of the potential problems, the notebook introduces an example of an automated HR tool that analyzes application photos to infer the gender of applicants. Such a tool could be used for diversity tracking or gender-specific hiring programs. At first sight, this application may seem harmless, as it does not make hiring decisions directly. But if the underlying model systematically misclassifies certain individuals, for example women of color, this can result in representational harms. Some people might be excluded from relevant programs and denied job opportunities as a result.

Unlike decision (or distributive) fairness, which was explored in the first practical study and evaluates whether decisions are equitably distributed across groups, representational fairness is concerned with how individuals are seen by a model. In tasks like gender classification, where gender is the prediction target and not a feature, biased outcomes do not directly deny access to opportunities. Instead, by failing to identify underrepresented groups accurately, they can indirectly lead to unfair results, such as denied job opportunities (Binns, 2018).

The goal of this study is to illustrate how bias can influence facial analysis systems and demonstrate why problems with representational fairness might need different handling than decision problems. On top of that, the study highlights how the choice of dataset for evaluation influences observed model performance. This builds on the theoretical framework presented in Section 2.2 and Part 4 of the notebook, particularly the concepts of representation bias and evaluation bias. The study analyzes two different gender classification models across three datasets. By shifting the focus from decision outcomes to recognition performance, the study shows that algorithmic bias can cause harm even when no explicit choice is made. A model that performs well on average can still fail for specific groups in discriminatory ways. This highlights the importance of intersectional analysis and the need to assess fairness not only in what models decide, but also how people are represented.

Dataset and Preprocessing

To evaluate representational fairness in gender classification, this study analyzes model performance across three facial image datasets, the FairFace dataset, the UTKFace dataset, and a manually created subset of UTKFace referred to as UTKFace adjusted. The datasets differ in

terms of demographic balance and label consistency, so this study can reveal data related as well as structural sources of bias.

The FairFace dataset was developed with fairness considerations in mind. It provides a balanced distribution of faces across gender and ethnicity groups and provides metadata labels for race, gender, and age (Karkkainen & Joo, 2021). For this study, a stratified sample of 100 images per ethnicity was selected, ensuring equal group representation. The full sample included 700 images. The number of images per group can be adjusted. A preset value of 100 is a good compromise between obtaining meaningful results and keeping runtime manageable. In the educational setting of the notebook, students cannot be expected to wait too long for the results. The gender labels were adapted to match the model's expected input format (e.g. 'Male' was changed to 'Man'), and full file paths were created to process the images automatically.

The UTKFace dataset also contains a wide variety of images with differences in pose, lighting, expression and image quality. In contrast to FairFace, UTKFace is more imbalanced regarding ethnicities. The metadata is encoded in the filename and includes information on age, gender, and ethnicity (UTKFace Dataset, n.d.). From this dataset, a stratified sample of 200 images per ethnicity was drawn, resulting in 800 images in total. The original numeric labels in UTKFace were translated into readable gender and ethnicity labels to match those used in FairFace.

With the provided metadata on ethnicity and gender, both datasets are useful for fairness evaluations. Nevertheless, they also suffer from potential sources of representation and measurement bias. For example, the UTKFace dataset includes mislabeled images, and neither dataset offers consistent image conditions across all subgroups. This increases the risk that observed differences in performance are partly due to data quality rather than model behavior.

To address this issue and have a dataset that fits the HR screening scenario, a third dataset was manually created. UTKFace adjusted only includes frontal, well-lit images with relatively neutral expressions, drawn from the original UTKFace dataset. The selected images are balanced across gender, ethnicity, and age groups. UTKFace adjusted only includes people between 16 and 70 years of age, reflecting a working-age population. Labels were manually checked, and only plausible image-label combinations were kept. This controlled subset allowed a more accurate assessment of whether observed disparities were caused by poor data conditions or model limitations. By comparing results across these three datasets, the study not only evaluates fairness, but also shows how representation bias and evaluation bias can interact across different datasets.

Methodology

This study evaluates the representational fairness of gender classification models by applying two pre-trained models to three facial image datasets. The goal is not to optimize model performance, but to identify disparities in how well different demographic groups are recognized. The two used models used are DeepFace and OpenCV GenderNet (Taigman et al., 2014; Levi & Hassner, 2015).

Unlike fairness assessments that are focused on decisions, this study deals with representation, a different fairness concern. Gender is not an input feature but the prediction target. The model does not allocate resources or risks, it classifies who someone is. As Binns (2018) emphasizes, representational harms differ from distributive harms and require a different procedure. Fairness metrics such as demographic parity, predictive parity or calibration are designed for decision-making systems with a neutral target variable. They want to balance outcomes or opportunities between groups. In tasks like gender classification, with the sensitive attribute being the prediction target, these metrics are not appropriate. Representational harms occur when certain subgroups are systematically misclassified or excluded from recognition completely, even if the overall model performance appears to be acceptable.

To evaluate such harms, we use group-based performance metrics. Accuracy per group and subgroup, false positive rate, and false negative rate help to identify whether the model performs equally well across demographic groups. Separation-based metrics are appropriate to reveal differences in recognition performance across demographic groups, which indicates representational unfairness. The intersectional view is particularly important, as disparities can remain hidden when results are only reported in aggregated form (see aggregation bias) (Suresh & Guttag, 2021). Instead of only looking at broad groups like women or people of color, an intersectional perspective helps to uncover where performance differences really occur, for example, for Black women (Buolamwini & Gebru, 2018).

The analysis consists of two stages. First, DeepFace is evaluated on the FairFace and UTKFace datasets. The results give insights into model behavior under the two different datasets. To account for data quality and noise, the model is then evaluated with the third dataset. UTKFace adjusted was manually created to simulate conditions that better reflect the HR use case.

In the second stage, the same adjusted dataset is then used to evaluate OpenCV GenderNet. This enables model comparison. We can check if the observed biases are specific to one model or if they could reflect broader structural issues across gender classification systems. The

OpenCV model can be considered simpler than DeepFace because it uses a basic classification pipeline without additional preprocessing steps like face alignment. It was trained specifically for gender classification on a limited dataset (Adience), using a small convolutional network (Levi & Hassner, 2015). In contrast, DeepFace was designed for high-accuracy face recognition and includes multiple processing steps to standardize and analyze faces (Taigman et al., 2014). While DeepFace is more complex in design, both models show fairness limitations.

This two-model, three-dataset approach gives the students detailed insights into representational fairness. It shows that systems that do not directly influence decision-making can still be discriminatory or lead to representational harms.

Results and Interpretation

The assessment of representational fairness in gender classification revealed significant disparities in model performance across demographic groups. Especially the intersectional analysis of gender and ethnicity gave valuable insights. Even under more controlled conditions with the adjusted dataset the differences persisted, which is a sign that bias is not only data-driven but part of the models themselves. A detailed summary of all evaluation results across datasets and demographic groups is provided in Table 5. The results are rounded to two decimal places for better readability.

The DeepFace model showed strong performance for some groups but performed poorly for others. Overall accuracy was higher on UTKFace (80.2%) than on FairFace (68.9%). The difference in accuracy between demographic subgroups was significant with both datasets. The model performed much better for men than for women. Within the female group, women of color had particularly low recognition performance. On FairFace, the accuracy for men was 96.7%, while for women it was only 37.9%. Similarly, on UTKFace, men reached 93.7%, while women were correctly classified 64.7% of the time. A closer look at the intersectional subgroups shows even more severe gaps. In the FairFace dataset, White men were classified with 100% accuracy, while Black women reached only 16.3%, followed by Southeast Asian (21.2%) and Indian women (31.3%). False negative rates (women misclassified as men) were alarmingly high in these groups, with over 80% for Black women. Similar, but not as extreme disparities were found in the UTKFace dataset. Even though the overall accuracy for women was higher than in FairFace (FairFace: 37.9% vs. UTKFace: 64.7%), the group-level patterns remained the same. This suggests that even when models are applied to datasets that are less balanced and reflect more natural variation, like UTKFace, systematic misrecognition persists.

To evaluate whether these disparities are mainly due to poor data quality or labeling noise, the DeepFace model was also applied to the manually created subset, UTKFace adjusted. This subset more accurately reflects the images an automated HR tool would be working with. Some group-level results improved, for example, White women reached 89.1% accuracy. The disparities between groups nonetheless persisted. Black women, despite better image conditions and verified labels, were recognized with only 36.4% accuracy. This is even lower than in the original, noisier UTKFace dataset (51%). The result highlights that cleaner evaluation data does not automatically lead to fairer outcomes. It suggests that the model itself, shaped by its training data, disadvantages certain groups, regardless of the image quality on the test set.

To test whether the observed bias patterns are specific to DeepFace or generally found in gender classification models, the second model OpenCV GenderNet, was evaluated on the adjusted dataset. Although this model is simpler it achieved higher overall accuracy (85%) than DeepFace and also showed more balanced performance across demographic groups. Accuracy for Black women increased to 59.1%, and for Indian women to 80%. The overall gender gap was smaller, and false negative rates were lower across all ethnicities, especially for women. However, the model still performed best for White individuals. Black women remained the group with the lowest performance. These results show that model architecture and training data influence performance, but bias patterns persist across models. Similar findings were observed in the Gender Shades study by Buolamwini and Gebru (2018). They analyzed commercial classifiers from IBM, Microsoft, and Face++ and they all performed poorly for women of color, despite different development and training data. This points to a deeper, systemic issue in how such models are trained and evaluated in general.

A consistent pattern was found across all models and datasets. Gender classification systems work well for dominant groups, in particular White men. They do not work well for women of color. These failures are examples of representational unfairness. The models do not explicitly make discriminatory decisions, they fail to recognize certain individuals disproportionately. This can lead to real-world harm. In the context of automated HR systems, misclassification can result in people being excluded from gender-specific support programs or addressed incorrectly. All of this occurs without any human actively making an unfair decision. An intersectional evaluation was essential to uncover these issues. Looking only at broad categories like women or people of color would have covered some disparities (e.g. Black women). This shows the importance of fairness assessment on a subgroup-level and why average accuracy is not enough. In summary, the results demonstrate that representational fairness cannot be

ensured through high overall accuracy. Bias comes from the way models are developed, trained and evaluated. Addressing bias requires not only technical interventions but also understanding of which groups are overlooked by the system.

Table 5

Accuracy scores gender classification study

Accuracy Scores	FairFace	UTKFace	UTKFace adj.	UTKFace adj. (OpenCV)
Overall	0.69	0.80	0.78	0.85
Men (M)	0.97	0.94	0.99	0.95
Women (W)	0.38	0.65	0.57	0.76
Black (M/W)	0.56 (0.94/0.16)	0.76 (0.98/0.5)	0.68 (1/0.36)	0.78 (0.96/0.59)
White (M/W)	0.75 (1/0.47)	0.86 (0.93/0.76)	0.94 (0.99/0.89)	0.93 (0.94/0.93)
Indian (M/W)	0.65 (0.96/0.31)	0.83 (0.94/0.7)	0.72 (1/0.45)	0.88 (0.95/0.8)
Asian (M/W)	-	0.77 (0.9/0.64)	0.77 (0.96/0.59)	0.81 (0.93/0.7)
E. Asian (M/W)	0.75 (0.92/0.59)	-	-	-
S.E. Asian (M/W)	0.59 (1/0.21)	-	-	-
M. Eastern (M/W)	0.81 (1/0.41)	-	-	-
Latino (M/W)	0.72 (0.94/0.51)	-	-	-

Note. Based on data from the FairFace dataset (Karkkainen & Joo, 2021) & UTKFace dataset (UTKFace Dataset, n.d.). Rounded to 2 decimal places.

Discussion

This study was designed not only as a fairness analysis, but also as a didactic tool to help students to critically deal with ethical challenges in machine learning. Unlike the previous study on student success prediction, which focused on decision fairness, this study illustrates the problem with representational fairness. The main concern is not distributive equality of outcomes or opportunities, it is about visibility and recognition. With the group-based evaluation of performance across gender and ethnicity, students learn that fairness is not only about decisions or allocations, but also about how accurately and equally people are represented in algorithmic systems. The results show that gender classification models perform significantly worse for marginalized groups and that these gaps persist across datasets and models.

Beyond representation bias, the study also wanted to raise awareness to evaluation bias. The comparison of results across three different datasets shows how different fairness issues can appear depending on the test data used. In all three cases, representational harms were present, but the severity differed. This demonstrates how fairness assessments can depend on the chosen evaluation context. If an automated HR system was tested only on a narrow subset, for example mostly White men, the underlying biases of the model might not be noticed, and the model could be seen as highly accurate. Students are encouraged to reflect on whom the model fails, but also on how the evaluation process itself can shape the fairness assessment.

The third form of bias from Section 2.2.1 that gets addressed is aggregation bias. The study highlights how overall accuracy can be misleading, and how disparities only become visible when data is disaggregated by subgroups. While gender classification accuracy can seem acceptable overall, intersectional analysis reveals that certain groups are misrecognized consistently, even when the model performs well for other groups. This shows the importance of intersectional inspection when evaluating fairness, and the danger of relying only on global performance metrics.

In addition to these three types of bias, the study also reflects elements of historical bias and measurement bias. Historical bias is present in the form of structural inequalities. For example, the underrepresentation of women of color in online image datasets and the reliance on binary gender categories (Scheuerman et al., 2020; Suresh & Guttag, 2021). Measurement bias arises through inconsistent labeling and differences in image quality, which can systematically disadvantage certain groups. All in all, the gender classification study represents nearly all relevant forms of bias discussed in Section 2.2.1. This was one of the main reasons for using it in the notebook. It offers a scenario that helps students recognize how different types of bias can occur and interact with each other.

The original idea of this study was to evaluate face recognition models and analyze group differences in performance. However, these models turned out to be less suitable for the pedagogical goals of the notebook. When tested on the FairFace dataset the FaceRecognition model surprisingly showed the lowest accuracy for White individuals (67%) and higher results for Southeast Asian (85%), Indian (79%), and Middle Eastern (76%) faces. One possible explanation is that the images of White people in FairFace are particularly heterogeneous in quality, which makes detecting more difficult. This is only a hypothesis and would need to be systematically tested for reliable conclusions. Nonetheless, the fact that White people had the

lowest accuracy only when using FairFace and across multiple models (dlib and DeepFace had similar results) suggests that the dataset plays a role.

When FaceRecognition was tested on UTKFace and the adjusted version, performance increased, and group differences almost disappeared. All ethnicities in UTKFace had detection rates greater than 95%. In the end none of the face recognition analysis were suitable for the educational purpose of this notebook. The recognition models did not provide clear or consistent examples of unfairness, as disparities between ethnicities were either too small or in the case of the FairFace dataset contradicted expected patterns. Representational harms are typically experienced by marginalized or minority groups, the lowest accuracy for White individuals does not work as an example. On top of that, gender did not show big gaps in performance. Women were sometimes even recognized better than men. To present meaningful and impressive disparities that illustrate the problem of representational fairness to the students would have required manipulation of the data.

In contrast, gender classification provided a clear example of representational unfairness. Without any manipulation the models showed significant differences in accuracy between gender and ethnicity subgroups. These differences were robust across datasets and models and therefore made it possible to illustrate the ethical issues in a direct way. The decision to focus on gender classification was made to provide students with a clear, data-driven example of how representation bias can manifest.

2.3 Self-Fulfilling Predictions and Feedback Loops

Machine learning systems do not operate in isolation. They often influence and are influenced by the environment they try to model. A central mechanism is the feedback loop. It describes the interaction between the model outputs, user behavior, data collection and algorithm updates, each being able to initiate or reinforce feedback cycles. Feedback loops arise when predictions or classifications made by an algorithm influence the behavior of users, decision-makers, or institutions, which then shape the new data that is fed back into the system (Barocas et al., 2023). This cyclical relationship can lead to the amplification of existing biases, especially when the training data itself is already shaped by historical inequalities. Feedback loops can come from underrepresentation (e.g. lack of images of darker-skinned females in facial recognition datasets) as well as overrepresentation (e.g. crime datasets that disproportionately reflect policing of certain communities) (Ntoutsi et al., 2020). Such loops can emerge across all stages of the machine learning lifecycle. Figure 2 illustrates how different types of bias are connected and can influence each other (Mehrabi et al., 2021). The danger of these dynamics is that biased historical decisions can shape future data, which leads to increasingly biased models. In the beginning, it is simply an imperfect representation of reality, but it can evolve into a self-reinforcing system that amplifies biases (Barocas & Selbst, 2016).

Self-fulfilling predictions and feedback loops are closely related phenomena in machine learning. While there are conceptual differences, they often occur together in practice, reinforcing and amplifying each other's effects. Self-fulfilling predictions occur when a prediction alters behavior, resource allocation, or decision-making in a way that the predicted outcome becomes reality. Not because the model was accurate, but because the prediction influenced the result. Feedback loops describe broader cyclical processes in which model outputs or user behavior shape new data that is then used to update or retrain the system, further influencing future outputs. In many cases, a self-fulfilling prediction can initiate a feedback loop, but they can also occur when there is no model retraining or technical feedback. For example, if a university predicts that a particular student is likely to drop out and withholds any financial aid or mentoring support, the student may really drop out, confirming the prediction (Barocas et al., 2023). Unlike feedback loops, which typically involve multiple iterations over time, self-fulfilling predictions can unfold in a single causal chain, still leading to lasting consequences. The concept of self-fulfilling predictions is not exclusively found in machine learning systems. Similar dynamics have been discovered, for example in psychology.

Stereotype threat describes the phenomenon that individuals perform worse when they are reminded of negative stereotypes about their identity group (Howard & Borenstein, 2018).

From an ethical perspective, self-fulfilling predictions in machine learning are relevant because they can influence social reality while appearing neutral. Instead of only describing the world, such predictions can shape it, for example by affecting how resources are allocated or how people are treated. This blurs the line between observation and intervention, especially when decisions based on these predictions change the conditions in ways that make the predictions come true. In these cases, algorithmic systems can create the very evidence that seems to confirm their accuracy. Self-fulfilling predictions represent a key challenge in machine learning. Predictive systems appear to be objective and data-driven, but in practice, they can end up reinforcing existing inequalities. These effects are often difficult to detect, since the mechanisms behind them are hidden or indirect. It may seem that these systems are only predicting future events, while they are actually shaping the outcomes they claim to foresee (Barocas et al., 2023; Osoba & Welser, 2017).

The next section will illustrate the interaction between self-fulfilling predictions and feedback loops in predictive policing systems.

2.3.1 Predictive Policing – Interplay of self-fulfilling prediction and feedback loop

Predictive Policing offers a real-world example of how self-fulfilling predictions and feedback loops can interact, reinforcing existing social disparities through algorithmic decision-making. These systems try to forecast where crimes are likely to occur or who might be involved. Typically, these predictions are based on historical crime data collected by law enforcement. However, this data does not only reflect criminal activity, but also patterns of police presence, enforcement strategies, and community reporting behavior. As a result, predictions made by these systems are not neural reflections of reality. They are shaped by institutional choices that can also inherit biases (Lum & Isaac, 2016).

Place-based systems such as PredPol can illustrate this issue. By analyzing past crime locations, types, and times, these models flag high-risk areas for increased police patrols. When more officers are deployed based on these predictions, more crime will be discovered in those areas. Not necessarily because more crime occurs, but because more is detected. These additional incidents are then recorded and used to train the system again. This results in the reinforcement of the original prediction and a feedback loop. Without adjustment for the sampling bias, the algorithm learns a distorted version of reality. Even small initial differences in crime rates

between neighborhoods can be amplified drastically through this process, leading to over-policing of already marginalized communities. This dynamic has also been analytically demonstrated by Ensign et al. (2017). They modeled feedback effects in predictive policing using a Pólya urn framework and proposed mathematical correction methods. One of them is down-weighting the discovered incidents. Their analysis distinguishes between discovered incidents, which are influenced by targeted policing, and reported incidents, which come from citizens and are less affected by police deployment. They show that feedback loops are mostly driven by biased data collection and not actual crime rates (Ensign et al., 2017).

The whole cycle starts with a self-fulfilling prediction. The forecast of higher crime risk leads to increased patrols, which create the very conditions that validate the forecast. The result is a system that appears to confirm its own accuracy. In reality, it reproduces and amplifies existing inequalities, which in this case affects Black and low-income neighborhoods that have historically been more policed (Lum & Isaac, 2016).

Person-based predictive tools raise similar concerns. These are systems that assign risk scores to individuals based on social network analysis, past interactions with the police, or commercial data such as the Strategic Subjects List used in Chicago. The scores can influence police decisions such as surveillance, stops or arrests. A person with high-risk may be more likely to get investigated, which increases the chance of detecting minor offenses and reinforcing the data that labels them risky in the first place (Robinson & Koepke, 2016).

From an ethical point of view, predictive policing systems are particularly worrying. Errors or biases in predictive policing can have serious consequences for individuals and communities. Applications that do not work correctly, such as voice assistants or recommendation systems, are a problem, but the stakes for mistakes in predictive policing are higher. Errors can lead to unjustified surveillance, wrongful arrests, and even incarceration (Howard & Borenstein, 2018). Over-policing already marginalized groups can lead to psychological and social consequences, such as increased stress, stigma, mental health issues, and a higher risk of police violence (Lum & Isaac, 2016).

One of the main concerns for most of the machine learning applications is the lack of transparency and accountability. This also makes predictive policing systems more problematic. Most of the time vendors claim proprietary rights over the algorithms and even police departments may not fully understand how the systems function or how risk levels are calculated. Informed public debate about their deployment is rare. Predictive Policing tools are

often used without community input or oversight. In Fresno, California, for example, the Beware system was piloted. When asked about how the threat levels are determined, police officers were unable to answer. Individuals had no chance of knowing why they were labeled as dangerous and because of that they also could not challenge it. This naturally leads to concerns about arbitrariness, potential harmful misclassifications and there cannot be trust in these systems.

In summary, predictive policing systems do not just predict crime, they help produce the evidence to justify their own predictions. The interplay of self-fulfilling predictions and feedback loops creates a powerful illusion of accuracy and objectivity, while often strengthening structural inequalities. In the US, these systems were often deployed without sufficient transparency, evaluation, or oversight and their actual impact on community safety is not clear. Existing studies offer little evidence that predictive policing improves public safety. Independent evaluations have found no clear reduction in crime rates. While some studies report moderate improvements in prediction accuracy, critics argue that these gains often do not translate into meaningful reduction of crime. Also, similar improvements could have been achieved with traditional methods (Robinson & Koepke, 2016). This raises concerns about the justification for deploying such systems and shows the importance of assessing machine learning systems from an ethical perspective, ideally before their actual use in the real world.

2.3.2 Practical Study: Self-Fulfilling predictions in predictive policing

Objective

The goal of this third and final study is to demonstrate how self-fulfilling predictions and feedback loops can influence machine learning outcomes (full notebook with code in Appendix H). These mechanisms are not theoretical concerns, they can shape the data that future models are built on. This is illustrated using a hypothetical predictive policing scenario. Using crime data from Berlin subdistricts, we simulate how algorithmic forecasts can influence policing behavior, which then affects future data collection and reinforces the original assumptions.

This study builds on the conceptual foundations from Section 2.3. Clustering is first applied to six years of crime data to identify high-risk areas. More police are sent into these hotspots, which increases the chance of detecting crime. A feedback loop is then simulated by artificially increasing the crime numbers in these high-risk clusters over five years. The updated data is clustered again to visualize the effect of biased data collection.

Through this procedure, the study shows how a self-fulfilling prediction can evolve into a feedback loop. The aim is to make students aware of the ethical challenges and social consequences of predictive systems that can shape reality and not only describe it. It is also important for the students to acknowledge that these systems often seem to be neutral and objective, while their influence remains hidden and subtle.

Dataset and Preprocessing

The study uses real-world crime data that is published publicly by the Berlin police through the “Kriminalitätsatlas” platform (Polizei Berlin, 2023). The original dataset contains case numbers (Fallzahlen) as well as frequency values (Häufigkeitszahlen) for a variety of crime categories, covering the years 2015 to 2023. In this simulation, frequency values were used to ensure comparability across the Berlin subdistricts with different population sizes. As predictive policing systems aim to find areas with higher relative risk, frequency-based metrics provide a more useful input in this instance than absolute case counts.

The dataset that is used in this study is a processed version titled ‘HZ_2018-2023’. It focuses on the years 2018 to 2023, since crime data older than 6 years is not that relevant to detect current crime hotspots. All aggregated district data above the subdistrict level (Bezirksregionen) has also been removed. Quarters such as Mitte or Neukölln have been excluded to avoid double-

counting. The subdistrict level (e.g. Tiergarten Süd or Reuterkiez) was kept enabling a detailed spatial analysis.

The original dataset contains a wide range of crime-related variables, structured across three levels of aggregation. A global total (‘Straftaten insgesamt‘), intermediate categories (e.g. ‘Diebstahl insgesamt’, ‘Körperverletzung insgesamt’) and specific offenses (e.g. ‘Fahrraddiebstahl’, ‘Brandstiftung’). Only a subset was used for the clustering process to simulate a realistic place-based predictive policing system. The top-level total (‘Straftaten insgesamt’) was excluded to avoid redundant information. In most other cases the intermediate categories were retained while the specific offenses were excluded. This was done to maintain clarity and to reduce redundancy, as the spatial patterns of aggregated and disaggregated categories were often almost identical. For example, the hotspot patterns of ‘Sachbeschädigung’ versus ‘Sachbeschädigung durch Graffiti’, or ‘Branddelikte insgesamt’ versus ‘Brandstiftung’, were nearly the same. The same was true for ‘Raub’ versus ‘Straßenraub’, and ‘Körperverletzung insgesamt’ with its subcategories. In these cases, keeping only the broader category was sufficient for capturing the spatial information without introducing unnecessary duplication. Theft-related crimes (‘Diebstahl’) were treated differently. Here the broad category was excluded, and instead four distinct subtypes were used as separate features. Unlike the other crime types, these subcategories display different spatial patterns across Berlin. ‘Fahrraddiebstahl’ is highly concentrated in central districts, whereas ‘Wohnraumeinbruch’ and ‘Diebstahl von Kraftwagen’ are more prevalent in the peripheral areas. ‘Diebstahl an/aus Kfz’ is again more shifted towards the inner city. These spatial differences would be masked by a single aggregated ‘Diebstahl’ variable. The feature ‘Kieztaten’ was also removed because it may contain incidents already included in other categories. To maintain consistency and avoid redundant information, only non-overlapping crime types were kept.

Geographic information for the visualization of the subdistricts was taken from the official LOR shapefiles (“Lebensweltlich orientierte Räume”) provided by the Berlin Senate (Amt für Statistik Berlin-Brandenburg, 2021). These shapefiles were joined with the crime data using the six-digit LOR keys. Before applying the clustering algorithm, all numerical features were standardized using the ‘StandardScaler’ from sklearn. This makes sure that crime types with naturally higher rates do not dominate the clustering process. The resulting dataset reflects normalized crime intensities across Berlin’s subdistricts and is the foundation for the clustering and simulation of feedback dynamics.

Methodology

To illustrate how predictive policing systems can reinforce their own assumptions through feedback loops, crime data from Berlin subdistricts between 2018 and 2023 was used. The procedure is inspired by the real place-based system PredPol, which predicts the spatial distribution of future crimes based on historical data (Lum & Isaac, 2016). The basic version in the notebook for computer science students first detects crime hotspots through clustering. Then feedback is simulated, and the new data is clustered again to show changes in the hotspot distribution.

In the first step, after the dataset was prepared as described in the previous section, a KMeans clustering algorithm was applied to divide the subdistricts into distinct crime levels. After testing different configurations, four clusters were selected. The four-cluster solution offered the clearest contrasts between pre- and post-feedback results. Since the main purpose of this study is to illustrate self-fulfilling predictions and feedback loops, the configuration with the clearest outcomes was chosen.

After clustering, the centroids of each cluster were analyzed, and the clusters were ranked by their total crime volume. Each was assigned a color to represent its risk level (red = highest total crime, orange = high, yellow = medium, green = low). This representation of the model's initial prediction of spatial risk serves as the baseline for further simulation. In a real predictive policing system, such risk assessments could be used to allocate police resources.

To simulate the effect of a self-fulfilling prediction evolving into a feedback loop, data were artificially manipulated. One of the reasons to use a predictive policing system is to allocate police resources more efficiently. Areas classified as high-risk by the model receive increased police presence. This in turn leads to more crimes being detected, which does not imply that more crimes are committed, but more are being recorded. The assumption is that for each of five simulated years, all crime counts in the high-risk cluster were increased by 10%, and those in the second-highest cluster were increased by 5%. This simulates the collection of biased data due to increased surveillance and policing in these areas.

After five simulation cycles, the dataset was clustered again using the same KMeans algorithm, features and scaling. The goal of this second clustering was to show how the altered data, shaped by the initial clustering and the resulting data collection, impacts the spatial distribution of clusters. It demonstrates how model outputs can shape the new input data used for future predictions.

In the end both the original and simulated cluster maps were visualized and compared. The shift in distribution illustrates how feedback mechanisms can change the spatial crime patterns. With the Berlin data, the initial map showed multiple high-risk areas, while the simulated map was more centralized with a stronger focus on specific regions. The detailed discussion of the results follows in the next section.

Results and Interpretation

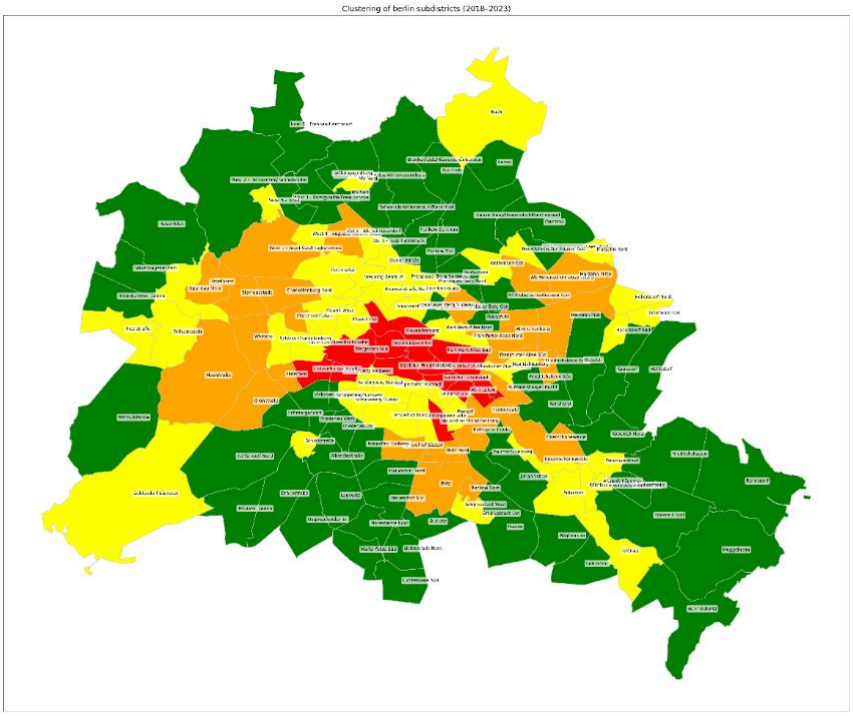
The initial clustering of the Berlin subdistricts, based on crime data from 2018 to 2023, resulted in the risk map displayed in Figure 3. Subdistricts were grouped into four risk categories, from low (green) to high crime intensity (red). As expected, the highest-risk clusters (red and orange) were located in central urban areas such as Kreuzberg, Mitte, and Friedrichshain. These districts have higher rates of offenses such as bicycle theft, drug-related crimes, assault, and robbery. The lower-risk clusters (yellow and green) are more in the peripheral areas, where car theft and burglary are more prevalent.

After simulating five years of predictive policing feedback, the clustering was repeated with the modified crime data. The resulting map (Figure 4) shows a clear change in spatial risk distribution. A single high-risk cluster emerged in Südliche Luisenstadt, Kreuzberg, while surrounding districts were reclassified to a lower risk category. This results in an even more centralized pattern, with most of the peripheral areas classified at low-risk. This change is not due to actual fluctuations in criminal activity, but the effect of the simulated predictive policing system. The two high-risk clusters receive intensified police presence and therefore more crime detection. This shift illustrates the effect of self-fulfilling predictions. What the model initially identified as risky led to more observations of crime in this area, which influenced future risk assessments. Over time, this results not in a more accurate reflection of crime distribution, but a distorted version of it. The initial assumptions of the model become increasingly true due to the biased data collection. The model's performance becomes more and more misleading over time. The model validates itself by generating the data that confirms its own prediction.

Enlarged versions of Figures 3 and 4 are included in Appendices A and B for better readability.

Figure 3

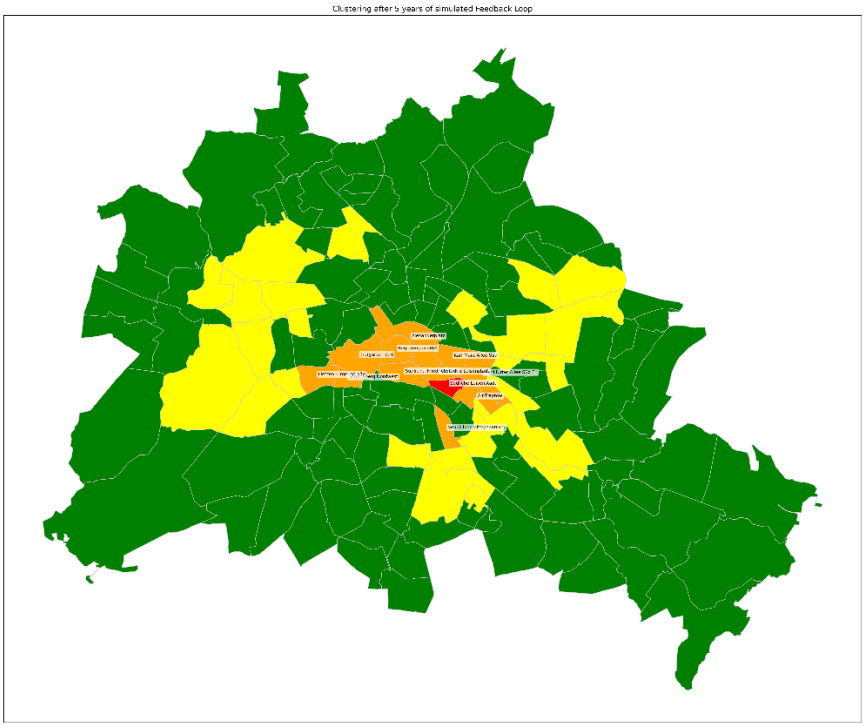
Berlin crime hotspots before simulated feedback loop (2018-2023)



Note. Based on data from the “Kriminalitätsatlas“ (Polizei Berlin, 2023) and “Lebensweltlich orientierte Räume” (Amt für Statistik Berlin-Brandenburg, 2021).

Figure 4

Berlin crime hotspots after simulated feedback loop



Note. Based on data from the “Kriminalitätsatlas“ (Polizei Berlin, 2023) and “Lebensweltlich orientierte Räume” (Amt für Statistik Berlin-Brandenburg, 2021).

The difference between the two maps highlights the risk of feedback loops in predictive systems when no countermeasures are applied. The initial clustering gave a reasonably balanced view of crime distribution across Berlin. After the feedback simulation, the balance was lost. The model focused completely on the city center, while other areas were out of the model's attention. This shift did not occur because crime decreased in those areas, but because the model's focus went to high-risk areas.

These results show how self-fulfilling predictions and feedback loops can shape a system's output over time. In a real-world setting such effects are often hard to detect, as the systems appear to work objectively and purely data-driven. This study makes these effects visible and shows students how predictive models can produce the conditions they predicted. It illustrates once more that fairness, transparency and accountability are essential problems in machine learning that need to be addressed.

Discussion

This study shows the influence of self-fulfilling predictions and feedback loops on predictive systems. What starts as a data-driven prediction of crime risk can evolve into a process that centralizes crime prevention. The simulation illustrates the risk of using model output without considering the consequences. In the context of predictive policing, these consequences can be severe, as certain districts can turn more and more into crime hotspots. The shift to more centralized hotspots did not come from actual changes in crime patterns, but from the model's influence on data generation. The study illustrates a self-fulfilling prediction that turns into a feedback loop that seems to confirm the model's prediction while, in fact, strengthening the bias. The illusion of objectivity can become dangerous when these systems are not questioned.

Ensign et al. (2017) showed that even small differences in crime rates lead to runaway feedback effects that cannot be mitigated by simply using more data. They demonstrated that the model can only converge to the true crime distribution if the influence of discovered incidents (crime identified directly through police) is adjusted, for example, through weighting or rejection sampling. In our study, no correctional measures are used, which is why the resulting risk landscape does not reflect the true crime rates. Without acknowledging and correcting the feedback effects, predictive policing models cannot be considered neutral measurement tools (Ensign et al., 2017).

From an ethical standpoint, these dynamics are concerning. Predictive systems are often deployed with the expectation that their outputs are neutral. However, as already pointed out,

they can have an influence on social structures. Areas that are heavily policed anyway will receive more attention after being labeled high-risk. Other areas will fall off the radar, not because crime decreases, but because attention shifts elsewhere. These dynamics can lead to resource misallocation, unjustified surveillance, stronger prejudice, and increased distrust among affected communities (Robinson & Koepke, 2016).

The purpose of this study, as part of the notebook for computer science students, is to make these abstract risks visible. By showing how decisions driven by model output can influence future outputs, the study encourages students to think about the long-term effects of algorithmic predictions. It highlights that a critical view on machine learning systems is essential. Not only in terms of statistical performance, but also social impact and feedback dynamics between model and environment. Understanding these interactions is essential for the next generation of data scientists or machine learning engineers. Machine learning systems do not operate in a vacuum. They interact with the world and can influence decision-making. Especially in sensitive areas it is important to understand the impact of machine learning systems. The three studies in the notebook together provide a foundation for recognizing the challenges with these systems.

3 Improving fairness in machine learning – actions and pitfalls

As machine learning models are increasingly used in high-stakes scenarios, questions of fairness and bias in these models have gained growing attention. However, improving fairness is not an easy, one-time intervention with clear guidelines. Fairness interventions are a continuous and context-dependent process. They must consider technical, social and legal dimensions, and they often involve trade-offs between competing goals such as accuracy and equity. This chapter explores the main obstacles when building fair models or assessing existing systems. Based on the three case studies introduced in the previous chapter, it shows how different challenges emerge in real-world applications and shows why there is no single solution to algorithmic fairness. The focus of this work is creating awareness for ethical problems in machine learning (Chapter 2) and challenges that occur when trying to create fair models (Chapter 3). Chapter 3.2 provides an overview of general practical implications for improving fairness in machine learning systems. Concrete methods are only briefly mentioned and not discussed in detail. Model cards and the Z-Inspection® framework are examined in more detail as exemplary approaches.

3.1 Challenges in achieving fair machine learning

Ensuring fairness in machine learning is a complex task that involves more than simply focusing on model performance and technical adjustments. Fairness is context-sensitive, multidimensional, and often incompatible with other optimization goals. The following section highlights major challenges in the development of fair machine learning systems. These challenges will be illustrated using the case studies introduced in Chapter 2.

3.1.1 Conceptual limitations of fairness definitions

Fairness in machine learning is often operationalized through mathematical definitions. While these metrics are useful tools to detect and quantify disparities, they rest on simplifying assumptions that may not always lead to the best results in practice.

A general conceptual limitation of statistical fairness definitions is their reliance on correlations rather than causality. Most machine learning models are trained on observational data and learn statistical associations, not causal relationships. This distinction is important. A model might find that patients with asthma are less likely to die from pneumonia. Not because asthma is protective, but because these patients receive more aggressive treatment. Using this correlation to change care decisions could lead to harmful outcomes. The problems that arise when correlation is mistaken for causality can also be illustrated through the study on predictive

policing. Crime prediction models may learn that certain neighborhoods are high-risk, not due to higher intrinsic crime rates, but because of more police presence. If such correlations are treated as causal, this can reinforce biased enforcement strategies and fail to address the actual causes of crime. In general, fairness interventions that are based purely on observational patterns risk reinforcing existing inequalities when they ignore the underlying causal mechanisms (Barocas et al., 2023). This is why causal reasoning should be integrated whenever possible. As mentioned in Section 2.1.3.1, statistical fairness measures are better understood as diagnostic tools rather than direct solutions. However, as noted in Section 2.1.3.3, causal approaches come with their own challenges. For example, they require strong domain expertise and are not always feasible in practice.

An already discussed issue is the mathematical incompatibility between fairness metrics. For example, it is generally impossible to satisfy equalized odds and calibration across groups simultaneously in cases where base rates differ, unless the sensitive attribute and the target variable are unrelated, or the model is perfectly accurate (Chouldechova, 2017; Kleinberg et al. 2016). This incompatibility is not only technical but shows tensions in the normative conception of fairness. Equalized odds focuses on error rate parity, aiming for procedural equality at the group level. Calibration on the other hand emphasizes predictive accuracy, ensuring that a score means the same thing regardless of group membership. These goals often pull in different directions. Adjusting a model to equalize error rates typically alters the risk scores, breaking calibration (Pleiss et al., 2017). The first case study on student success prediction illustrates this trade-off. Applying threshold-based post-processing to achieve equalized odds led to worse calibration.

Another limitation of many statistical fairness definitions lies in what Corbett-Davies et al. (2023) describe as the “problem of inframarginality”. Fairness notions such as statistical parity or predictive equality operate on infra-marginal statistics. They assess model behavior away from the decision boundary. Many critical fairness considerations however are concerned with what happens at the margin, where decisions are uncertain and small changes can determine who receives a positive outcome. These marginal cases are often most sensitive to fairness concerns, because they reveal whether the same standards are applied to different groups in cases close to the decision boundary. When groups have different risk distributions, as is nearly always the case, using the same threshold for all can violate statistical parity, but adjusting thresholds to enforce parity can result in utility loss and unintended harm. To illustrate this, Corbett-Davies et al. (2023) use the example of diabetes screening. In threshold-based decisions

individuals receive a positive decision (being admitted, screened, approved) if their estimated risk or score exceeds a certain cutoff value. In the case of diabetes screening, using a uniform threshold of for example 1.5% risk means that more individuals from groups with higher baseline risk (e.g. Asian Americans) are screened. While this maximizes utility, it violates statistical parity, which requires equal screening rates across groups. Enforcing statistical parity would require raising the threshold for Asian Americans and lowering it for others. This would lead to denying tests to high-risk individuals and more screening of low-risk people, which undermines fairness as well as health goals. This tension arises because many fairness criteria do not focus on the marginal benefit of decisions but rather on overall error rates that ignore where in the distribution these errors occur. As a result, interventions based on such fairness metrics can lead to decisions that harm all groups involved. High-risk individuals may miss out on screenings, leading to worse health outcomes. Low-risk individuals may receive unnecessary procedures, costing time and money. In such cases, fairness interventions are pareto-dominated. Meaning that there exists an alternative policy (e.g. uniform thresholds) that would improve outcomes for all groups involved, without worsening conditions for any one group. This leads to the broader concern about whether fairness constraints truly serve the population they aim to protect. When the context is not taken into account the people that should be protected by the fairness intervention may unintentionally be harmed (Corbett-Davies et al., 2023).

Corbett-Davies et al. (2023) demonstrate that many fairness constraints lead to pareto-dominated policies. This challenges the assumption that satisfying fairness metrics necessarily leads to fairer outcomes. Enforcing fairness notions can reduce the overall utility of a system. For example, in a fair college admission scenario the objectives academic preparedness and diversity need to be balanced. In a setting where only a limited number of students can be admitted, these objectives can come into tension. Strict implementation of fairness definitions may reduce academic preparedness and student diversity. Instead of achieving a desirable trade-off, fairness-constrained models often move decisions away from the pareto frontier (the set of optimal policies that maximize competing objectives). Similar findings apply to medical resource allocation and pretrial risk assessments, where fairness-constrained models perform worse for everyone, including the disadvantaged groups they wanted to support (Corbett-Davies et al., 2023).

These insights highlight that fairness should not be seen as a static constraint that can be optimized in isolation. Approaching fairness instead requires balancing competing goals, such as accuracy, equity, and social impact, based on the context. Stiff mathematical definitions for

decisions are not the solution for improving fairness. Flexible and outcome-oriented approaches are needed, that can account for social context and utility. This consequentialist perspective underlines the importance of assessing real-world impacts on affected groups or people. As the case studies show, fairness goals and trade-offs vary a lot. The procedure for decision fairness in the first study and representational fairness in the second study differed and demanded acknowledgment of the context to understand fairness issues.

3.1.2 Context dependence of fairness metrics

A core difficulty in achieving fairness in machine learning that was already mentioned several times in this work is that neither the definition nor the appropriate fairness metrics are universal. Fairness is strongly context-dependent and influenced by the application domain, societal norms, and the form of potential harms (Osoba & Welser, 2017). What is a fair decision in one setting may be ethically inappropriate in another. Fairness is a sociotechnical construct and cannot be captured through mathematical definitions alone (Selbst et al., 2019).

This becomes clear when comparing the three case studies. In the first study, which dealt with predicting student dropout, the main question was which observational fairness definition should be satisfied. The choice between these metrics depends on which approach to fairness is prioritized. Optimizing separation-based metrics led to a trade-off with the other definitions. The right choice was not obvious without looking at the context and thinking about what fairness depends on in educational decision-making.

The second study, on gender classification, moved from decision to representational fairness. It showed that the approach to representational fairness differs. For instance, fairness metrics like statistical parity or calibration were insufficient. These metrics assume that the prediction target is neutral and quantifiable, which is problematic when the target itself is a sensitive attribute. As Binns (2018) argues, representational harms require alternative forms of assessment that take visibility and cultural context into account. Metrics such as group-wise accuracy or separation-based metrics remain useful as diagnostic tools, especially from an intersectional perspective to uncover disparities (Buolamwini & Gebru, 2018).

The predictive police case presents yet another fairness context. Here, predictive models are applied in a setting with deep historical and political implications. The fairness assessment must reflect societal concerns, such as over-policing in marginalized neighborhoods. A technically fair model that minimizes aggregate errors can still reproduce existing inequalities. Therefore, evaluating fairness demands sensitivity to local context and historical injustices (Barocas et al.,

2023). Since this study uses clustering rather than classification, it requires a different procedure. Fairness in clustering cannot be captured by standard classification fairness metrics.

Across all three cases, the main insight is that fairness metrics cannot be applied in isolation to solve a problem. Their relevance and limitations are dependent on the context the system operates in. Fairness metrics are best used as a diagnostic instrument that helps to find and reflect disparities, but do not offer direct conclusions on their own. Addressing fairness requires looking at the whole sociotechnical system (Selbst et al., 2019).

3.1.3 Representation and measurement issues

Another fundamental challenge in achieving fairness lies in how machine learning systems represent the world. Not only through the data but also the choice and construction of target variables and labels. While much of the fairness discourse is focused on model performance and output, unfairness is often introduced earlier. It can come from representational problems in the data, from measurement issues or from the way labels are defined.

This can be seen in the first study. The label ‘dropout’ appears to be an objective outcome, but it is shaped by institutional and social factors such as financial problems or systemic discrimination. These influences can force capable students to leave university even though they have the potential to succeed. If a model is trained to predict dropout without considering these underlying causes, it could learn to associate social disadvantage with academic failure. Even when following support offers are well-intentioned, they could reinforce structural inequalities if the model treats historically disadvantaged groups as inherently high-risk, without addressing the reasons for that risk. This can lead to interventions that stabilize disparities rather than challenge them, especially with missing explanations for the support (Barocas et al., 2023).

In the gender classification study, different issues with representation are prominent. Here the task is not the prediction of an outcome, but to assign a category (gender) based on images. A challenge is the assumption that gender is an observable, fixed and binary category. This ignores that gender is socially constructed and fluid. Even if a model achieves parity in classification accuracy across subgroups, it can still be harmful towards people that do not fit this binary framework. Moreover, the labels in datasets like UTKFace and FairFace are often inferred or assigned by annotators, they are not self-identified. This can raise questions of label validity (Scheuerman et al., 2020).

The predictive policing case illustrates yet another problem with representation and measurement. The model is trained to predict crime risk based on historical police data, in this case recorded incidents. However, these are not neutral reflections of true crime rates, they are shaped by prior patterns of surveillance, reporting behavior, and over-policing in certain neighborhoods. Using this data as ground truth means the model is trained to predict police activity and not necessarily where crime occurs. The model will learn and reinforce those patterns, creating feedback loops that amplify rather than correct injustice (Lum & Isaac, 2016).

The three case studies show that fairness cannot be addressed without considering representation and measurement. Fairness requires questioning the validity and meaning of what we ask the model to predict. Metrics like calibration or error rate parity may signal technical fairness, but if the target variable is poorly designed, the system can remain unfair in practice. Addressing this challenge involves rethinking the labels themselves. For instance, the selection of more meaningful target variables or designing systems that allow contestation of predictions by affected groups (Selbst et al., 2019). Fairness is not just about fixing biased outputs but about rethinking what we measure and why.

3.1.4 Lack of standards, transparency, and accountability

One more obstacle in building fair machine learning systems lies in the absence of standards, transparency, and accountability. Despite the increasing attention to ethical machine learning, there is still no widely accepted framework for deciding when a system is fair or not, how it should be evaluated, or who is responsible for its behavior (Ntoutsi et al. 2020; Olteanu et al., 2019). This lack of standardization is problematic for the trustworthiness of these systems.

In the student dropout study, there is no consensus or clear recommendation on which fairness metric to prioritize. As already discussed in Chapter 3.1.2, the appropriate fairness metric depends strongly on the specific context and the type of harm being addressed. Without standards or guidance, the modelling choices and evaluation criteria remain arbitrary and context-dependent. This makes design and assessment of fairness complicated (Pagano et al., 2022). The context-dependence is one of the reasons why it is hard to establish universal standards for fairness evaluation in machine learning. And missing standards in turn strengthen issues around transparency, accountability, and explainability.

These issues are extremely relevant, as they directly affect how decisions are experienced by the individuals affected by automated decision-making. In all three case studies, affected groups would have little or no understanding of how predictions are generated or why they are

categorized in a certain way. In the student success prediction study, students labeled with high-risk of dropping out may receive targeted support without really understanding the reason behind this classification. If no explanation is provided, this can lead to confusion, stigmatization, or mistrust, especially when the student's self-perception differs from the prediction.

In the gender classification study, the models used operate as black boxes. Unlike models that produce decisions based on observable features or thresholds, these models provide no information on how the model came to its prediction. As a result, individuals affected by the prediction cannot understand or challenge the decision. In the HR hiring tool scenario, the lack of transparency can mean that individuals may not even be aware that their application was pre-filtered by a gender classifier. This form of hidden automation removes the possibility for contestation and consent. Such systems operate invisibly but still influence opportunities and outcomes without the affected people knowing (Eubanks, 2018, as cited in Barocas et al., 2023).

The predictive policing study highlights the accountability problem. It is often unclear who is responsible for the outcomes of algorithmic predictions. The developers of the system, the police officers using it, or the institution deploying it. As already mentioned, predictive policing systems are often proprietary and therefore not transparent. Even police officers themselves do not always understand how the predictions were generated (Robinson & Koepke, 2016). The lack of transparency leads to limited explainability. The less we can understand these systems, the fewer possibilities of meaningful fairness interventions there are.

Efforts to improve transparency and accountability have led to proposals such as model cards (Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2018). These aim to document model assumptions, limitations, and performance differences across groups. While these tools can help to increase awareness and comparability, they are not widely adopted in practice yet as there are no legal requirements. Section 3.2 discusses model cards in a bit more detail.

All in all, a lack of standards and with that limited transparency not only makes evaluation of fairness difficult but also undermines trustworthiness of machine learning systems. Without clear accountability, affected individuals have no opportunity to challenge decisions that might harm them. Fairness in machine learning therefore cannot be solved with purely mathematical metrics. It needs institutional support that requires machine learning systems to be explainable, transparent and accountable.

3.1.5 Sociotechnical and institutional barriers

Whereas Section 3.1.2 addressed how fairness metrics must be tailored to specific contexts, this section focuses on institutional and structural challenges that arise independently of the choice of metrics. Fairness is not only context-dependent, but also part of a complex sociotechnical system that determines what is practically achievable. While legal and regulatory aspects are certainly relevant, they are not part of this thesis, as they can change quickly and differ between geographical areas.

Ensuring fairness in machine learning is not only a matter of refining algorithms or tuning metrics. It involves including the sociotechnical system. Challenges in fairness do not only come from data and model, but also from the structures and norms that surround the system (Ntoutsi et al., 2020; Selbst et al., 2019).

This is particularly evident in the third case study on predictive policing. Even if the model itself worked perfectly from a technical point of view, it would still operate within historically biased policing practices and decision-making frameworks. Technical interventions cannot eliminate such structural factors. They even risk legitimizing practices that lead to unequal outcomes (Lum & Isaac, 2016). In high-stakes domains such as criminal justice, the broader sociotechnical environment is especially critical in shaping the fairness of algorithmic systems.

Moreover, machine learning systems often interact with multiple stakeholders that can have conflicting goals. In the dropout prediction system, for example, educators may be more concerned with fairness and policymakers like people responsible for the budget may be more focused on efficiency. Implementing fairness constraints could conflict with legal requirements or university goals. For instance, providing support to all students flagged as high-risk to drop out might be advisable from an ethical standpoint, but it could be financially unsustainable. As a result, the institution might choose to support only a small group of students.

The gender classification study reveals another structural challenge. In the imagined HR application scenario, the gender recognition model is likely provided by an external vendor. This makes it difficult for the organization that uses the tool to modify the underlying data or modeling decisions. Institutions become dependent on pre-trained black-box systems, where neither the labeling process nor the model architecture is transparent. This lack of control limits the ability to identify or mitigate representational harms. Even more if the model is part of a broader recruitment workflow without clear oversight or responsibility.

Another institutional barrier is the lack of interdisciplinary expertise. Fairness assessments require knowledge beyond computer science, including perspectives from law, sociology, and ethics. As Selbst et al. (2019) point out, when teams do not integrate such perspectives, they often adopt overly narrow notions of fairness that miss real-world consequences. Cross-disciplinary collaboration is essential not only for identifying relevant harms, but also to put fairness interventions into context.

A barrier to recognizing fairness concerns in the first place lies in the lack of diversity within development teams. Many teams, especially in the tech sector, are composed of individuals with similar social and cultural backgrounds. They can overlook problems that affect marginalized groups or misinterpret the impact of certain model decisions. Biases in training data or model assumptions may remain unnoticed if no one within the team has the experience or perspective to question them. A broader range of perspectives helps to identify harms that might appear neutral to people that will not be affected. Increasing team diversity is therefore a step toward more robust and socially aware machine learning systems (Lee 2018; Osoba & Welser, 2017; Yapo & Weiss, 2018).

Even when fairness improvements are technically possible, they are often difficult to implement once a system is already in use. As Calegari et al. (2023) emphasize, adding fairness interventions after deployment is usually unrealistic due to the operational complexity. This includes challenges such as integrating changes into existing workflows, coordinating across departments, and dealing with performance trade-offs. As a result, fairness concerns that are not addressed early in the development process often remain unresolved in practice.

In sum, achieving fairness needs more than model improvements. It demands addressing the institutional conditions under which these models are designed and used. Without that, even technically flawless models can reinforce unfair structures or create new forms of inequality.

3.1.6 The five abstraction traps

Many contextual, technical, or institutional challenges in achieving fairness in machine learning have been discussed so far in this chapter. A lot of these challenges are not isolated technical problems, they come from a tendency in machine learning to abstract away social context to build generalizable, modular systems. To better understand why well-intended fairness interventions often fail in practice, Selbst et al. (2019) describe five abstraction traps. Although some of the underlying ideas have already appeared in previous sections, these traps offer a

useful perspective. They can be used as a conceptual tool for recognizing limitations in the way fairness is approached.

The framing trap

This trap occurs when fairness is treated as a problem to be solved within the model, rather than looking at the broader sociotechnical system in which the model operates. For instance, the fairness debate is often concentrated on whether the predictions match the labels. This assumes that the choice of labels, features, and problem formulation are unproblematic. In the dropout prediction case, the label dropout could reflect structural disadvantages (see 3.1.3). Yet these broader consequences are not captured in standard fairness metrics. A focus that only looks at model performance ignores the sociotechnical system in which these decisions take place (Selbst et al., 2019).

The portability trap

Machine learning often wants generalizable solutions that can be reused across contexts. But fairness is not as portable as code is. Fairness metrics that work in one setting do not need to make sense in another. This can be seen in the gender classification study, where metrics like statistical parity or calibration failed to capture representational harm. As discussed in 3.1.2 fairness metrics must be carefully chosen given the context, there are no universal solutions (Selbst et al., 2019).

The formalism trap

Attempts to mathematically define fairness often ignore its procedural, normative, and contested nature. While these definitions are useful diagnostic tools, they cannot fully capture what fairness means in a specific context. As discussed in 3.1.1, different fairness metrics focus on different values and can be incompatible. Formalizing fairness can create the illusion that fairness is solved once a particular metric is optimized (Selbst et al., 2019).

The ripple effect trap

Even a technically fair model can disrupt the social system it is embedded in. For example, in predictive policing, the algorithmic forecast could influence officer behavior or alter community dynamics in unpredictable ways. Fairness assessments that ignore these effects miss the full picture. The goal should not only be local fairness at the moment of prediction but also to understand how the model and social system influence each other (see 3.1.5) (Selbst et al., 2019).

The solutionism trap

This trap occurs when technical solutions are applied to problems that might not need to be solved technically. Rather than asking whether machine learning should be used in a domain at all, fairness research often begins with the assumption that automation is beneficial. Yet in the gender classification study, using a model to sort applications by gender is susceptible to errors and therefore offers little practical value. The effectiveness of predictive policing systems to reduce crime in comparison to traditional methods is also not proven (Robinson & Koepke, 2018). Machine learning systems are not always the best answer, and to ask whether an automated solution is necessary at all, should be the first step.

In fact, Selbst et al. (2019) recommend turning the traps around and approaching fairness in reverse order. That means first asking if the technical solution is appropriate for the problem in the first place (solutionism). Then think about how the system influences the social environment it operates in (ripple effect). Afterwards check which notions of fairness are relevant, and if they are contestable (formalism). Then ask if the solution fits the given context after all (portability) and lastly if the framing of the problem reflects the actual goals and actors involved (framing). Keeping these traps in mind does not guarantee a fair system, but it enables a more critical development or assessment process. These traps complement the challenges discussed in earlier sections and clarify that fairness in machine learning is not simply a technical, but a complex sociotechnical problem. The traps can also be seen as a shift from challenges towards practical implications, which the next section will discuss in more detail.

3.2 Practical implications for handling fairness concerns

This chapter highlights general insights on how fairness in machine learning can be improved in practice. It does not provide a list of concrete technical interventions but rather shows some key implications that are proposed in literature. As fairness in machine learning is a young discipline, there is still a lot of movement and debate in this field. Therefore, this chapter focuses on broadly agreed implications.

It is important to note that many of the recommendations discussed in the following sections are unlikely to be widely implemented without legal enforcement or regulatory pressure. Legal regulation, however, is not in scope of this thesis. Still, some companies may voluntarily adopt ethical self-regulation. This is not purely out of intrinsic interest in fairness, but to avoid binding legal oversight (Boonstra et al., 2024).

3.2.1 Fairness is a process, not a fix

Improving fairness in machine learning is not a one-time technical fix, it is an ongoing iterative process. As discussed in the previous section many fairness interventions fail because they oversimplify or overlook the social context in which the machine learning system is deployed. Fairness is not a property of the algorithm alone. It comes from how the system is designed, implemented, and evaluated, in interaction with its context and the people it affects (Selbst et al., 2019). This process-based view implies that fairness is relevant in the entire machine learning lifecycle. From problem definition and data collection to model development and deployment to the monitoring of the systems. Each phase brings distinct risks for fairness. Therefore, addressing fairness needs more than optimizing a single metric, it requires a sociotechnical perspective that considers the broader context in which the system operates.

Corbett-Davies et al. (2023) advocate for a consequentialist perspective that aligns with this view. They argue that fairness constraints often lead to pareto-dominated outcomes. As a result, evaluating fairness should consider the real-world impact of algorithmic systems. Chapter 3.1 came to a similar conclusion. Fairness is deeply context-dependent and shaped by many influences. Fairness is not guaranteed with metrics but rather with developing a fair process, which includes aspects such as stakeholder inclusion, transparent documentation, or interdisciplinary collaboration. This shift from solution-based to process-based approaches is supported by many authors (Binns, 2018; Mitchell et al., 2019; Selbst et al, 2019). It shows that fairness is not a final state to be achieved, but more a commitment to continuous reflection.

The following sections explore how this process-oriented understanding of fairness can be translated into practice.

3.2.2 Integrating fairness into the machine learning lifecycle

As discovered in the previous section, fairness needs to be integrated throughout the entire lifecycle of a system. Each stage has different challenges but also opportunities to mitigate bias (Mitchell et al., 2019). A useful distinction is made between pre-processing, in-processing, and post-processing methods. The strategies target different stages in the development process and highlight that fairness can be addressed in many ways. This section provides a rough overview of each phase, the technical details of specific mitigation methods are beyond the scope of this thesis and will not be discussed in depth.

Pre-processing approaches aim to detect and reduce bias in the data before model training. This can include sampling strategies to balance datasets, transforming features to remove correlations with sensitive attributes, or relabeling instances. These techniques are useful when access to the model is limited, as they intervene before training. The downside is that they often require strong assumptions about what makes a dataset fair (Calegari et al., 2023; Ntoutsi et al., 2020).

In-processing methods try to improve fairness by modifying the algorithm itself. For example, this can mean adding regularization terms that penalize unfair outcomes or use adversarial learning that reduces the influence of sensitive attributes. These methods tend to be more powerful and flexible, but they require knowledge about the model internals and are usually more complex to implement (Calegari et al., 2023). Moreover, the trade-off between fairness and accuracy becomes more visible in this phase, as fairness constraints are built directly into the model's learning objectives. This means that any gain in fairness can have a measurable impact on predictive performance, making the relationship between the two easier to observe (Zafar et al., 2017).

Post-processing techniques adjust model outputs to satisfy fairness criteria after training. For instance, threshold adjustments (e.g. Fairlearn's TresholdOptimizer), probabilistic relabeling, or group-specific calibration (Calegari et al., 2023). These methods are independent of the model and easier to implement, especially when the model is already trained and cannot be modified. However, they can affect consistency and calibration and rely on access to protected attributes. The retrospective adjustment can also raise legal and ethical concerns in some contexts (Ntoutsi et al., 2020).

These methods can be combined and there exist a huge variety of approaches for many contexts. They all require a conscious decision about which type of fairness is most relevant and what trade-offs are acceptable in a given context. As discussed in Chapter 3.1, these decisions must take the sociotechnical scenario into account (Barocas et al., 2023; Corbett-Davies et al., 2023).

On top of that, fairness considerations should not end with the deployment. Real-world environments are dynamic, and systems can behave differently when they interact with people or institutions. Therefore, continuous evaluation and monitoring are crucial in keeping these systems fair. This is especially true in high-stakes areas where the cost of unfairness is high (Osoba & Welser, 2017).

To be able to integrate fairness into the machine learning lifecycle and to choose the right intervention method, transparency is needed. Insights into design choices, data assumptions, and model behavior are important information, but often not available. Documentation practices such as model cards (Mitchell et al., 2019) or datasheets for datasets (Gebru et al., 2018) have been proposed to fight the opaqueness of machine learning models. These tools help to communicate system limitations and threats to fairness to developers and stakeholders. A more detailed discussion of such tools and the importance of transparency follows in the next section.

3.2.3 Transparency, accountability, and explainability

One of the most important factors in improving fairness in machine learning is communication. Transparent and explainable models are essential to make sure that affected individuals and stakeholders understand, trust, and can contest the outputs of algorithmic systems. Without communication about fairness in machine learning models, these systems remain black boxes that are not accessible to those that are impacted the most (Barocas et al., 2023; Ntoutsi et al., 2020; Suresh & Guttag, 2021).

Transparency begins with documenting the development process, data origin, and assumptions made at each step. Tools such as datasheets for datasets (Gebru et al., 2018) and model cards (Mitchell et al., 2019) have been proposed to increase and standardize transparency, by including information on datasets, model usage, and performance across demographic groups. These tools help users or auditors to assess the fairness of a model and if it is suitable for a given context. Mitchell et al. (2019) emphasize that model cards should include intersectional evaluation, intended use, and limitations. With this information the likelihood of using systems in unsuitable contexts is reduced. Transparency is not only about documentation, access and communication are essential as well. The public demand for transparency and the corporate

interest to keep models proprietary is a tension to be solved, especially in public sector applications such as policing or education (Lee, 2018).

Transparency alone is not sufficient. Even when models are well-documented, they may not be understandable. Explainability refers to the capacity of a system to show how and why a decision was made. Ntoutsi et al. (2020) distinguish further between explainability and interpretability. Explainability referring to how understandable a model is, while interpretability focuses on predicting changes in outcomes. Many current machine learning models fall short on both. Interpretable models, such as rule-based systems, are preferable in sensitive contexts. Explanation techniques like LIME can help with the interpretation of black-box models, but they are mostly useful for domain experts (Ntoutsi et al., 2020). A central challenge is that explainability is often in conflict with other requirements, such as performance or privacy. Greater explainability can reduce model complexity and lead to performance trade-offs. Privacy concerns could limit how much information can be shared, especially when sensitive attributes are involved in the decision process (Schmidt et al., 2024).

Another essential factor of fairness in machine learning systems is stakeholder communication. The people affected by machine learning systems often have little awareness of being evaluated, and therefore no chance to contest outcomes (Barocas et al., 2023). For instance, in the gender classification study individuals were sorted by gender without ever being informed that this took place. Or in predictive policing systems police officers often do not understand how exactly predictions are generated (Robinson & Koepke, 2016). This missing transparency is related to questions about accountability. It is not clear who is responsible for algorithmic decisions, the developer, the deploying institution, or the user. Transparency and explainability are important steps toward accountability, as they make visible how decisions are made, based on what assumptions (Pagano et al., 2022). However, they are not guarantees of accountability. It needs clearly defined responsibilities that are in the best case integrated into the institutional structure. This includes documenting who made which design choice, establishing roles for oversight and mechanisms to appeal decisions. As Osoba and Welser (2017) emphasize, accountability together with transparency is essential to achieve fairness in machine learning systems.

Ultimately, improving transparency, explainability, and accountability requires a cultural shift towards inclusive and participatory system development. This includes diverse teams, stakeholder involvement, and ethical awareness throughout the machine learning lifecycle

(Howard & Borenstein, 2018; Selbst et al., 2019; Yapo & Weiss, 2018). These aspects will be discussed in the next section.

3.2.4 Tools and institutional practices

Another implication for improving fairness in machine learning is the use of institutional practices and supportive tools. More and more tools emerge to support developers in identifying and addressing fairness concerns. Toolkits such as AIF360 (LF AI & Data Foundation, n.d.), Fairlearn (Fairlearn Organization, n.d.), or SageMaker Clarify (Amazon Web Services, n.d.) offer ways to investigate models for disparities across demographic groups and visualize fairness metrics across subpopulations. These tools can help uncover hidden biases in training data or model predictions. They can also support fair model development or post-hoc adjustments. However, the usefulness of these tools depends on how they are implemented and interpreted. Without involvement from domain experts and stakeholder input, they can become more symbolic than effective interventions.

Institutional practices play a central role in making fairness considerations essential in machine learning development. As already discussed in Section 3.2.3, transparent documentation helps to assess the suitability and fairness of a model. Formats like model cards make it easier for reviewers or users to understand what a system is designed for. Beyond documentation, internal governance mechanisms also support the development of fair machine learning systems. Organizations could establish fairness review boards, ethics committees, or impact assessment protocols that provide oversight. This would make sure that decisions about fairness are not left to individual developers alone but are handled collectively and transparently (Raji et al., 2020). In addition, teams could receive training on fairness, ethics, and the social implications of algorithmic decision making, to ensure that these aspects are considered (Barocas et al., 2023).

The composition of teams is another crucial factor to fair machine learning systems. Diverse and inclusive teams are important to detect unfairness in development or assessment of applications. As Lee (2018) illustrates, the lack of diversity in the tech industry has contributed to serious blind spots. For example, the case where a photo tagging algorithm labeled Black individuals as “gorillas”. A more diverse team, with greater sensitivity to social and cultural issues, could have recognized and avoided such harm. Diverse teams are better equipped to identify representational harms and to anticipate how systems might fail for different populations (Lee, 2018).

In a similar way, interdisciplinary collaboration is important. Fairness in machine learning cannot be addressed by computer scientists alone. Insights from areas like law, ethics, or sociology, and domain-specific experts are essential to understand all aspects of fairness in a broad sociotechnical system (Suresh & Guttag, 2021).

Additionally, it can be recommended to involve those affected by algorithmic systems. Participation in development allows for a better and context-aware understanding of fairness (Howard & Borenstein, 2018). Fairness should not be defined exclusively by engineers or researchers. It must also take the experiences of the impacted communities into account. Institutions would need to open up for these types of dialogues (Barocas et al., 2023).

On top of that the timing of addressing fairness concerns is essential. Ensuring fairness after a system is already deployed in an organization is often not effective, due to operational constraints. Instead, institutions should integrate fairness considerations from the beginning of the development process (Calegari et al., 2023).

External audits can also be a way to assess and improve fairness for organizations, complementing internal processes (Law et al., 2020). Section 3.2.5 will present the Z-Inspection® framework as one example of an external evaluation approach for complex, high-impact machine learning systems.

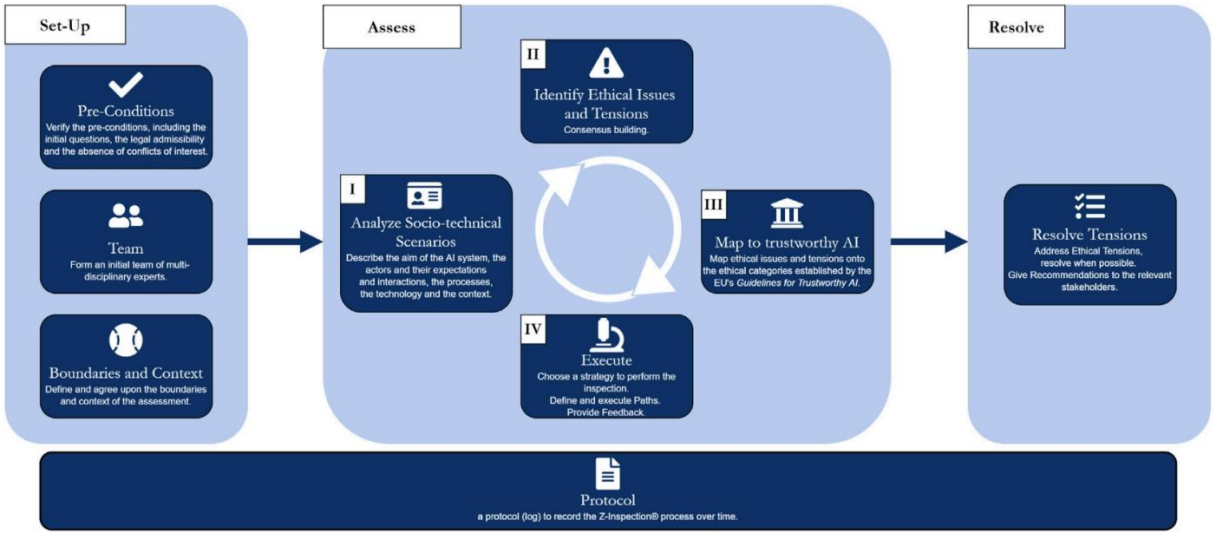
3.2.5 Z-Inspection®

Many of the practical implications discussed in this chapter, such as treating fairness as a process, the importance of interdisciplinary collaboration, and the importance of paying attention to the whole sociotechnical system, are part of the Z-Inspection® framework. Z-Inspection® is a holistic, context-sensitive methodology for assessing the trustworthiness of machine learning systems and is focused on ethical concerns. To help with the challenges of applying abstract principles of AI ethics in real-world settings, it provides a structured but also flexible approach to evaluate fairness (Zicari et al., 2021; Boonstra et al., 2024).

Z-Inspection® consists of three core phases, shown in Figure 5, which guide an interdisciplinary team of experts through the evaluation of a given machine learning use case.

Figure 5

The Z-Inspection® Process



Note. From Zicari et al. (2021).

The set-up phase defines the scope of the assessment, selects a diverse group of experts and identifies relevant stakeholders. Independence and no conflict of interests are important, especially in high-risk domains such as healthcare.

In the assess phase sociotechnical scenarios are developed to check the consequences of the system. Ethical tensions are identified in working groups, not abstract but through specific dilemmas. These issues are then mapped from an open description to the structured vocabulary of the ethics guidelines for trustworthy AI. The guidelines include four key ethical principles and seven requirements for machine learning systems (European Commission, 2019). The mapping helps to align the diverse perspectives and find a shared understanding.

In the resolve phase ethical issues are discussed with the goal of developing concrete recommendations. These include technical mitigations, governance reforms, or even the recommendation to not deploy a system because fairness or rights cannot be ensured. This process involves finding consensus and balancing competing ethical principles, for example fairness versus accuracy (Zicari et al, 2021).

A key factor of Z-Inspection® is the importance of interdisciplinary collaboration. As already emphasized in Sections 3.2.1 and 3.2.4, addressing fairness in machine learning requires integrating diverse perspectives and knowledge. Z-Inspection® implements this through collective ethical reflections in each phase (Boonstra et al., 2024).

Moreover, the framework shows how fairness cannot be reduced to metrics alone. For example, in the assessment of a machine learning system designed to detect cardiac arrests from emergency calls, fairness concerns emerged due to poorer model performance for non-native Danish speakers. These issues were not only technical but also affected trust in the system, raised questions of accountability, and highlighted the need for stakeholder involvement during development (Zicari et al., 2022). In another case, Z-Inspection® was combined with the Dutch Fundamental Rights and Algorithms Impact Assessment (FRAIA) to evaluate a machine learning system for monitoring nature reserves. This assessment uncovered tensions between ethical goals and legal norms, for instance ensuring environmental protection versus the appropriate use of satellite data (Boonstra et al., 2024). These examples demonstrate that fairness must be seen in the broader institutional, legal, and societal contexts.

Machine learning systems are not static and new ethical issues can arise when a system is used in another context or by different users. In Section 3.2.2 the need for continuous evaluation was highlighted. The iterative and participatory approach of Z-Inspection® reflects this need by emphasizing that fairness must be reassessed throughout the entire machine learning lifecycle.

In sum, Z-Inspection® illustrates how abstract values like fairness, transparency, and accountability can be translated into a practical and dynamic evaluation process. It stresses the central message of this chapter. Fairness in machine learning is not a property of models alone, but of the sociotechnical system they are part of. This is why fairness must be addressed through collective, interdisciplinary, and context-aware methods like Z-Inspection®.

4 Creating a notebook

The practical outcome of this thesis is an interactive Jupyter notebook for computer science students. It was developed to support students in understanding and reflecting on the ethical challenges of machine learning. The notebook is designed as a teaching tool for university-level courses and aims to raise awareness, deliver key concepts, and encourage critical thinking about fairness in machine learning. It is primarily aimed at computer science students, because familiarity with Jupyter notebook and an understanding of Python code are required. However, students from other disciplines, such as ethics, with basic programming knowledge, could also find the notebook valuable. Typographic emphasis was used as a stylistic device to highlight key aspects and terms. The estimated completion time is between two and three hours, depending on how thoroughly it is worked on.

The content of this thesis is transferred into the notebook in a compact form. It is divided into seven separate parts and the structure follows this thesis closely. The division into multiple notebook pages was made deliberately to reduce cognitive load and maintain attention. To increase active engagement, in the sections about the practical studies, the notebook alternates between explanatory markdown sections and executable code blocks. This structure enables students not only to observe the results of code but also to run it themselves and understand the logic behind each step. Each of the seven parts concludes with one true/false question and two multiple-choice questions to encourage reflection on the learned content. The solutions, as well as a short user guide, are provided in a separate PDF file. The notebook has a consistent and clean layout, using the formatting options available in Jupyter markdown sections. The practical studies included in the notebook were selected to connect the theoretical concepts discussed in this thesis with possible real-world use cases. Attention was paid to choosing datasets that students would find relevant and interesting.

The full notebook is available on GitHub: https://github.com/LukasWel/ethical-challenges-in-ml. Images of the notebook can be found in Appendix C-I.

Part 1: Motivation

The first part introduces the topic and explains why fairness in machine learning is highly relevant for computer science students. The widely discussed COMPAS case was chosen as a starting point because it illustrates many relevant aspects. The problem is understandable without any prior knowledge, but it also introduces deeper topics such as the tension between

equalized odds and calibration. The section guides students step by step through the case and includes reflective questions to encourage critical evaluation of the results (see Appendix C).

Part 2: Measuring fairness

This part lays the theoretical foundations for the following notebook pages. It defines fairness in the machine learning context and distinguishes it from discrimination. Afterwards this part discusses the attempts to measure fairness. It introduces confusion matrices and essential statistical metrics before observational, similarity-based, and causal definitions of fairness are described (see Appendix D).

Part 3: Practical application of fairness metrics

This section of the notebook translates the theoretical concepts from Part 2 into practice. It uses the case study from Section 2.1.4 of the thesis. Students are guided through the code and are asked to manually calculate accuracy, false positive and negative rates, positive predictive value, and statistical parity based on the confusion matrices. These calculation exercises should increase the understanding of metrics and how they are related to the output of a confusion matrix. No automated tools or fairness libraries were used for these steps, to make sure students recognize the underlying ideas of the metric calculations. Each of the six observational fairness metrics from Part 2 is then evaluated in detail. Special attention is given to the concept of calibration, which is less intuitive and not based on the core statistical measures. Different forms of calibration (calibration, group calibration, well-calibration) are explained and compared to prevent confusion. Finally, the Fairlearn ‘ThresholdOptimizer’ is introduced as an exemplary post-processing technique to illustrate the potential and limitations of fairness optimization. Threshold optimization is useful to demonstrate the trade-offs between conflicting fairness definitions. It also shows the students that such technical mitigation strategies exist, even though this thesis and also the notebook do not explore them in more detail (see Appendix E).

Part 4: Bias in machine learning

The fourth part of the notebook provides an overview of the most important types of bias in machine learning. As in this thesis, this section follows the framework proposed by Suresh and Guttag (2021). The focus lies on the stages in the machine learning lifecycle where bias can be introduced. Students are asked to match different bias types to the respective steps in the lifecycle, to test their understanding. The section concludes with a short summary of Hall et al.’s (2022) insights on bias amplification (see Appendix F).

Part 5: Identifying bias in gender classification

After the theoretical consolidation of different forms of bias in Part 4, Part 5 again translates the ideas into a practical scenario. The second case study addresses representational fairness in the context of gender classification. The section starts by clarifying the distinction between decision fairness and representational fairness and explains why context is essential for fairness assessment. As described in Section 2.2.3, three datasets and two models are analyzed. The structure mirrors that of the student dropout study. The analysis is broken into small and understandable steps that allow students to follow the logic behind the fairness evaluation (see Appendix G).

Part 6: Self-fulfilling predictions and feedback loops

This part wants to introduce the students to the concepts of self-fulfilling predictions and feedback loops. After a short theoretical introduction, the third and final case study is presented (see Section 2.3.2). It uses Berlin crime data to demonstrate how predictive policing systems can reinforce existing biases and the interplay between self-fulfilling predictions and feedback loops. The goal is to show how machine learning can create new forms of bias through its influence on human decision-making and to sensitize students for that (see Appendix H).

Part 7: Challenges and recommendations for fairness in machine learning

The final section summarizes the key challenges in developing fair machine learning systems. Although the thesis does not aim to provide concrete solutions in detail, it provides general recommendations. The Z-Inspection® process is introduced as a framework for evaluating machine learning systems in a multidisciplinary and contextual approach. The recommendations serve as a conclusion to the notebook and can also be seen as a good transition to further exploration of fairness in practice (see Appendix I).

The notebook will be submitted together with this thesis, the datasets used, and a separate PDF file containing instructions and solutions. It represents a condensed and interactive version of this thesis, that presents all core findings in a structured and interactive way. The notebook can complement existing teaching materials and can be a valuable addition to university classes. By combining theoretical input, practical exercises, and reflective questions, the notebook enables students to actively question and explore the challenges of fairness in machine learning.

5 Conclusion and Outlook

The aim of this thesis was to develop a way of introducing computer science students to core ethical challenges in machine learning. To achieve this goal an interactive Jupyter notebook was developed that consolidates the content of the thesis. The focus was to raise awareness of fairness issues, introduce foundational concepts and enable students to critically reflect on the social impact of algorithmic decision-making systems. The notebook addresses central concepts such as bias, fairness, and self-fulfilling predictions in an engaging way, supplemented with practical case studies. It emphasizes more established theoretical foundations in the relatively new and therefore still evolving field of fair machine learning. The notebook supports students in developing a conceptual framework that they can apply and expand as technologies and debates about fairness progress.

The three integrated case studies illustrate key tensions. It has been demonstrated by the student dropout study that different fairness metrics can contradict each other and are not universally compatible. With the help of calculation exercises and model evaluation, students experienced trade-offs and learned why fairness in machine learning cannot be reduced to a single number. The gender classification study highlighted that fairness is not only about decisions, but also about representation. This case showed the importance of context, diverse datasets, and the need to critically assess which groups are being misrepresented in datasets and model outputs. The predictive policing study explored the interplay of self-fulfilling predictions and feedback loops. It illustrated how machine learning models can reinforce existing inequalities when they influence the environment from which they learn. This highlights unique dynamics that can appear in machine learning systems.

Together, these studies promote a broader understanding of the sociotechnical nature of fairness and the limitations of purely technical solutions. They also show that good intentions alone are often insufficient, and that even seemingly neutral applications can produce unfair outcomes. This highlights that critical reflection only on model performance is not enough. Data sources, design choices, and the social context in which systems are developed and deployed need to be assessed as well. The developed notebook consists of interactive elements, theoretical explanations, code, and reflection tasks. This encourages active engagement and learning, in contrast to passive consumption of information. The interactive notebook offers a more engaging alternative to purely theoretical materials and can be used in different university classes where students already have basic computer science knowledge.

While the notebook provides a structured introduction to fairness in machine learning, it does not claim to offer exhaustive solutions. Many questions and aspects of fairness are still under debate and therefore not part of this thesis. For example, legal aspects are not included, but regulatory developments such as the EU AI Act (Artificial Intelligence Act, n.d.) begin to take shape and influence future practice. Given the fast-moving and still unsettled research landscape as well as the lack of widely accepted standards, this thesis deliberately avoids deep dives into experimental mitigation techniques or complex causality models. Instead, it provides students with a pragmatic entry point to a highly relevant field. By becoming familiar with foundational concepts and challenges early on, students will be prepared to critically assess algorithmic systems in their future roles or at least recognize when fairness concerns could arise.

Ultimately, fairness in machine learning is not a fixed goal, but an ongoing process. It demands technical understanding, ethical reflection, and interdisciplinary collaboration. This thesis contributes to that process by enabling future practitioners to view fairness assessments not as an irrelevant task, but as an integral part of responsible machine learning development.

References

Alelyani, S. (2021). Detection and evaluation of machine learning bias. *Applied Sciences, 11*(14), 6271. https://doi.org/10.3390/app11146271

Amazon Web Services. (n.d.). *Amazon SageMaker Clarify.* Retrieved May 21, 2025, from https://aws.amazon.com/de/sagemaker-ai/clarify/

Amt für Statistik Berlin-Brandenburg. (2021). *Lebensweltlich orientierte Räume (LOR), Version 2021.* https://www.berlin.de/sen/sbw/stadtdaten/stadtwissen/sozialraumorientierteplanungsgrundlagen/lebensweltlich-orientierte-raeume/

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Artificial Intelligence Act. (n.d.). *The EU Artificial Intelligence Act.* Retrieved May 21, 2025, from https://artificialintelligenceact.eu/

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities.* MIT Press. https://fairmlbook.org/

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy*. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp.149–159). PMLR. https://proceedings.mlr.press/v81/binns18a/binns18a.pdf

Boonstra, M., Bruneault, F., Chakraborty, S., Faber, T., Gallucci, A., Hickman, E., Kema, G., Kim, H., Kooiker, J., Hildt, E., Lamadé, A., Mathez, E. W., Möslein, F., Pathuis, G., Sartor, G., Steege, M., Stocco A., Tadema, W., Tuimala J., … Zicari, R. V. (2024).

Lessons learned in performing a trustworthy AI and fundamental rights assessment [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2404.14366

Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification.* In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Calegari, R., Castané, G. G., Milano, M., & O'Sullivan, B. (2023). Assessing and enforcing fairness in the AI lifecycle. In *Proceedings of the Thirty-Second International Joint Conference in Artificial Intelligence (IJCAI-23)* (pp. 6554–6562). https://doi.org/10.24963/ijcai.2023/735

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153-163. https://doi.org/10.1089/big.2016.0047

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research, 24*, 1–117. http://jmlr.org/papers/v24/22-1511.html

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). https://doi.org/10.1145/2090236.2090255

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). *Runaway feedback loops in predictive policing* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1706.09847

European Commission. (2019). *Ethics guidelines for trustworthy AI.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fairlearn Organization. (n.d). *Fairlearn*. Retrieved May 21, 2025, from https://fairlearn.org/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for datasets* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1803.09010

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies, 156*(1), 33–63. https://www.jstor.org/stable/41487720

Hall, M., Jenni, S., Raji, I. D., & Bethge, M. (2022). *A systematic study of bias amplification* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2201.11706

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics, 24*, 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Kahneman, D. (2012). *Schnelles Denken, langsames Denken* (R. Friedrich & H. Gronemeyer, Trans.). Siedler Verlag. (Original work published in 2011)

Karkkainen, K., & Joo, J. (2021). FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1548–1558). https://doi.org/10.48550/arXiv.1908.04913

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1609.05807

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 4069–4079). https://dl.acm.org/doi/10.5555/3294996.3295162

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society, 16*(3), 252–260.

https://doi.org/10.1108/JICES-06-2018-0056

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 34–42). IEEE. https://talhassner.github.io/home/publication/2015_CVPR/2015_CVPR.pdf

LF AI & Data Foundation. (n.d.). *AI Fairness 360*. Retrieved May 21, 2025, from https://ai-fairness-360.org/

Loftus, J. R., Russel, C., Kusner, M., & Silva, R. (2018). *Causal reasoning for algorithmic fairness* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1805.05859

Lum, K. & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 220–229). https://doi.org/10.1145/3287560.3287596

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), Article e1356.

https://doi.org/10.1002/widm.1356

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*, Article 13. https://doi.org/10.3389/fdata.2019.00013

Osoba, O. A., & Welser, W. IV. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence.* RAND Corporation. https://doi.org/10.7249/RR1744

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Cruz, G. O. R., Peixoto, R. M., Guimarães, G. A. S., Santos, L. L. dos, Araujo, M. M., Cruz, M., Oliveira, E. L. S. de, Winkler, I., & Nascimento, E. G. S. (2022). *Bias and unfairness in machine learning models: A systematic literature review* (Version 4). arXiv. https://doi.org/10.48550/arXiv.2202.08176

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 5680–5689). https://papers.nips.cc/paper_files/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1d7-Abstract.html

Polizei Berlin. (2023). *Kriminalitätsatlas Berlin 2023.* https://www.berlin.de/polizei/service/kriminalitaetsatlas/

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 33–44). https://doi.org/10.1145/3351095.3372873

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data, 7*(11), Article 146. https://doi.org/10.3390/data7110146

Robinson, D., & Koepke, L. (2016). *Stuck in a pattern: Early evidence on "predictive policing" and civil rights.* Upturn. https://www.upturn.org/work/stuck-in-a-pattern

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), Article 58. https://doi.org/10.1145/3392866

Schmidt, J., Pietsch, V., Nocker, M., Radar, M., & Montuoro, A. (2024). Navigating the trade-off between explainability and privacy. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024)* (Vol. 1, pp. 726–733). https://doi.org/10.5220/0012472200003660

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 59–68). https://doi.org/10.1145/3287560.3287598

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Vol. 17, pp. 1–9). https://doi.org/10.1145/3465416.3483305

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1701–1708). IEEE. https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf

UTKFace Dataset. (n.d.). *UTKFace: A large-scale face dataset.* Retrieved May 21, 2025, from

https://susanqq.github.io/UTKFace/

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (pp. 1–7). https://doi.org/10.1145/3194770.3194776

Wachter, S., Mittelstadt, B., & Russel, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, Article 105567. https://doi.org/10.1016/j.clsr.2021.105567

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 5365–5372). University of Hawai´i at Mãnoa. https://scholarspace.manoa.hawaii.edu/items/bc761360-eae84bb7-ba2e4fdd8e9bcd6d

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)* (pp. 1171–1180). https://doi.org/10.1145/3038912.3052660

Zicari, R. V., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Hickman, E., Gallucci, A., Gilbert, T. K., Hagendorff, T., van Halem, I., Hildt, E., Holm, S., Kararigas, G., Kringen, P., Madai, V. I., Mathez, E. W., Tithi, J. J., Vetter, D., Westerlund, M., & Wurth, R. (2022). *How to assess trustworthy AI in practice* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2206.09887

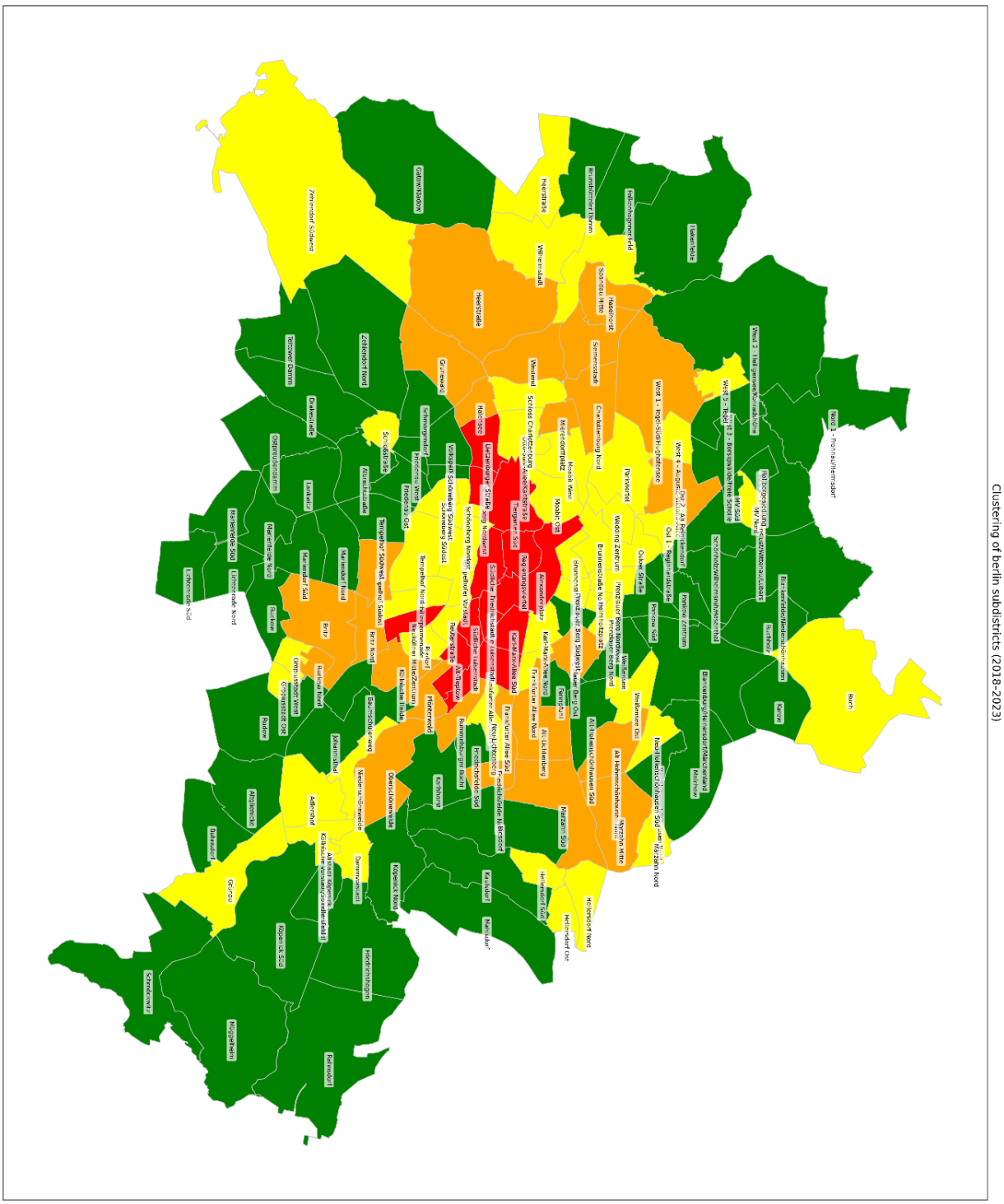
Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., Gerke, S., Gilbert, T. K., Hickman, E., Hildt, E., Holm, S., Kühne, U., Madai, V. I., Osika, W., Spezzatti, A., Schnebel, E., Tithi, J. J., Vetter, D., Westerlund, M., … Kararigas, G. (2021). On assessing trustworthy AI in healthcare: Machine learning as

a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics, 3,* Article 673104. https://doi.org/10.3389/fhumd.2021.673104

Appendix A: Full-Size Version of Figure 3

Figure A1

Enlarged version of Figure 3 from Chapter 2.3.2: Berlin crime hotspots before simulated feedback loop (2018-2023)

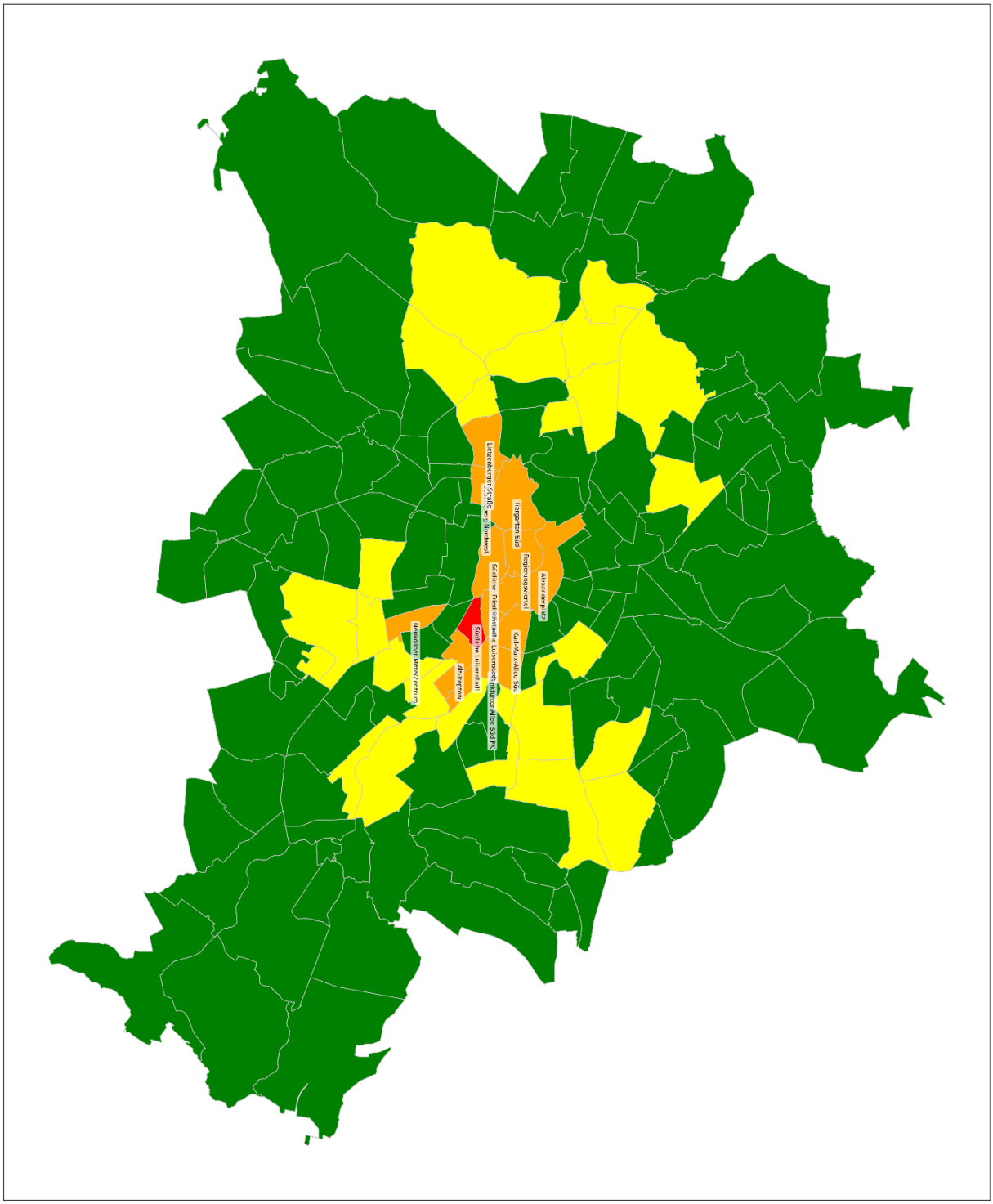


Note. Full notebook, including all steps and additional visualizations, is available in the accompanying Jupyter Notebook on GitHub: https://github.com/LukasWel/ethical-challenges-in-ml

Appendix B: Full-Size Version of Figure 4

Figure B1:

Enlarged version of Figure 4 from Chapter 2.3.2: Berlin crime hotspots after simulated feedback loop



Note. Full notebook, including all steps and additional visualizations, is available in the accompanying Jupyter Notebook on GitHub: https://github.com/LukasWel/ethical-challenges-in-ml

Appendix C: Notebook Part 1

Ethical Challenges in Machine Learning

A practical guide addressing issues such as bias, fairness, and self-fulfilling predictions

This notebook has two main goals:

1. **Raising awareness** of the ethical risks in machine learning (ML) and why fairness must be considered when designing or using ML systems.
2. Demonstrate why **achieving fairness** is complex, often involving trade-offs and difficult decisions.

Structure

1. Motivation
2. Measuring Fairness
3. Practical Application of Fairness Metrics
4. Bias in Machine Learning
5. Identifying Bias in Gender Classification
6. Self-Fulfilling Predictions & Feedback Loops
7. Challenges and Recommendations for Fairness in Machine Learning

---

Part 1: Motivation

Machine learning applications are becoming more important in modern decision-making. They influence everyday tasks like search rankings or personalized recommendations, but also high-stake scenarios with big impact on individuals, such as hiring, healthcare, education or law enforcement. Errors in such domains need to be recognized and mitigated.

> If machine learning models are not carefully designed and monitored, they can **reinforce existing biases**, **contribute to discrimination**, and produce **unfair outcomes** that affect some societal groups more than others.[1]

What makes this even more concerning is that it often happens **unintentionally**.

Before we get deeper into what fairness means and how it can be accessed, we begin with one of the most widely discussed real-world examples of unfairness in algorithmic systems.

Case Study: COMPAS - Risk Prediction in Criminal Justice

Introduction

Imagine you are working for a court. You are introduced to a tool called **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions). It predicts the likelihood that a defendant will commit another crime.

The goal: **help judges make better & fairer decisions.**

Step 1: First Impression

COMPAS uses a questionnaire with **130+ factors** to predict a risk score between 1 (low) and 10 (high).

Here is what the developer claims:[2]

- It **does not use race** as an input.
- It has been **statistically validated for overall predictive accuracy**, meaning that COMPAS risk scores correlate with actual reoffending rates across the entire population.
- It aims to **help reduce human biases** in the justice system.

> **What is a Sensitive Attribute?**[1]
>
> A sensitive attribute relates to protected or vulnerable characteristics of individuals, such as race, gender, age, religion, disability status, or sexual orientation.
>
> Discrimination based directly on sensitive attributes is considered ethically unacceptable and often prohibited by law.

Reflection:

Based on this description, would you trust COMPAS as a fair and objective tool?

Step 2: Exploring the Risk Scores

Below, you see histograms of risk scores from two groups:[2]

- First Black defendants
- Second White defendants

Reflection:

Do you notice any differences between the groups?

Step 3: Investigating Individual Cases

You now review some real examples where COMPAS was used:[2]

Prior Offenses	2 armed robberies, 1 attempted armed robbery	4 juvenile misdemeanors
Subsequent Offenses	1 grand theft	None

Prior Offenses	1 attempted burglary	1 resisting arrest without violence
Subsequent Offenses	3 drug possessions	None

Prior Offenses	1 domestic violence assault, 1 grand theft, 1 petty theft, 1 drug trafficking	1 petty theft
Subsequent Offenses	1 grand theft	None

Reflection:

Do the assigned risk scores match your expectations based on prior and subsequent offenses?

Step 4: Findings from ProPublica Investigation

In 2016 **Angwin et al.** analyzed the fairness of COMPAS and found significant **racial bias** in its predictions. The most important aspects are summarized below (full article).[2]

Key Findings:

	White Defendants	Black Defendants
Predicted High Risk, No Reoffense (False Positive)	23.5%	44.9%
Predicted Low Risk, Reoffended (False Negative)	47.7%	28.0%

- Black defendants were **almost twice as likely** to be incorrectly labeled high-risk compared to white defendants.
- White defendants were **more often** incorrectly labeled low-risk.
- These disparities remained **even after accounting for** prior offenses, age, and gender.
- Although race was **not an explicit input**, bias emerged indirectly through correlated variables such as education, employment, or neighborhood (proxies).

Broader Ethical Issues Identified

- **Seemingly neutral algorithms can reinforce societal inequalities**

  > Even if a model does not use race directly, it can lead to structural disadvantages through proxy variables like income, neighborhood, or education level. This makes it possible for biases to persist invisibly within algorithmic decision-making.

- **Intended fairness is not sufficient**

> Although the system excluded race to reduce bias, it still produced racially biased results. This shows that simply omitting sensitive features does not automatically prevent discrimination when proxy variables exist.

- **Historical data can encode and amplify structural inequalities**

> Machine learning models often learn patterns from past decisions, which may reflect biased practices.

Without critical oversight, such models can replicate or even reinforce these biases.

- **Lack of transparency and explainability**

> As a proprietary system, COMPAS offers no insight into how its risk scores are generated.

This opacity makes it difficult for affected individuals to understand or contest decisions — reducing trust in the system.

- **No clear accountability**

> When algorithmic decisions lead to harmful outcomes, responsibility is often unclear.

Is it the developers, the institutions that deploy the system, or the data providers who should be held accountable? Accountability is often necessary to implement corrective actions.

- **Subtle and invisible bias**

> Algorithmic bias often operates below the surface through indirect correlations and statistical patterns.

Affected individuals often don't realize that they have been treated unfairly and such systems can remain unchallenged for long periods.

Conflict Between Fairness Definitions

The COMPAS case also illustrates that **different definitions of fairness can be in conflict**.[3]

- The company behind COMPAS argued that the tool was **calibrated**: Among individuals with the same risk score, the probability of reoffending was similar across racial groups.
- Angwin et al. (2016) emphasized **unequal error rates**: Black defendants had much higher false positive rates, and white defendants had higher false negative rates — a violation of **equalized odds**.

> An introduction into fairness metrics such as calibration and equalized odds will follow in the next part of this notebook.

Final Reflection

The COMPAS case shows that even when algorithms are intended to be neutral and fair, they can still replicate and even reinforce societal inequalities.

The identified issues highlight ethical challenges that can arise when deploying machine learning systems in high-stakes areas.

> This demonstrates why an **ethical perspective is not optional, but essential** when developing and using machine learning, especially when these systems have a direct impact on people's lives.

Machine learning makes decisions based on **statistical inference**. Algorithmic decisions use **generalizations** and fail to treat people as individuals by design. While such generalizations can be statistically sound and necessary, they can only be morally acceptable if they are sufficiently **accurate** and do **not create systematic disadvantages**.[4]

In the next part of this notebook, we will take a closer look at what fairness in machine learning actually means and why defining fairness is itself a complex task.

Quiz

1. True or False: Excluding sensitive attributes like race guarantees that a machine learning model will be fair.

1. ☐ True
2. ☐ False

2. Which of the following best describes why the COMPAS tool was criticized by Angwin et al.? *(Select one option)*

1. ☐ It was completely inaccurate in predicting any reoffending
2. ☐ It explicitly used race as an input feature
3. ☐ It showed different error rates between racial groups
4. ☐ It was free and open-source, causing legal concerns

3. Which of the following actions would most likely help prevent biased outcomes like those found in the COMPAS case? *(Select one option)*

1. ☐ Removing sensitive attributes from the model
2. ☐ Using more training data, regardless of their correlations
3. ☐ Carefully auditing how features correlate with sensitive attributes
4. ☐ Optimizing the model only for highest predictive accuracy across the entire population

---

Sources:

1. Mehrabi et al., 2021
2. Angwin et al., 2016
3. Barocas et al., 2023
4. Binns, 2018

In [ ]:

Appendix D: Notebook Page 2

Part 2: Measuring Fairness

Fairness and Discrimination

Fairness and discrimination are closely related, but not the same.

> **Fairness** describes the goal of treating individuals and groups equitably in decision-making processes. **Discrimination** refers to a violation of this goal, often resulting in unjust or unequal outcomes[1].

In the context of machine learning, fairness is used to assess whether algorithmic decisions disadvantage certain groups — especially based on sensitive attributes like race, gender, or age. Several factors influence fairness, such as:

- **Transparency**
- **Accountability**
- **Explainability**
- **Bias**

> **Among these, bias plays the most significant role in contributing to discrimination.**

There is **no universal definition of fairness** in machine learning (neither in other disciplines). Competing perspectives exist, often shaped by legal, cultural, or domain-specific goals.[2]

- **Individual fairness**: Similar individuals should be treated similarly.
- **Group fairness**: Different demographic groups should receive similar outcomes.[3]

Each approach has **trade-offs**. What is fair in credit scoring may not be fair in university admissions or criminal justice.[4]

Discrimination in machine learning is different from traditional human discrimination. Algorithms do not have intent or moral awareness. However, they can still produce discriminatory outcomes when trained on biased or unbalanced data. We can distinguish two main types of algorithmic discrimination:[3]

- **Direct discrimination**: Using protected attributes (e.g. gender or race) explicitly in decision-making.
- **Indirect discrimination**: Using seemingly neutral features (e.g. zip code) that correlate with protected attributes and act as proxies.[2]

Discrimination occurs at multiple levels:[1]

- **Structural**: Systemic inequality embedded in laws or history
- **Organizational**: Biased rules or decision processes in institutions
- **Interpersonal**: Individual-level stereotypes or assumptions

Machine learning systems can unintentionally replicate or amplify discrimination from any of these levels. Understanding how fairness and discrimination interact is essential for designing ethical models. The next section introduces common fairness metrics that allow us to detect and evaluate discrimination in practice.

Fairness Metrics

As already said, there is no universal definition of fairness in machine learning. Over the past years, many different fairness metrics have been proposed to evaluate algorithmic decision-making. The large number of fairness definitions can be overwhelming, also because there is no clear consensus on when to use which metric. Simply satisfying as many notions as possible is not an option, as some definitions are **mathematically incompatible**.[5,6]

> **Fairness metrics should be seen as diagnostic tools, not automatic solutions. They can help to identify potential sources of unfairness and discrimination but do not directly fix them.**

Choosing the right metric depends on:

- **The application domain**
- **Ethical priorities**
- **Context-specific trade-offs**

The following section focuses on **observational fairness metrics**. As many of these definitions are derived from the confusion matrix, first a small reminder about the **confusion matrix** and some **core statistical measures**.

The selected fairness metrics presented in this part of the notebook represent the main ideas behind measures of fairness, as many of them are similar in their approach. For a more extensive list of fairness metrics see the paper by Verma & Rubin (2018). Also note that most research on fairness metrics (and also this notebook) focuses on classification algorithms.

Confusion Matrix

The confusion matrix compares the predicted and true class labels. It forms the basis for many fairness and performance metrics:

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Accuracy

Accuracy is the most basic evaluation metric in classification:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

> **Limitation**: Accuracy can be misleading in imbalanced datasets and tells us nothing about fairness across groups.

Core Statistical Measures

These 8 measures are derived from the confusion matrix and form the basis for many fairness definitions:[7]

Measure	Formula	Description
Positive Predictive Value/Precision (PPV)	TP / (TP + FP)	How many predicted positives are correct?
False Discovery Rate (FDR)	FP / (TP + FP)	How many predicted positives are wrong?
Negative Predictive Value (NPV)	TN / (TN + FN)	How many predicted negatives are correct?
False Omission Rate (FOR)	FN / (TN + FN)	How many predicted negatives are wrong?
True Positive Rate/Recall/Sensitivity (TPR)	TP / (TP + FN)	How many actual positives are caught?
False Negative Rate (FNR)	FN / (TP + FN)	How many positives were missed?
False Positive Rate (FPR)	FP / (FP + TN)	How many negatives were wrongly predicted as positive?
True Negative Rate/Specificity (TNR)	TN / (TN + FP)	How many actual negatives are caught?

> These values are computed **per group** (e.g. male/female) to assess fairness.

Observational Fairness

Observational fairness refers to fairness definitions that rely only on observed data — specifically on statistical relationships between:

- $\hat{\mathbf{Y}}$: the model prediction
- **Y**: the ground truth
- **A**: the sensitive attribute (e.g. race, gender)

These definitions **do not require access to causal knowledge** or model internals. They are easy to compute and widely used in fairness audits.

Most observational fairness metrics are based on combinations of the 8 statistical measures from the confusion matrix.

We group them into **three main categories**:[1]

- Independence
- Separation
- Sufficiency

Independence

Requires:

$$\hat{Y} \perp A$$

The predicted outcome should be statistically independent of the sensitive attribute.[1]

This means that all groups should receive positive predictions at equal rates — **regardless of their actual outcome (Y)**.[7]

Common Metric

- **Statistical Parity / Demographic Parity**[7]

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$$

Pros

- Simple and intuitive
- Easy to implement and visualize

Cons

- Ignores true outcome (Y)
- Can result in unfair treatment if base rates differ
- Overlooks explainable or justified outcome differences
- Blind to structural and historical context

> **Example**: Consider a machine learning system that predicts the likelihood of reoffending (such as COMPAS). If men statistically reoffend more often than women, enforcing equal prediction rates across groups could lead to harsher treatment of women without justification.[4]

Separation

Requires:

$$\hat{Y} \perp A \mid Y$$

Given the true label Y, predictions should be independent of A.

This ensures **equal error rates** across groups.[1]

Common Metrics

- **Equalized Odds**:[7]
  Equal FPR and FNR across groups
- **Predictive Equality**:[7] Equal FPR only
- **Equal Opportunity**:[7] Equal FNR only

Pros

- Captures error disparities between groups
- Especially relevant when different types of errors (false positives vs. false negatives) have unequal real-world consequences (e.g. false arrests vs. wrongful releases)

Cons

- May reduce overall accuracy
- Relies on a valid and unbiased ground truth (Y)
- Cannot be satisfied together with other metrics under realistic conditions

> Trade-offs between fairness and performance can be visualized using **ROC curves**. Disparities in error rates often reflect and reinforce historical inequalities.[1]

Sufficiency

Requires:

$$Y \perp A \mid \hat{Y}$$

Given the prediction, the true outcome should be independent of A.

This means that **predictions are equally reliable across groups**.[1]

Common Metrics

- **Calibration Concepts:**[7]
  - **Calibration**: On average, across the whole population, predicted probabilities match actual outcomes.
  - **Group Calibration**:

  Predicted probabilities match actual outcomes **within each demographic group** (e.g. gender, race).

    - **Well-Calibration**: The strongest form — predicted probabilities perfectly match actual outcomes at a fine-grained level **within each group** and for **each score value**.

  In fairness assessments, **group calibration** is mostly used because it checks whether predictions are **equally interpretable across groups**.

  > **Example:** If a model predicts a 70% risk of default, about 70% of individuals assigned a score of 0.7 should actually default — regardless of their group membership.

- **Predictive Parity**:[7]
  Equal PPV across groups

Pros

- Predictions are equally reliable across groups
- Ensures consistent interpretation of predicted scores
- Often achievable without explicit fairness constraints

Cons

- Relies on a valid and unbiased ground truth (Y)
- Can reproduce harmful disparities if the ground truth itself reflects historical bias
- Can conflict with separation-based metrics

> Sufficiency ensures that prediction scores are consistent across groups, but it does not address deeper structural inequalities (but this is true for all statistical fairness definitions).[1]

Incompatibility of Metrics

One challenge in fair machine learning is that **not all fairness metrics can be satisfied at the same time**. In many real-world situations, especially because the sensitive attribute A and the true outcome Y are almost always statistically dependent, fairness definitions make **conflicting assumptions**. Conflicts arise because:

- **Independence** requires predictions to ignore group membership

$$\hat{Y} \perp A$$

- **Separation** requires equal error rates, which depend on group-specific outcome distributions

$$\hat{Y} \perp A \mid Y$$

- **Sufficiency** focuses on reliability of predictions across groups

$$Y \perp A \mid \hat{Y}$$

These conditions can't hold simultaneously unless:[5,6]

- Predictions are **perfect**, or
- The sensitive attribute **has no statistical relationship** with the target variable (which is rare)

Example: **COMPAS**[8]

- The model was **calibrated** → it satisfied sufficiency (PPV equal across groups)
- But it had **unequal error rates** (FPR/FNR differed by race) → it violated separation
- This illustrates how calibration alone cannot guarantee fairness.[1]

Takeaway: Fairness is not one-size-fits-all. Trade-offs between fairness goals are inevitable. Which metric to use depends on context, goals, and ethical priorities.

> **Note:** Observational fairness metrics are useful diagnostic tools to detect statistical disparities between groups.
> However, they only rely on observed relationships between predictions, outcomes, and sensitive attributes — and ignore other relevant features that may contribute to unfairness.
> As a result, they often miss the actual mechanisms behind discrimination. These metrics are blind to structural inequalities, biased decision processes, and causal factors outside the model.
> They can confirm unequal treatment, but not explain *why* it happens.
> **Similarity-based** and **causal fairness** approaches aim to address these limitations by evaluating the decision-making process itself and identifying justified and unjustified sources of disparity.

Similarity-Based Fairness

While observational fairness focuses on statistical group-level patterns in model outcomes, **similarity-based fairness** evaluates the fairness of the **decision process itself** by assessing whether similar individuals receive similar outcomes — regardless of their group membership.

> This reflects the intuitive idea that fairness means treating comparable cases consistently, based on individual characteristics.[2]

Note: Defining what counts as *similar* can itself be subjective and context-dependent.

Key Methods

Causal Discrimination (pairwise test)[7]

- Two individuals who differ **only** in a sensitive feature (e.g. gender) should receive **the same outcome**.
- Captures the idea that sensitive attributes **should not causally influence** decisions.
- Often unrealistic in practice to find two individuals that differ only in one dimension, as sensitive attributes are usually correlated with other features.

Fairness through Unawareness[9]

- Ensures fairness by **removing sensitive attributes** from the data (blinding).
- Based on the logic: "if the model doesn't see it, it can't discriminate."
- Limitation: **proxy variables** can still leak bias (e.g. in the U.S. zip code → race).
- Can lead to **miscalibration** or **reduced accuracy** for some groups if relevant factors are ignored.

Fairness through Awareness[7]

- Uses a **similarity function** (e.g. distance metric) to compare individuals.
- Individuals who are "close" in feature space should be treated similarly.
- Requires including **all relevant attributes**, including sensitive ones.
- More flexible than blinding, but depends on how similarity is defined.

Pros

- Evaluates fairness at the **individual level**, not just between groups
- Focuses on the **decision process**, not only on outcomes
- Allows for **context-sensitive** definitions of fairness
- Bridges the gap between observational and causal fairness approaches

Cons

- Defining **similarity functions** is subjective and challenging
- Sensitive attributes often correlate with other features, complicating comparisons
- Requires **complex feature engineering** and **domain knowledge**

> Similarity-based fairness emphasizes how decisions are made rather than just focusing on the final outcomes. It requires careful thinking about which individuals should be treated similarly.[7]

Causal Fairness

Causal fairness uses **causal models** to understand how variables (e.g. gender, income, education) influence decisions and to **distinguish fair from unfair causal effects**.

> Instead of just asking *who gets what*, causal models ask *why* certain outcomes occur — and whether sensitive attributes **legitimately** influence decisions.

Causal models are represented as **directed graphs** and allow for:[2]

- Reasoning about **hypothetical scenarios**
- Identifying and blocking **unfair influence paths**
- Designing **interventions** to improve fairness

However, modeling causality for **social attributes** (e.g. race, gender) is challenging:[1]

- These categories are **socially constructed**, not fixed
- Meanings vary across time, cultures, and contexts
- Looping effects exist: being labeled can influence future behavior

---

Key Methods

Counterfactual Fairness[9]

- Asks: *Would the decision have been different if the person had belonged to another group when everything else is equal?*
- If the answer is *no* the decision is fair
- Enables **individual-level fairness auditing**
- Requires detailed causal models and well-defined counterfactuals

Path-Specific Fairness[10]

- Recognizes that **some paths** from sensitive attributes to outcomes may be acceptable and others not
- Allows defining which **causal paths are fair**
- Offers a **flexible and context-aware** balance between utility and fairness

---

Pros

- Enables **targeted diagnosis** of unfair influence
- Works even with indirect or subtle bias
- Supports **intervention design**

Cons

- Requires **domain knowledge** and modeling effort
- Difficult to apply when variables are **entangled** or **not clearly defined**

> Causal fairness provides the **most powerful and flexible tools** for fairness analysis, but also the most demanding in terms of assumptions and modeling effort.

The next section provides a practical example showing how different fairness metrics can stand in **conflict** with each other and why **no single metric** is sufficient for evaluating fairness in complex systems. In practice, a combination of observational, similarity-based, and causal fairness assessments can offer a more comprehensive understanding of bias and discrimination in ML systems.

Quiz

1. True or False: If a model satisfies statistical parity, it also guarantees equal error rates across groups.

1. ☐ True
2. ☐ False

2. Which of the fairness metrics requires predictions to be independent of the sensitive attribute, regardless of the true outcome? *(Select one option)*

1. ☐ Equalized Odds
2. ☐ Statistical Parity
3. ☐ Calibration
4. ☐ Predictive Parity

3. Which statement about the incompatibility of fairness metrics is correct? *(Select one option)*

1. ☐ All fairness metrics can usually be satisfied simultaneously in real-world settings
2. ☐ If predictions are imperfect and sensitive attributes influence the outcome, different fairness goals can be in conflict
3. ☐ Independence, Separation, and Sufficiency are not in conflict when sensitive attributes are statistically related to the outcome
4. ☐ Sufficiency and Separation can always be satisfied together if enough data is available

---

Sources:

1. Barocas et al., 2023
2. Mehrabi et al., 2021
3. Calegari et al., 2023
4. Binns, 2018
5. Chouldechova, 2017
6. Kleinberg et al., 2016
7. Verma & Rubin, 2018
8. Angwin et al., 2016
9. Kusner et al., 2017
10. Corbett-Davies et al., 2023

In [ ]:

Appendix E: Notebook Page 3

Part 3: Practical Application of Fairness Metrics

In this notebook section, we apply the fairness metrics you've learned about to a real-world scenario. The goal is to:

- Train a predictive model
- Evaluate its statistical and fairness-related performance
- Apply fairness interventions
- Reflect on trade-offs and challenges in fair machine learning

We use a dataset on student outcomes to predict whether a student will graduate or drop out. A university could use this information and focus on supporting students that are predicted to drop out.

> The potential benefit of such an application is clear - but it also raises ethical concerns:
>
> - If the model **wrongly predicts graduation**, a student in need may receive **no support**.
> - If the model **wrongly predicts dropout**, a student may receive **unnecessary intervention**.
>
> It is essential to consider **which kinds of errors matter more**, and **for whom**. In this case, missing a student who is about to drop out (false negative) may be **more harmful** than offering unnecessary support (false positive).

We will explore how fairness metrics behave in this context and see why **no single metric** is sufficient to assess fairness.

The dataset used in this study is the Student Performance Dataset from the UCI Machine Learning Repository. It contains information about students enrolled in higher education institutions in Portugal. The dataset includes dempgraphic data (e.g. gender, age, marital status), academic performance (e.g. grades, failures), and institutional factors (e.g. scholarship status, application preferences).[1]

First we **import** the necessary libraries. Use pip install for packages you do not have.

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.metrics import confusion_matrix, accuracy_score
from fairlearn.postprocessing import ThresholdOptimizer
```

Load data

Next we load the dataset. **Adjust the path** to where you saved the data. Nationality is removed, because it is strongly imbalanced (most students are from Portugal), providing little analytical value and potentially introducing unwanted bias. To have a binary classification task we only keep `Graduate` and `Dropout` as target variables.

In [2]:
```
# Load dataset
df = pd.read_csv(r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\Student_Study\student_data.csv', sep=';') # Adjust to you
df = df[df['Target'] != 'Enrolled'] # Keep only Graduate and Dropout
df['Target'] = df['Target'].map({'Graduate': 1, 'Dropout': 0}) # Binary mapping
df = df.drop(columns=['Nacionality']) # Drop imbalanced feature
```

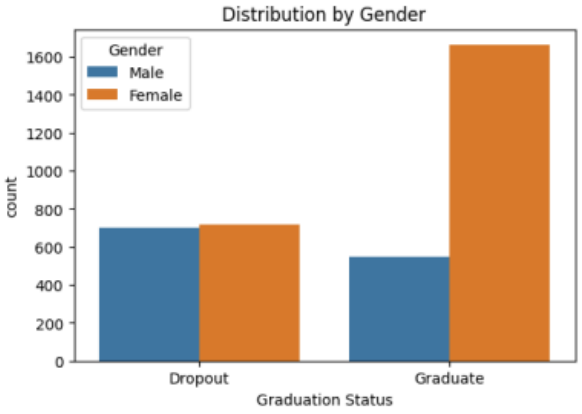
Explore the data

Let's start by examining the distribution of our target variable (`Graduate` vs `Dropout`) and how it differs across gender.

In [3]:
```
# Create copy for vizualization
df_plot = df.copy()
df_plot['Target'] = df_plot['Target'].map({1: 'Graduate', 0: 'Dropout'}) # Mapping to names again
df_plot['Gender'] = df_plot['Gender'].map({1: 'Male', 0: 'Female'})

# Visualization
display(df_plot['Target'].value_counts(normalize=True)) # Display percentage
plt.figure(figsize=(6,4))
sns.countplot(data=df_plot, x='Target', hue='Gender')
plt.title('Distribution by Gender')
plt.xlabel('Graduation Status')
plt.show()
```

```
# Cross-table
counts = pd.crosstab(df_plot['Gender'], df_plot['Target'])
display(counts)
```

```
Target
Graduate    0.60854
Dropout     0.39146
Name: proportion, dtype: float64
```



Target	Dropout	Graduate
Gender		
Female	720	1661
Male	701	548

While dropout rates appear balanced between men and women, graduation rates are significantly higher among women.
This indicates an **unequal base rate** (prevalence) of the positive class (`Graduate`) between groups.

> **Why this matters:**

Unequal base rates are known to cause **incompatibility between different fairness metrics**.

For example, it may become impossible to satisfy **Calibration** and **Equalized Odds** at the same time.[2,3]

This structural imbalance must be considered when evaluating fairness – it does not necessarily indicate bias, but it limits which fairness notions can realistically be satisfied.

Preprocessing & Model Training

We now prepare the data and train a **Random Forest** classifier. We **do include gender as a feature**, so that the model is allowed to treat groups differently. This will allow us to explore fairness conflicts that arise because of group-specific predictions.

In [4]:

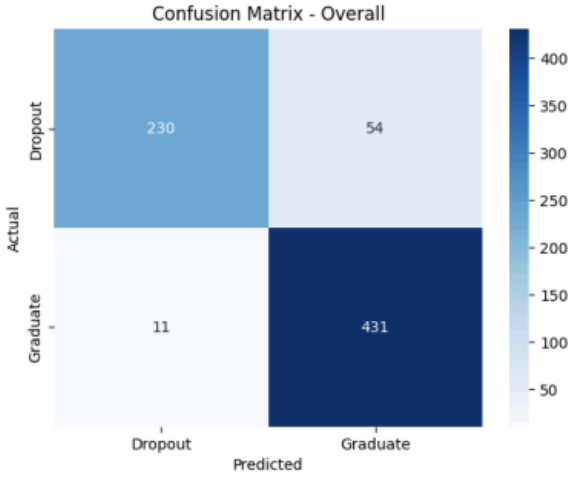
```
# Define features and target variable
X = df.drop(columns=['Target'])
y = df['Target']

# Separate columns by data type
categorical = X.select_dtypes(include='object').columns
numerical = X.select_dtypes(include=['int64', 'float64']).columns

# Preprocessing: scaling for numeric, one-hot encoding for categorical features
preprocessor = ColumnTransformer([
    ('numerical', StandardScaler(), numerical),
    ('categorical', OneHotEncoder(handle_unknown='ignore'), categorical)
])

# 80% training and 20% test set split
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, test_size=0.2, random_state=42)
gender_test = X_test['Gender'] # save gender labels for later group fairness analysis

# Pipeline: Preprocessing + Random Forest
model = Pipeline([
    ('preprocessing', preprocessor),
    ('random_forest', RandomForestClassifier(random_state=42))
```

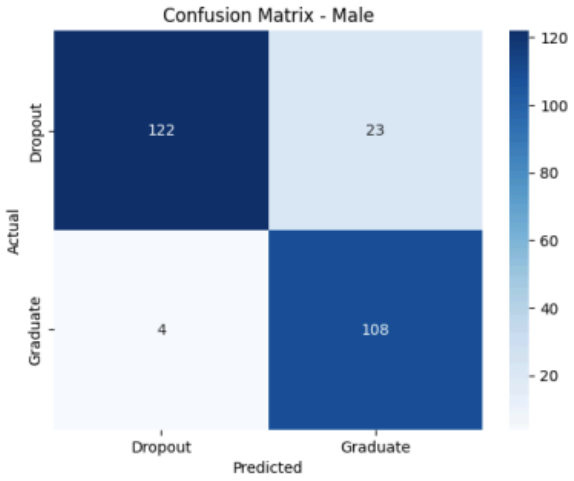
```
])

# Train model and create predictions
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
y_prob = model.predict_proba(X_test)[:, 1] # probability for class 1 (graduate)
```

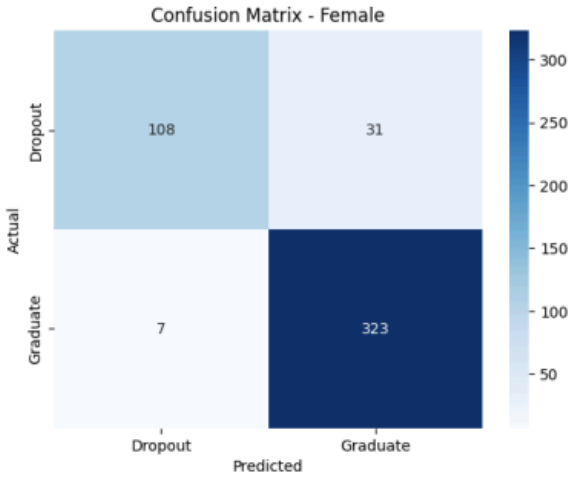
Evaluate Performance with Confusion Matrices

We begin by visualizing the confusion matrix for the entire dataset and then for each gender.

In [5]:
```
# Confusion matrix for all students
cm_all = confusion_matrix(y_test, y_pred)
sns.heatmap(cm_all, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Dropout', 'Graduate'],
            yticklabels=['Dropout', 'Graduate'])
plt.title('Confusion Matrix - Overall')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()

# Confusion matrix for male students
filter_male = gender_test == 1
cm_male = confusion_matrix(y_test[filter_male], y_pred[filter_male])
sns.heatmap(cm_male, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Dropout', 'Graduate'],
            yticklabels=['Dropout', 'Graduate'])
plt.title('Confusion Matrix - Male')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()

# Confusion matrix for female students
filter_female = gender_test == 0
cm_female = confusion_matrix(y_test[filter_female], y_pred[filter_female])
sns.heatmap(cm_female, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Dropout', 'Graduate'],
            yticklabels=['Dropout', 'Graduate'])
plt.title('Confusion Matrix - Female')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```







Accuracy Calculation Exercise

Based on the confusion matrices, try to manually calculate the **overall accuracy** and **accuracy per gender**. After you are done run the code below to compare your results with the printed values.

In [6]:
```
# Compute accuracy
acc_overall = accuracy_score(y_test, y_pred)
acc_male = accuracy_score(y_test[filter_male], y_pred[filter_male])
acc_female = accuracy_score(y_test[filter_female], y_pred[filter_female])
```

Solution Accuracy Calculation

In [7]:
```
print('Overall Accuracy:', acc_overall)
print('Accuracy Male:' , acc_male)
print('Accuracy Female:' , acc_female)
```

```
Overall Accuracy: 0.9104683195592287
Accuracy Male: 0.8949416342412452
Accuracy Female: 0.9189765458422174
```

Core Statistical Measures

Please calculate the following metrics based on the confusion matrices.

False Positive Rate (FPR), **False Negative Rate (FNR)** & **Positive Predictive Value (PPV)**

Do this: For the **entire dataset** & separately for **male** and **female** students. Once you are done, compare your results with the summary table below.

In [8]:
```
# Compute core statistical measures
def core_metrics(y_true, y_pred):
    tn, fp, fn, tp = confusion_matrix(y_true, y_pred).ravel()
    return {
        'TPR': tp / (tp + fn),
        'FPR': fp / (fp + tn),
        'TNR': tn / (fp + tn),
        'FNR': fn / (tp + fn),
        'PPV': tp / (tp + fp),
        'NPV': tn / (tn + fn),
        'FDR': fp / (tp + fp),
        'FOR': fn / (tn + fn)
    }

metrics_all = core_metrics(y_test, y_pred)
metrics_male = core_metrics(y_test[filter_male], y_pred[filter_male])
metrics_female = core_metrics(y_test[filter_female], y_pred[filter_female])
```

Summary Table: Core Statistical Measures

In [9]:
```
# Measures into DataFrame
df_measures = pd.DataFrame(
    [metrics_all, metrics_male, metrics_female],
    index=['Total', 'Male', 'Female']
).round(3)

display(df_measures)
```

	TPR	FPR	TNR	FNR	PPV	NPV	FDR	FOR
Total	0.975	0.190	0.810	0.025	0.889	0.954	0.111	0.046
Male	0.964	0.159	0.841	0.036	0.824	0.968	0.176	0.032
Female	0.979	0.223	0.777	0.021	0.912	0.939	0.088	0.061

Fairness Metrics

We will now compute the 6 observational fairness measures and discuss the results.

Statistical (Demographic) Parity[4]

Goal: The model should assign positive predictions ($\hat{Y} = 1$) equally often to both groups. We calculate the difference in positive prediction rates between male and female students:

$$\text{Statistical Parity Difference} = P(\hat{Y} = 1 \mid \text{Male}) - P(\hat{Y} = 1 \mid \text{Female})$$

The closer this value is to **0**, the more statistically fair the model is in terms of group treatment.

You can compute **Statistical Parity** manually using the group-specific confusion matrices.

Remember:
The number of **positive predictions** in a group equals the sum of **True Positives (TP)** and **False Positives (FP)**.

So for each group:

$$P(\hat{Y} = 1 \mid \text{Group}) = \frac{TP + FP}{TP + FP + TN + FN}$$

Now use the confusion matrices from earlier and calculate statistical parity manually. Once you are done you can run the code below to check your results.

In [10]:
```
# Ground truth + predictions per group
y_true_m = y_test[filter_male]
y_pred_m = y_pred[filter_male]
```

```
y_true_f = y_test[filter_female]
y_pred_f = y_pred[filter_female]

# Compute statistical parity
def compute_statistical_parity(y_true, y_pred):
    tn, fp, fn, tp = confusion_matrix(y_true, y_pred).ravel()
    total = tp + fp + tn + fn
    return (tp + fp) / total


sp_male = compute_statistical_parity(y_true_m, y_pred_m)
sp_female = compute_statistical_parity(y_true_f, y_pred_f)
sp_diff = sp_male - sp_female

print('Proportion of positive predictions for males:', sp_male)
print('Proportion of positive predictions for females:', sp_female)
print('Statistical Parity Difference:', sp_diff)
```

```
Proportion of positive predictions for males: 0.5097276264591439
Proportion of positive predictions for females: 0.7547974413646056
Statistical Parity Difference: -0.2450698149054616 5
```

Interpretation

- **Statistical parity** is violated.
- Female students are **more likely** to receive a positive prediction (`Graduate`) than male students.
- The model assigns **opportunities unequally across groups**, even if those predictions are correct.

Context matters:
In this scenario, a positive prediction does **not** lead to a benefit, but **excludes** students from support programs.
Therefore, male students might actually receive **more interventions** due to the model's behavior — the fairness impact depends on **how the prediction is used**.

Also note:

- The **actual base rates** for graduation **differ significantly** between groups — female students do graduate more often in the data.
- This makes it **difficult** to satisfy statistical parity without distorting reality.

Fairness metrics like Statistical Parity must therefore be interpreted **carefully** and **in context**.

> **Note: How to interpret fairness metric differences**
>
> There is **no universally accepted threshold** that determines when a fairness metric is considered "violated". The values below are **rule-of-thumb** guidelines that can be found in some toolkits and papers:[4]
>
> - **|Difference| < 0.01** → practically identical
> - **0.01 ≤ |Difference| < 0.05** → small difference, often acceptable
> - **0.05 ≤ |Difference| < 0.1** → moderate disparity, may require justification
> - **|Difference| ≥ 0.1** → strong disparity, often considered unfair
>
> These guidelines are not strict rules. Whether a difference is acceptable depends on:
>
> - **Context of the decision** (e.g. education, finance, healthcare)
> - **Scale and consequences** (e.g. minor errors vs. systemic harm)
> - **Baseline risks** (e.g. FNR of 0.02 vs 0.2)
>
> Even small disparities can become **ethically or legally significant** when they reinforce historical disadvantages or affect critical outcomes. Always interpret metric differences **in context** rather than relying purely on numerical thresholds.
>
> For the following analysis, a threshold of 0.05 is used to flag fairness issues, following the paper by Verma & Rubin (2018).

Equal Opportunity[4]

Goal: The model should detect actual positive cases (graduates) equally well for both groups.

This means the **False Negative Rate (FNR)** should be equal across groups:

$$\text{Equal Opportunity Difference} = \text{FNR}_{\text{Male}} - \text{FNR}_{\text{Female}}$$

The closer this value is to **0**, the more fair the model is in terms of *equal opportunity to be correctly identified as positive*. A classifier with equal FNR will mathematically also have equal TPR (since $\text{TPR} = 1 - \text{FNR}$).

From the summary table:

- FNR (Male): 0.036
- FNR (Female): 0.021

→ Difference: **0.015**

False negatives occur slightly more often for male students.
The **difference of 0.015** is **very small** and may be considered **acceptable** under common fairness thresholds (e.g. ±0.05).
However, even small disparities can be relevant depending on the context and scale of decisions.

Predictive Equality[4]

Goal: The model should produce false positives equally often across groups.
This is measured using the **False Positive Rate (FPR)**:

$$\text{Predictive Equality Difference} = \text{FPR}_{\text{Male}} - \text{FPR}_{\text{Female}}$$

From the summary table:

- FPR (Male): 0.159
- FPR (Female): 0.223

→ Difference: **–0.064**

False positives are **more frequent among female students**.
This violates Predictive Equality, since **the model incorrectly predicts Graduate more often for women**.

Equalized Odds[4]

Goal: The model should make both types of errors — false positives and false negatives — **at the same rate across groups**.

We compute:

$$|\text{FNR}_{\text{Male}} - \text{FNR}_{\text{Female}}| + |\text{FPR}_{\text{Male}} - \text{FPR}_{\text{Female}}| = |0.036 - 0.021| + |0.159 - 0.223| = 0.015 + 0.064 = 0.079$$

The model shows a **moderate violation of Equalized Odds**.
Although FNR is relatively balanced, the **unequal false positive rates** are the main driver of this fairness gap.

Predictive Parity[4]

Goal: Among those predicted to graduate ($\hat{Y} = 1$), the probability of actually graduating should be the same across groups.

This is measured by comparing the **Positive Predictive Value (PPV)**:

$$\text{Predictive Parity Difference} = \text{PPV}_{\text{Male}} - \text{PPV}_{\text{Female}}$$

From the summary table:

- PPV (Male): 0.824
- PPV (Female): 0.912

→ Difference: **–0.088**

The model's **positive predictions are more reliable for female students**.
In other words, when the model predicts graduation, it is more often correct for women than for men.

This violates **Predictive Parity** because positive predictions are **not equally trustworthy** across groups.

Group Calibration[4]

Goal: The predicted probability of graduation should have the **same meaning** for each group.

That means: for a given predicted score s, the probability of actual graduation should be the same regardless of group membership:

$$P(Y = 1 \mid \hat{S} = s, \text{Group} = \text{Male}) = P(Y = 1 \mid \hat{S} = s, \text{Group} = \text{Female})$$

> $\hat{S}$ represents the predicted score (graduation probability) output by the model. s is a specific realization of this score for an individual.

This is known as **Group Calibration** or **Test Fairness**.

Fairness Metric:

$$\text{Group Calibration Difference} = \text{Average}_s \left| P(Y = 1 \mid \hat{S} = s, \text{Male}) - P(Y = 1 \mid \hat{S} = s, \text{Female}) \right|$$

The closer this value is to **0**, the better the model is calibrated across groups.

Example:

> Imagine two students – one male, one female – both receive a predicted graduation probability of **0.8**. **Group Calibration** means that for both groups, about **80% of students** with a predicted **score of 0.8** actually graduate.
>
> If, however:
>
> - only **60% of men** with score 0.8 graduate,
> - but **90% of women** with the same score do,
>
> then the model is **not calibrated across groups**. It **overestimates success for men** and **underestimates it for women**.

Related Concepts

- **Calibration**:

$$P(Y = 1 \mid \hat{S} = s) = s$$

→ The predicted score reflects the actual success probability *on average*, across the whole population.

- **Well-Calibration**:

$$P(Y = 1 \mid \hat{S} = s, \text{Group} = \text{Male}) = P(Y = 1 \mid \hat{S} = s, \text{Group} = \text{Female}) = s$$

→ The strongest form: the predicted score equals the actual success rate and is consistent across all groups.

In fairness assessments, **group calibration** is most relevant.
It tells us whether predicted probabilities can be interpreted **consistently and fairly** across different groups.

To better understand the differences between the types of calibration, the following section explains their calculation with an example.

How to compute group calibration[4]

1. **Bin the predicted scores** into intervals (e.g. 0.0–0.1, 0.1–0.2, ..., 0.9–1.0)
2. For each **group** and each **bin**:
   - Compute the **mean predicted probability**
   - Compute the **actual proportion of graduates (Y = 1)**

If the observed graduation rate differs across groups for the same predicted score, **group calibration is violated**.

Example: What Calibration Means in Practice

Let's look at a small group of 10 students – each with a predicted probability and actual graduation outcome:

Student	Gender	Predicted Probability	Actual Outcome
A	Male	0.8	0
B	Male	0.8	1
C	Male	0.8	0
D	Female	0.8	1
E	Female	0.8	1
F	Female	0.8	1
G	Male	0.4	0
H	Female	0.4	1
I	Female	0.4	0
J	Male	0.4	1

Calibration (overall)

Calibration means that across **all students**, the predicted probabilities match the actual outcomes.

- For all students with a score of 0.8 (Bin 0.8):
  - 4 out of 6 students actually graduate → 67% success rate.
- For all students with a score of 0.4 (Bin 0.4):
  - 2 out of 4 students actually graduate → 50% success rate.

Interpretation:

- A score of 0.8 should ideally correspond to an 80% graduation probability.
- Here, it corresponds to 67% → the model is **not calibrated** overall.

Group Calibration

Group Calibration requires that predicted scores have **the same meaning** across groups.

- Among students with score 0.8:
  - **Males:** 1/3 graduated → 33%
  - **Females:** 3/3 graduated → 100%
- Among students with score 0.4:
  - **Males:** 1/2 graduated → 50%
  - **Females:** 1/2 graduated → 50%

Interpretation:

- At 0.8 there is a **large difference** between males and females.
- At 0.4 no group difference appears.
- **However:**
  → **Group Calibration evaluates the overall behavior across all scores.**
  → Because there is a strong mismatch at 0.8, the model is **not group-calibrated**.

Well-Calibration

Well-Calibration is the strictest requirement:
The predicted score must match the actual probability **for each group separately**.

- Males, score 0.8 → predicted 80%, actual 33%
- Females, score 0.8 → predicted 80%, actual 100%
- Males, score 0.4 → predicted 40%, actual 50%
- Females, score 0.4 → predicted 40%, actual 50%

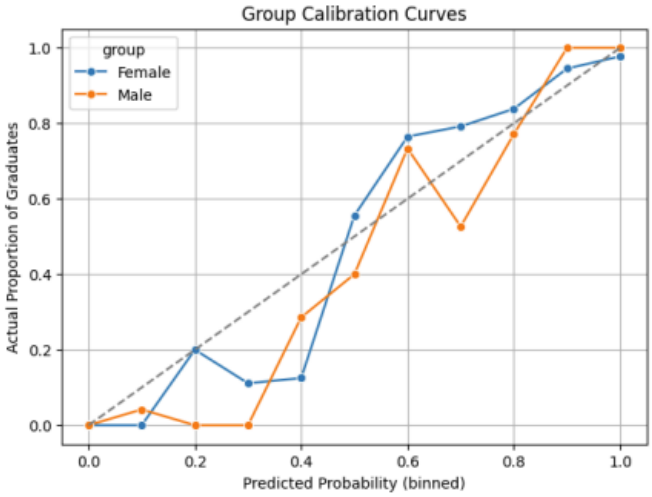
Interpretation:
The model is **not well-calibrated**, because neither group matches the predicted scores exactly.

Group Calibration Analysis

Now let's return to our student dropout prediction model and calculate **group calibration** based on the predicted probabilities.

In [11]:
```
# Create DataFrame with probabilites, true values and group membership
df_calib = pd.DataFrame({
    'prob': y_prob,
    'true': y_test,
    'group': gender_test
})

df_calib['group'] = df_calib['group'].map({0: 'Female', 1: 'Male'}) # Mapping back to names for visualization

# Bin predicted probabilities into intervals
df_calib['bin'] = (df_calib['prob'] * 10).round(0) / 10

# Compute average observed outcomes per bin and group
grouped = df_calib.groupby(['group', 'bin'])['true'].mean().reset_index()

# Visualization
plt.figure(figsize=(7,5))
sns.lineplot(data=grouped, x='bin', y='true', hue='group', marker='o')
plt.plot([0,1], [0,1], linestyle='--', color='gray') # perfect calibration line
plt.title('Group Calibration Curves')
plt.xlabel('Predicted Probability (binned)')
plt.ylabel('Actual Proportion of Graduates')
plt.grid(True)
plt.show()

# Separate male and female graduation rates by bin
calib_male = grouped[grouped['group'] == 'Male'].set_index('bin')['true']
```

```
calib_female = grouped[grouped['group'] == 'Female'].set_index('bin')['true']

# Find bins that exist for both groups
same_bins = calib_male.index.intersection(calib_female.index)

# Compute mean absolute difference across common bins
group_calibration_diff = (calib_male[same_bins] - calib_female[same_bins]).abs().mean()

print('Group Calibration Difference:', round(group_calibration_diff, 3))
```




```
Group Calibration Difference: 0.101
```

Each curve shows the **actual graduation rate** in each score bin, **separately for male and female students**.

- A perfectly calibrated model would lie on the **gray diagonal line**
- Large deviations from this line → the model **over- or underestimates** the true graduation probability
- If the **curves for different groups diverge** → the same predicted score corresponds to **different actual outcomes** across groups

In our case, we observe:

- Systematic differences between the male and female curves
- A **mean absolute difference** in calibration of approximately **0.101**

> **Note:** Since we group predicted probabilities into bins, group calibration is only **approximated**. The finer the binning, the more precise the estimate. In practice, binning is necessary to make calibration measurable.

Conclusion:
The model is **not group-calibrated**.
The same predicted probability **does not mean the same thing** for all students – violating a key fairness criterion.

Summary of Fairness Metrics

Below is a brief overview of how the model performs with respect to six key fairness metrics:

Fairness Metric	Description	Absolute Difference	Fair?
Statistical Parity	Equal share of positive predictions across groups	**0.245**	Violated
Equal Opportunity	Equal False Negative Rate (FNR) across groups	**0.015**	Satisfied
Predictive Equality	Equal False Positive Rate (FPR) across groups	**0.064**	Violated
Equalized Odds	Both FNR and FPR are equal across groups	**0.079**	Violated
Predictive Parity	Equal Positive Predictive Value (PPV) across groups	**0.088**	Violated
Group Calibration	Predicted scores have same meaning across groups	**0.101**	Violated

---

Interpretation

- Only **Equal Opportunity** is approximately satisfied — the model detects actual graduates equally well in both groups.

- All other metrics show violations, especially **Statistical Parity** and **Group Calibration**.
- This illustrates that **not all fairness criteria can be satisfied at the same time** and that model evaluation must be context-aware.

Fairness metrics are diagnostic tools — not moral judgments.
They reveal **patterns** in model behavior, but **only a contextual interpretation** tells us whether those patterns are **unfair**.

For example, if we only looked at **Equal Opportunity**, we might conclude the model is fair.
But a deeper look shows that **positive predictions are not equally meaningful** for male and female students, and that **opportunities are distributed unevenly**.

Fairness assessments must therefore go beyond numbers and ask:
What does this prediction mean — and for whom?

In the next section, we will explore how to mitigate such fairness violations using **threshold adjustments**.

Threshold Adjustment to Enforce Equalized Odds

Now that we have diagnosed fairness violations, we apply the `ThresholdOptimizer` from Fairlearn[5] to mitigate them.

Our goal is to enforce **Equalized Odds**:

- Equal **False Positive Rates (FPR)** across groups
- Equal **False Negative Rates (FNR)** across groups

This is done via **post-processing** – without changing the underlying model – by adjusting decision thresholds *per group* based on fairness constraints.

In [19]:
```python
# Function to compute metrics based on core statistical measures
def compute_fairness(y_true, y_pred, group):
    metrics_male = core_metrics(y_true[group == 1], y_pred[group == 1])
    metrics_female = core_metrics(y_true[group == 0], y_pred[group == 0])

    return {
        'Equalized Odds': abs(metrics_male['TPR'] - metrics_female['TPR']) + abs(metrics_male['FPR'] - metrics_female['FPR
        'Equal Opportunity': abs(metrics_male['TPR'] - metrics_female['TPR']),
        'Predictive Equality': abs(metrics_male['FPR'] - metrics_female['FPR']),
        'Predictive Parity': abs(metrics_male['PPV'] - metrics_female['PPV'])
    }

# ThresholdOptimizer
threshopt = ThresholdOptimizer(
            estimator=model,
            constraints='equalized_odds',
            predict_method='predict_proba',
            prefit=True)

threshopt.fit(X_test, y_test, sensitive_features=gender_test)
y_pred_adj = threshopt.predict(X_test, sensitive_features=gender_test)

# Compute Fairness Metrics
fairness_scores_adj = compute_fairness(y_test, y_pred_adj, gender_test)

# Statistical Parity
# Adjusted predictions per group
y_pred_adj_m = y_pred_adj[filter_male]
y_pred_adj_f = y_pred_adj[filter_female]

# Compute statistical parity with function and ground truth values from before
sp_male_adj = compute_statistical_parity(y_true_m, y_pred_adj_m)
sp_female_adj = compute_statistical_parity(y_true_f, y_pred_adj_f)
sp_diff_adj = sp_male_adj - sp_female_adj

# Accuracy
acc_overall_adj = accuracy_score(y_test, y_pred_adj)
acc_male_adj = accuracy_score(y_test[filter_male], y_pred_adj[filter_male])
acc_female_adj = accuracy_score(y_test[filter_female], y_pred_adj[filter_female])

# Output results
print('Fairness Metrics after ThresholdOptimizer:')
print('Statistical Parity Difference:', round(sp_diff_adj, 3))
print('Equal Opportunity:', round(fairness_scores_adj['Equal Opportunity'], 3))
print('Predictive Equality:', round(fairness_scores_adj['Predictive Equality'], 3))
print('Equalized Odds:', round(fairness_scores_adj['Equalized Odds'], 3))
print('Predictive Parity:', round(fairness_scores_adj['Predictive Parity'], 3))
print('Overall Accuracy:', round(acc_overall_adj, 3))
print('Accuracy Male:', round(acc_male_adj, 3))
print('Accuracy Female:', round(acc_female_adj, 3))
```

```
Fairness Metrics after ThresholdOptimizer:
Statistical Parity Difference: -0.201
Equal Opportunity: 0.003
Predictive Equality: 0.005
Equalized Odds: 0.008
Predictive Parity: 0.142
Overall Accuracy: 0.905
Accuracy Male: 0.868
Accuracy Female: 0.925
```

Summary of Fairness Metrics *after* Threshold Adjustment

Fairness Metric	Abs. Diff. before Treshold Adjustment	Abs. Diff. after Treshold Adjustment	Fair?
Statistical Parity	**0.245**	**0.201**	Violated
Equal Opportunity	**0.015**	**0.003**	Satisfied
Predictive Equality	**0.064**	**0.005**	Satisfied
Equalized Odds	**0.079**	**0.008**	Satisfied
Predictive Parity	**0.088**	**0.142**	Violated
Group Calibration	**0.101**	-	Violated

> Note: The results of the ThresholdOptimizer may vary slightly each time you run this code, even with a fixed model. You will not get the exact same values, but the storyline stays the same.

Why Group Calibration Cannot Be Recomputed After Threshold Adjustment[6]

> The `ThresholdOptimizer` does **not** change the model's predicted probabilities — it only applies **different classification thresholds** for each group to satisfy fairness constraints like Equalized Odds.

As a result:

> - The original predicted scores (`y_prob`) no longer reflect the model's actual behavior.
> - The adjusted predictions (`y_pred_adj`) are binary outcomes, not probabilities.
> - Calibration compares **probabilities** with **actual outcomes** — but after threshold adjustment, the *same score* can lead to *different decisions* depending on group membership.
>
> **Group Calibration is not recomputed**, as the necessary input (group-consistent scores) is no longer valid.

Interpretation

After applying the `ThresholdOptimizer`, we observe:

- **Separation-based metrics** (Equal Opportunity, Predictive Equality, Equalized Odds) are now satisfied - the model treats both groups similarly in terms of error rates.
- **Statistical Parity** remains violated — male and female students still receive positive predictions at different rates.
- **Predictive Parity** worsens — positive predictions are less equally reliable across groups.
- **Group Calibration** is no longer meaningful to evaluate.

This illustrates a **key finding** in fairness research:

> **It is mathematically impossible to satisfy both Equalized Odds and Calibration when base rates differ.[2,3]**

Optimizing one fairness metric **necessarily introduces trade-offs** in others.

In this case, we prioritized **equal error rates** — which improved separation metrics, but sacrificed **score reliability** (Calibration, Predictive Parity).

Final Takeaway

Fairness is not something that can be "fixed" through technical adjustments alone. The same model may appear *fair* under one metric — and *unfair* under another.

In this section, we demonstrated how a model predicting student graduation can be adjusted to satisfy **Equalized Odds** by applying **different thresholds** for male and female students.

This decision improves **error rate parity**:

- **False negatives**, which represent students wrongly assumed to succeed, are now equally distributed — an ethically desirable outcome in our case.

Why this matters:
The goal is to **identify students at risk of dropping out** in order to offer support.
Failing to identify a student at risk (false negative) may result in **denied help**, while wrongly identifying someone (false positive) may *only* result in **unnecessary support**.

In this context, a **false negative** is more harmful than a false positive. Because of that, **Equal Opportunity** — ensuring equal False Negative Rates across groups — can be seen as the most relevant fairness criterion in this study.

> **Note:** From a purely technical perspective, applying the `ThresholdOptimizer` was not necessary — the model already showed **low and balanced false negative rates** between genders. The optimizer was used as an educational tool to:
>
> - Introduce the idea of fairness **interventions**
> - Highlight the **incompatibility** between different fairness metrics (e.g. Separation vs. Sufficiency)
> - Show that such interventions can have **unintended consequences**

Although the separation-based metrics improved, the intervention **worsened calibration and predictive parity** (scores are less reliable). Additionally, the **accuracy** slightly shifted — it increased for female students and decreased for male students. This underlines that **fairness interventions are not neutral**, they reallocate performance and create new trade-offs.

> We accept unequal treatment in scoring to achieve **more equal treatment in access to opportunity**.

Fairness is not a number and fairness metrics are not moral truths — they are diagnostic tools for reflection.

Each metric captures a different notion of fairness.
No model will ever be fair by all definitions at once.

What matters is asking:

- *What is the real-world impact of this model?*
- *Whose outcomes are improved or harmed?*
- *What fairness goal are important in this context?*

Next, we take a closer look at different forms of bias that can be relevant in machine learning systems. Knowing these bias types helps explain why fairness violations arise and how they can be addressed.

Quiz

1. True or False: Applying group-specific thresholds to enforce Equalized Odds can cause the same predicted score to lead to different decisions for different groups.

1. ☐ True
2. ☐ False

2. Why might it be ethically justified to accept group-specific thresholds in the student support example? *(Select one option)*

1. ☐ Because accuracy improves significantly for both groups
2. ☐ Because it ensures both groups have equal predicted scores
3. ☐ Because it equalizes access to support by reducing false negatives
4. ☐ Because statistical parity is the most important fairness criterion

3. What does this notebook illustrate about fairness in machine learning? *(Select one option)*

1. ☐ All fairness metrics should be optimized simultaneously to ensure equity
2. ☐ Fairness is primarily a technical problem that can be solved with enough optimization
3. ☐ Different fairness metrics reflect different values and may conflict with one another
4. ☐ Applying post-processing methods like threshold adjustment guarantees fairness

---

Sources:

1. Realinho et al., 2021
2. Chouldechova, 2017
3. Kleinberg et al., 2016
4. Verma & Rubin, 2018
5. Fairlearn Organization, n.d.
6. Pleiss et al., 2017

In [ ]:

Appendix F: Notebook Page 4

Part 4: Bias in Machine Learning

What is Bias?

Bias refers to systematic **distortions** that lead to **unfair outcomes**. In machine learning, bias often arises **unintentionally** and is closely tied to **discrimination**. Bias can occur at any stage of the ML pipeline and is often introduced through **training data, modeling decisions, or deployment context**.

Bias is not always harmful. However, when it influences decisions about people's lives, we must understand and address it. **Recognizing** and **reducing harmful bias is key** to building fair and trustworthy systems.[1,2]

Types of Bias in Machine Learning

There are many forms of bias that can affect data-driven systems. This notebook focuses on the seven bias categories introduced by **Suresh & Guttag (2021)**, which are especially relevant in the **machine learning lifecycle**. Each category highlights a specific entry point for bias and includes relevant subtypes.

1. Historical Bias[3]

Bias that already exists in the world, **before** any data collection, model training, or algorithmic decision-making takes place. It reflects **structural inequalities and social patterns** that are encoded in the data.

Note: According to Suresh & Guttag (2021), historical bias is the most fundamental form of bias. It includes **many other biases caused by users or society**. The subtypes below are just a small selection of common examples.

- **Subtypes:**[1]
  - *Temporal bias*: outdated data that does not reflect current realities
  - *Content production bias*: some groups produce more or different data (e.g. online content)
  - *Behavioral bias*: different behavior across platforms or contexts
  - *Social bias*: others' behavior influences personal input (e.g. ratings)
  - *Self-selection bias*: participation in data generation is non-random
  - *User interaction bias*: feedback loops reinforce earlier behaviors
- **Example:** Word embeddings link "doctor" to male and "nurse" to female, reflecting societal stereotypes.

---

2. Representation Bias[3]

Occurs when the data underrepresents parts of the target population, leading to models that generalize poorly for these groups.

- **Subtypes:**
  - *Population bias*: target population is misdefined (e.g. based on outdated census data)
  - *Sampling bias*: data collection fails to reflect diversity (e.g. only hospital patients)
  - *Coverage bias*: not all subgroups are equally included
  - *Subset bias*: small groups (e.g. pregnant women) are statistically drowned out
- **Example:** ImageNet contains mostly Western-centric images (45% from the U.S. vs. 1% from China), leading to reduced performance for underrepresented regions. → Problematic when a skewed dataset is used for model training

---

3. Measurement Bias[3]

Bias in how features or labels are defined, collected, or measured. It often stems from inaccurate proxies or unequal label quality across groups.

- **Subtypes:**[1,4]
  - *Label bias*: labels don't reflect ground truth equally (e.g. arrest = crime?)
  - *Omitted variable bias*: important explanatory variables are missing
  - *Instrument bias*: the measurement tool itself performs differently across groups
- **Example:** Using arrest records as a crime proxy leads to inflated risk scores for overpoliced communities. → Higher false positive rates for Black defendants in COMPAS

---

4. Aggregation Bias[3]

Happens when a single model is used across diverse subpopulations and uniform behavior is assumed.

- **Subtype:**[1]
  - *Simpson's Paradox*: trends reverse or disappear when data is aggregated
- **Example:** A model trained on general data misclassifies medical conditions in women because the average values are male-dominated.

---

5. Learning Bias[3]

Bias introduced through modeling decisions, such as optimizing only for overall accuracy.

- **Example:** A model may ignore underrepresented group patterns because they're harder to learn or contribute little to global accuracy.
- **Note:** This often interacts with representation or measurement bias.

---

6. Evaluation Bias[3]

Bias introduced when model evaluation uses benchmarks or metrics that do not reflect the real-world deployment population.

- **Example:** Gender classification performs worst for dark-skinned women due to underrepresentation in benchmark datasets.[5]
- **Impact:** Poor subgroup performance may remain unnoticed if metrics like overall accuracy are used.

---

7. Deployment Bias[3]

When a model is applied in a context that differs from its training or evaluation phase — often without appropriate human oversight.

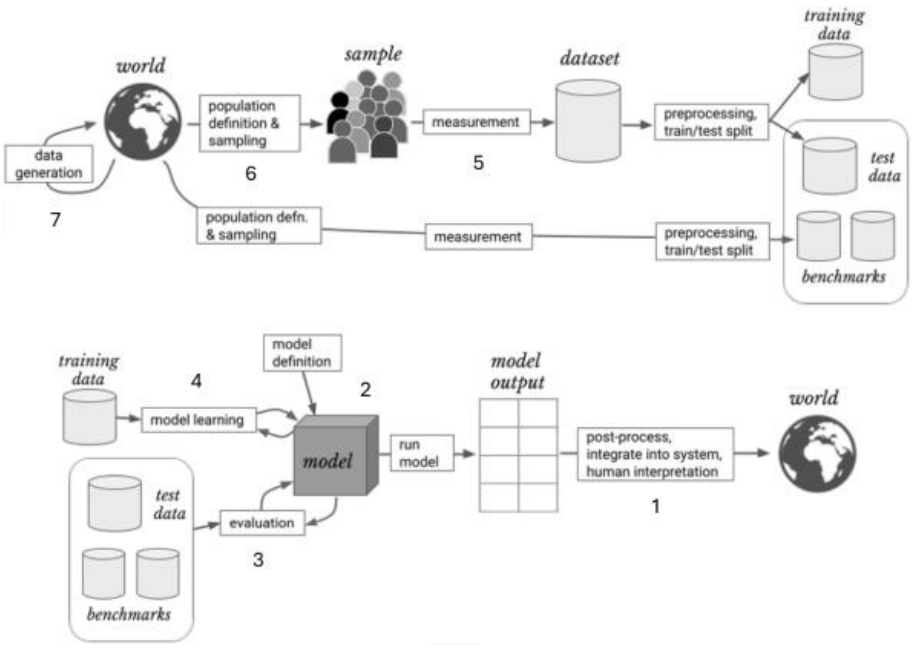
- **Example:** A recidivism prediction model is used to determine prison sentences, but it was designed only to assess risk.
- **Risk:** Even well-performing models can cause harm if deployed carelessly.

Note: These bias types are **not isolated**. They often interact and reinforce each other in **feedback loops**. These dynamics will be discussed on page 6 of this notebook.

Exercise: Localize Forms of Bias in ML-Lifecycle

Below is a simplified version of a figure from Suresh & Guttag (2021). The bias names have been replaced with numbers.

> **Task:** Match each number to the correct bias type. When you are done you can check your answers in the user guide.



Bias Amplification[6]

Bias amplification happens when the model not only reflects existing bias in the data but **intensifies** it — producing more skewed predictions than expected. Research by Hall et al. (2022) identifies several key conditions under which amplification is most likely to occur:

Strong Group Signals

A **group signal** refers to information in the data that reveals **group membership** (e.g. gender or ethnicity) even if this feature is not explicitly included.

- If the **group signal is strong** and easier to learn than the actual label,
  the model may rely on **group-related shortcuts** instead of task-relevant patterns.
- This leads to **overgeneralization** and **amplified disparities** in predictions.
- **Example:** If female students in the data graduate more often, and gender can be inferred from features like *field of study* or *high school GPA*, the model may overuse this information. This results in more graduation predictions just because the student appears to be female.

> **Training dynamics:** Bias amplification is often highest at the beginning of training when models rely on simple group cues. It may decrease mid-training as class-specific patterns emerge, and slightly rise again in late phases.

Model Capacity (V-Shaped Effect)

Model capacity describes a model's ability to learn complex patterns. It is influenced by architecture, depth, regularization, and other hyperparameters.

- **Low-capacity models** underfit and rely on simple features — often group-related ones.
- **High-capacity models** overfit, learning spurious or biased correlations.
- Amplification is highest at both extremes → the **relationship follows a V-shape**.
- Careful tuning and regularization (e.g. weight decay) can help reduce this effect — but often require trading off accuracy.

Data Size and Bias

The structure and size of the training data matter:

- **Small datasets** can lead to overfitting or reliance on noisy group cues.
- **Highly biased datasets** amplify existing imbalances.
- In contrast, **larger and more balanced datasets** reduce amplification risk.

Confidence and Calibration

Poor calibration means the model's prediction confidence does **not match** its actual accuracy.

- Overconfident models — especially on **underrepresented groups** — are more likely to amplify bias.
- A model that makes incorrect predictions with **high confidence** can mask its errors and further reinforce group disparities.
- This effect is particularly common in **high-capacity models**.

Important:
The findings above are based on **binary classification** and **image recognition tasks**.

> **More research is needed** to determine how generalizable these patterns are across domains and applications.

The next section puts the theoretical bias types from this page into practice. We analyze gender classification models to investigate the different forms of bias in a real-world scenario.

Quiz

1. True or False: Deployment bias occurs when the training data is not representative of the target population.

1. ☐ True
2. ☐ False

2. Which of the following is an example of measurement bias? *(Select one option)*

1. ☐ A model trained on mostly Western images underperforms in Asian countries
2. ☐ A recidivism model uses "arrest record" as a label for "criminal behavior"
3. ☐ A model misclassifies women due to male-dominated training data
4. ☐ A model is evaluated only on benchmark datasets that exclude minorities

3. Which statement about bias amplification is correct? *(Select one option)*

1. ☐ It only occurs when the training data is fully biased
2. ☐ It can occur even in small datasets due to overfitting
3. ☐ It always increases with model capacity
4. ☐ It can be avoided by removing all group-related features

---

Sources:

1. Mehrabi et al., 2021
2. Howard & Borenstein, 2018
3. Suresh & Guttag, 2021
4. Corbett-Davies et al., 2023
5. Buolamwini & Gebru, 2018
6. Hall et al., 2022

Appendix G: Notebook Page 5

Part 5: Identifying Bias in Gender Classification

Imagine an **automated HR system** that analyzes application photos to **predict the applicant's gender**. The company uses this information to track diversity statistics or to match candidates with gender-specific support programs.

At first this may seem harmless. However, if the model systematically misclassifies trans individuals or People of Color, they might be excluded from such programs, misrepresented in statistics, or addressed incorrectly.

In this case, it's not about deciding *for* or *against* someone (like in student support models), but about how individuals are **represented** by the system. This leads us to a different fairness concern.

Decision (or Distributive) Fairness vs. Representational Fairness[1]

Most fairness discussions in machine learning focus on **decision fairness**:
Is a model making **equitable decisions** for different demographic groups?

For example:

- Are women and men equally likely to receive a loan if equally qualified?
- Do people of color experience higher false positive rates in recidivism predictions?

In these cases, the **sensitive attribute (e.g., gender, race)** is an **input** — and the model is predicting a **neutral target** (e.g., risk, success, eligibility).

Fairness metrics like **Demographic Parity**, **Equal Opportunity**, or **Equalized Odds** are used to assess whether decisions or errors are **fairly distributed across groups**.

Representational Fairness: A Different Problem[1]

In **representational fairness**, the situation is reversed:

> The **sensitive attribute** (e.g. gender) is not an input — it's the **prediction target**.

The model isn't deciding something *about* a person, it's trying to recognize or classify **who someone is**.

This applies to tasks like:

- Gender classification
- Face recognition
- Emotion detection
- Attribute classification (e.g. age, skin tone)

Here, fairness is about whether the model performs **consistently across demographic groups** — not whether it makes "fair decisions."

How to Evaluate Representational Fairness

We don't optimize for fairness constraints.
Instead, we use **diagnostic performance metrics** to assess **bias in representation**. These metrics (see Table below) help identify **where** the model may fail **for whom**.

Among the observational notions of fairness, only **separation-based metrics** are useful, because they compare **error rates** across groups.
This is exactly what representational fairness is concerned with.

Metric	Purpose
Accuracy per group	Measures overall model performance for each demographic group
False Negative Rate per group (Separation)	Who is most often misclassified as not belonging to their actual group?
False Positive Rate per group (Separation)	Who is most often wrongly assigned to a group?
Equalized Odds (Separation)	Are both FNR and FPR equal across groups?
ROC Curves per group	Shows differences in sensitivity (TPR) and specificity (TNR)
Confusion Matrix per group	Reveals patterns in specific misclassifications

---

What Doesn't Work Well Here?

Some fairness metrics are **not appropriate** for representational fairness problems:

Metric	Why not?
Demographic Parity	Assumes equal predicted class distribution, which is not meaningful if class distributions are unequal or if the target is the sensitive attribute itself
Predictive Parity	Hard to interpret meaningfully when predicting identity attributes
Calibration	Requires probabilistic outputs, which are not available in most classification APIs like DeepFace or OpenCV

Final Takeaway

> **Representational fairness isn't about who gets what — it's about who gets seen, and how.**
> Fairness here means that a model recognizes people from all groups **equally well**, not that it distributes opportunities or resources.

Datasets to Evaluate Model Performance

To examine fairness in gender classification, we begin by comparing two datasets:

- **FairFace**: A dataset with a **balanced distribution** across **ethnicity and gender**, curated specifically for fairness research.[2]
- **UTKFace**: A **realistic but more imbalanced dataset**, containing a wide variety of ages, image qualities, and labeling inconsistencies.[3]

This initial comparison allows us to identify **disparities in model performance** and understand how **representation and data quality** influence results.

Later, we introduce a **manually curated subset** of UTKFace to simulate a controlled real-world scenario.

First we import the necessary libraries. Use pip install for packages you do not have.

In [1]:
```
import os
import cv2
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from deepface import DeepFace
from tqdm import tqdm
from sklearn.metrics import accuracy_score, confusion_matrix
```

```
WARNING:tensorflow:From C:\Users\lukas\anaconda3\envs\Uni\Lib\site-packages\tf_keras\src\losses.py:2976: The name tf.losse
s.sparse_softmax_cross_entropy is deprecated. Please use tf.compat.v1.losses.sparse_softmax_cross_entropy instead.
```

Load Data

Next we load and prepare the two datasets used for fairness evaluation. Each image is labeled with gender and race to allow subgroup-level performance analysis later.

Make sure to **adjust the folder paths** to the storage location on your machine.
If the image processing takes too long on your machine, consider **reducing the number of samples per group** to speed things up.

In [2]:
```
# Image Folders
images_fairface = r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\GenderClassification_Study\fairface-img-margin025-trainv
metadata_fairface = r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\GenderClassification_Study\fairface_label_val.csv' # A

images_utk = r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\GenderClassification_Study\UTKFace' # Adjust to your path

# Load FairFace-Metadata
df_fairface = pd.read_csv(metadata_fairface)

# Adjust column names
df_fairface = df_fairface.rename(columns={'file': 'image', 'gender': 'true_gender', 'race': 'true_race'})

# DeepFace expects "Man" or "Woman" so change "Male" to "Man" and "Female" to "Woman"
df_fairface['true_gender'] = df_fairface['true_gender'].map({'Male': 'Man', 'Female': 'Woman'})

# Complete image paths
df_fairface['image_path'] = df_fairface['image'].apply(lambda x: os.path.join(images_fairface, x.replace('val/', '')))
```

```
print('FairFace:', df_fairface['true_race'].value_counts())

# Sample 100 images per ethnicity
df_fairface = df_fairface.groupby('true_race').sample(n=100, random_state=42).reset_index(drop=True)


# Extract UTKFace ethnicity and gender labels from filenames
utk_data = []
for filename in os.listdir(images_utk):
    # UTKFace filenames have format: age_gender_race_(...).jpg
    name_parts = filename.split("_")
    # Make sure filename has age, gender, and race information
    if len(name_parts) >= 3:
        gender = int(name_parts[1]) # 0 = Man, 1 = Woman
        race = int(name_parts[2]) # 0 = White, 1 = Black, 2 = Asian, 3 = Indian
        full_path = os.path.join(images_utk, filename)
        utk_data.append([full_path, gender, race])

# Create DataFrame
df_utk = pd.DataFrame(utk_data, columns=['image_path', 'true_gender', 'true_race'])

# Assign labels
ethnicity_mapping = {0: 'White', 1: 'Black', 2: 'Asian', 3: 'Indian'}
gender_mapping = {0: 'Man', 1: 'Woman'}

df_utk['true_race'] = df_utk['true_race'].map(ethnicity_mapping)
df_utk['true_gender'] = df_utk['true_gender'].map(gender_mapping)
df_utk['ethnicity_gender'] = df_utk['true_race'] + ' ' + df_utk['true_gender']

print('UTKFace:', df_utk['true_race'].value_counts())

# Sample 200 images per ethnicity
df_utk = df_utk.groupby('true_race').sample(n=200, random_state=42).reset_index(drop=True)

# Overview
print('FairFace images loaded:', len(df_fairface))
print('UTKFace images loaded:', len(df_utk))
```

```
FairFace: true_race
White              2085
Latino_Hispanic    1623
Black              1556
East Asian         1550
Indian             1516
Southeast Asian    1415
Middle Eastern     1209
Name: count, dtype: int64
UTKFace: true_race
White     10078
Black      4526
Indian     3975
Asian      3434
Name: count, dtype: int64
FairFace images loaded: 700
UTKFace images loaded: 800
```

Model 1: DeepFace Gender Classification[4]

We now analyze gender classification performance using the **DeepFace** library.

For each of the three datasets, we evaluate the model across several dimensions:

- **Confusion matrices**
- **Overall accuracy**
- **Accuracy per gender and per ethnicity**
- **Accuracy by ethnicity × gender subgroup**
- **False Positive Rate** and **False Negative Rate** per group

This allows us to assess **representational fairness** by measuring whether DeepFace performs consistently across demographic groups or whether some individuals are systematically misclassified.

We start with applying the model to both datasets. Depending on your system this can take up to 10 minutes for each dataset. You can reduce the number of samples above to reduce the time.

In [3]:

```
def predict_gender_deepface(df, image_col='image_path'):
    predictions = []

    # Iterate over each image in the dataset with a progress bar (tqdm)
    for _, row in tqdm(df.iterrows(), total=len(df), desc='DeepFace Prediction'):
```

```
        try:
            # Analyze the image using DeepFace and extract the predicted gender
            result = DeepFace.analyze(row[image_col], actions=['gender'], enforce_detection=False)
            predictions.append(result[0]['dominant_gender'])
        except:
            # If analysis fails (e.g. no face detected) append None
            predictions.append(None)

    # Add predictions to the DataFrame
    df['predicted_gender'] = predictions

    # Drop rows where prediction failed
    return df.dropna(subset=['predicted_gender'])

# Apply model to both datasets
df_fairface = predict_gender_deepface(df_fairface)
df_utk = predict_gender_deepface(df_utk)
```

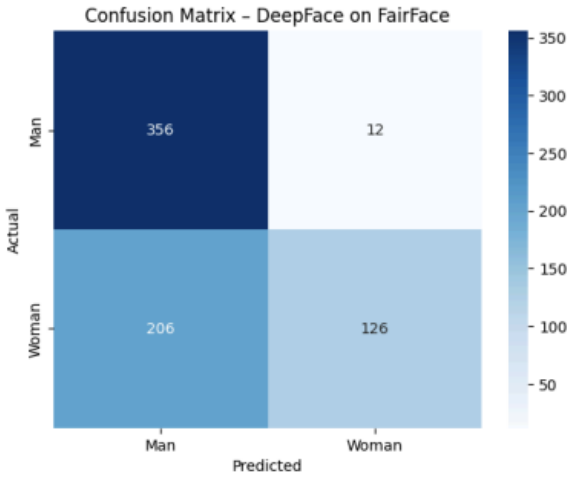
```
DeepFace Prediction: 100%|██████████| 700/700 [03:03<00:00,  3.80it/s]
DeepFace Prediction: 100%|██████████| 800/800 [03:36<00:00,  3.70it/s]
```

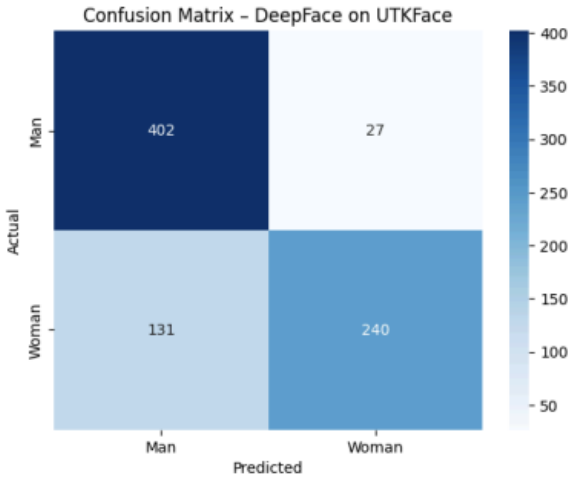
Confusion Matrices

Next we evaluate performance with the help of confusion matrices.

In [4]:

```
def plot_confusion_matrix(df, title):
    cm = confusion_matrix(df['true_gender'], df['predicted_gender'], labels=['Man', 'Woman'])
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
                xticklabels=['Man', 'Woman'],
                yticklabels=['Man', 'Woman'])
    plt.title(title)
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()

plot_confusion_matrix(df_fairface, title='Confusion Matrix - DeepFace on FairFace')
plot_confusion_matrix(df_utk, title='Confusion Matrix - DeepFace on UTKFace')
```





First Observations from Confusion Matrices

Both datasets show a similar pattern in DeepFace's gender predictions:

- **Women (positive class)** are more often **misclassified** than men.
- The number of **false negatives (FN)** (women predicted as men) is way **higher** than false positives (FP).

This suggests that the model struggles more to correctly identify women than men. In the FairFace dataset this imbalance seems the be even stronger.

These inital findings raise concerns about **representational fairness** and will now be explored in more detail using **accuracy, error rates and intersectional evaluation**.

In [5]:

```python
def evaluate_gender_classification(df, dataset_name):
    print('Evaluation Results for', dataset_name)

    # Overall Accuracy
    acc_overall = accuracy_score(df['true_gender'], df['predicted_gender'])
    print('Overall Accuracy:', round(acc_overall, 3))

    # Mark correct predictions
    df['correct'] = df['true_gender'] == df['predicted_gender']

    # Accuracy by Gender
    acc_gender = df.groupby('true_gender')['correct'].mean()
    print('\nAccuracy by Gender:')
    print(round(acc_gender, 3))

    # Accuracy by Ethnicity
    acc_ethnicity = df.groupby('true_race')['correct'].mean()
    print('\nAccuracy by Ethnicity:')
    print(round(acc_ethnicity, 3))

    # Accuracy by Ethnicity × Gender
    acc_ethn_gend = df.groupby(['true_race', 'true_gender'])['correct'].mean()
    print('\nAccuracy by Ethnicity × Gender:')
    print(round(acc_ethn_gend, 3))

    # Convert to binary to compute FNR & FPR
    df_bin = df.copy()
    df_bin['true_gender_bin'] = df_bin['true_gender'].map({'Man': 0, 'Woman': 1})
    df_bin['predicted_gender_bin'] = df_bin['predicted_gender'].map({'Man': 0, 'Woman': 1})

     # Compute FNR & FPR
    def compute_fnr_fpr(sub_df):
        tn, fp, fn, tp = confusion_matrix(sub_df['true_gender_bin'], sub_df['predicted_gender_bin']).ravel()
        fnr = fn / (tp + fn)
        fpr = fp / (fp + tn)
        return pd.Series({'FNR': round(fnr, 3), 'FPR': round(fpr, 3)})

    fnr_fpr_by_ethnicity = df_bin.groupby('true_race').apply(compute_fnr_fpr)

    print('\nFNR & FPR by Ethnicity:')
```

```
    print(fnr_fpr_by_ethnicity, '\n')

# Apply evaluation funtion to datasets
evaluate_gender_classification(df_fairface, dataset_name='FairFace')
evaluate_gender_classification(df_utk, dataset_name='UTKFace')
```

```
Evaluation Results for FairFace
Overall Accuracy: 0.689

Accuracy by Gender:
true_gender
Man      0.967
Woman    0.380
Name: correct, dtype: float64

Accuracy by Ethnicity:
true_race
Black              0.56
East Asian         0.75
Indian             0.65
Latino_Hispanic    0.72
Middle Eastern     0.81
Southeast Asian    0.59
White              0.74
Name: correct, dtype: float64

Accuracy by Ethnicity × Gender:
true_race        true_gender
Black            Man            0.941
                 Woman          0.163
East Asian       Man            0.918
                 Woman          0.588
Indian           Man            0.962
                 Woman          0.312
Latino_Hispanic  Man            0.939
                 Woman          0.510
Middle Eastern   Man            1.000
                 Woman          0.406
Southeast Asian  Man            1.000
                 Woman          0.212
White            Man            1.000
                 Woman          0.469
Name: correct, dtype: float64

FNR & FPR by Ethnicity:
                   FNR    FPR
true_race
Black            0.837  0.059
East Asian       0.412  0.082
Indian           0.688  0.038
Latino_Hispanic  0.490  0.061
Middle Eastern   0.594  0.000
Southeast Asian  0.788  0.000
White            0.531  0.000

Evaluation Results for UTKFace
Overall Accuracy: 0.802

Accuracy by Gender:
true_gender
Man      0.937
Woman    0.647
Name: correct, dtype: float64

Accuracy by Ethnicity:
true_race
Asian     0.765
Black     0.755
Indian    0.830
White     0.860
Name: correct, dtype: float64

Accuracy by Ethnicity × Gender:
true_race  true_gender
Asian      Man            0.896
           Woman          0.644
Black      Man            0.981
           Woman          0.510
Indian     Man            0.937
           Woman          0.697
White      Man            0.932
           Woman          0.756
Name: correct, dtype: float64

FNR & FPR by Ethnicity:
              FNR    FPR
true_race
Asian       0.356  0.104
Black       0.490  0.019
Indian      0.303  0.063
White       0.244  0.068
```

Interpretation: DeepFace Performance on FairFace vs. UTKFace

We applied DeepFace to two datasets with different characteristics and found **consistent bias**, but also **important differences**:

Overall Accuracy

- UTKFace achieves **higher overall accuracy (80.2%)** compared to FairFace (68.9%)
- However, this is **not necessarily a sign of better fairness**

Gender Disparities

- In both datasets, the model performs **significantly better for men**:
  - **FairFace**: Men = 96.7%, Women = **38%**
  - **UTKFace**: Men = 93.7%, Women = **64.7%**
- FairFace: The model misclassifies **more than 60% of women**

Ethnic & Intersectional Patterns

- In **FairFace**, some of the worst-performing subgroups include:
  - **Black Women** → 16.3% accuracy
  - **Southeast Asian Women** → 21.2%
  - **Indian Women** → 31.2%
- In contrast, **White Men consistently achieve 100% accuracy**
- **False Negative Rates** (women being classified as men) are alarmingly high for Black (83.7%), Southeast Asian (78.8%), and Indian women (68.8%)

UTKFace Comparison

- While UTKFace shows **better overall performance**, the patterns remain:
  - Accuracy is still highest for **men**
  - Accuracy for **Black women** is again low (51.0%)
  - FNRs for women are lower than in FairFace, but **still substantial**

Takeaway

DeepFace works almost perfectly for **dominant groups** like White Men, but fails dramatically for **marginalized subgroups** — particularly **Women of Color**.

> This is a clear example of **representational unfairness**:
> The model doesn't make unfair decisions — it **fails to recognize certain people at all**.

Why We Need a Clean Subset

These results highlight a key challenge:

> Model performance is not just a function of architecture — it is heavily influenced by **data quality**, **group representation**, and **label noise**.

Both datasets suffer from issues that may **exaggerate** fairness problems:

- Uncontrolled variations in pose, lighting, and expression
- Unequal representation of age, ethnicity, and gender combinations
- Mislabeled or ambiguous images

Motivation: A Realistic Evaluation Scenario

To isolate the model's behavior from such noise, we introduce a **manually curated subset** of UTKFace → UTKFace adjusted.
This subset better reflects the kind of conditions an **automated HR tool** might operate under:

- Frontal, well-lit images
- Neutral expressions
- Balanced distribution of gender and ethnicity
- Age range limited to 16–70 years

This allows us to answer a critical question:

> **Can we still observe representational bias under more controlled and "realistic" conditions?**

We will now load the adjusted dataset and look at the results.

In [6]:
```
# Image Folder
images_utk_adj = r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\GenderClassification_Study\UTKFace_Adjusted' # Adjust to

# Extract UTKFace ethnicity and gender labels from filenames
utk_data_adj = []
for filename in os.listdir(images_utk_adj):
    name_parts = filename.split("_")
    if len(name_parts) >= 3:
        gender = int(name_parts[1])
        race = int(name_parts[2])
        full_path = os.path.join(images_utk_adj, filename)
        utk_data_adj.append([full_path, gender, race])

# Create DataFrame
df_utk_adj = pd.DataFrame(utk_data_adj, columns=['image_path', 'true_gender', 'true_race'])

# Assign Labels
df_utk_adj['true_race'] = df_utk_adj['true_race'].map(ethnicity_mapping)
df_utk_adj['true_gender'] = df_utk_adj['true_gender'].map(gender_mapping)
df_utk_adj['ethnicity_gender'] = df_utk_adj['true_race'] + ' ' + df_utk_adj['true_gender']

print('UTKFaceAdjusted-Bilder geladen:', len(df_utk_adj))
```

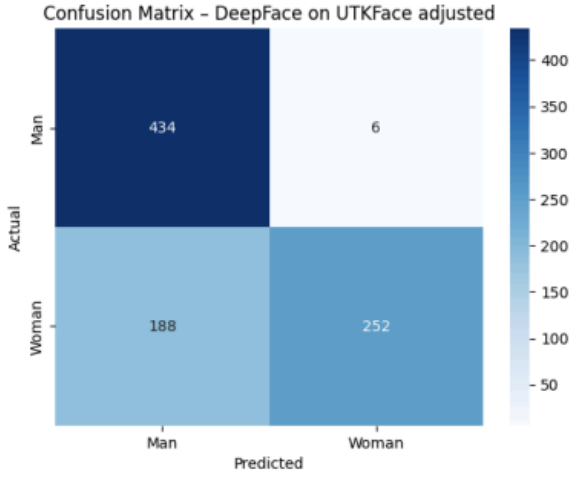
```
UTKFaceAdjusted-Bilder geladen: 880
```

Evaluation UTKFace adjusted

We now apply the model and use our functions to compute the **confusion matrix** and **performance metrics**.

In [7]:
```
# Apply model
df_utk_adj = predict_gender_deepface(df_utk_adj)

# Confusion Matrix
plot_confusion_matrix(df_utk_adj, title='Confusion Matrix – DeepFace on UTKFace adjusted')

# Evaluation
evaluate_gender_classification(df_utk_adj, dataset_name='UTKFace adjusted')
```

```
DeepFace Prediction: 100%|██████████| 880/880 [03:55<00:00,  3.73it/s]
```



```
Evaluation Results for UTKFace adjusted
Overall Accuracy: 0.78

Accuracy by Gender:
true_gender
Man      0.986
Woman    0.573
Name: correct, dtype: float64

Accuracy by Ethnicity:
true_race
Asian     0.773
Black     0.682
Indian    0.720
White     0.941
Name: correct, dtype: float64

Accuracy by Ethnicity × Gender:
true_race  true_gender
Asian      Man            0.955
           Woman          0.591
Black      Man            1.000
           Woman          0.364
Indian     Man            1.000
           Woman          0.445
White      Man            0.991
           Woman          0.891
Name: correct, dtype: float64

FNR & FPR by Ethnicity:
             FNR    FPR
true_race
Asian      0.409  0.045
Black      0.636  0.000
Indian     0.555  0.000
White      0.109  0.009
```

Bias Is Structural

Even under **controlled conditions** (frontal, high-quality, balanced images) DeepFace continues to show **strong disparities** in performance across demographic groups.

Key Findings

- **White Women** benefit from improved data: Accuracy rises to **89.1%**
- But **Black and Indian Women** still suffer from **major recognition failures**
- **The gender gap is smaller** for White individuals, but persists for others

This suggests that:

> **The bias is not just due to bad data — it's embedded in the model's behavior.**

Group Comparison Table

Metric	FairFace	UTKFace	UTK Adjusted
Accuracy (Women overall)	38%	64.7%	57.3%
Accuracy (Black Women)	16.3%	51.0%	36.4%
Accuracy (White Women)	46.9%	75.6%	89.1%

This reinforces the importance of:

- **Intersectional evaluation**
- **Group-level fairness diagnostics**
- And the need for **representational fairness** — not just high accuracy

---

Model 2: OpenCV Gender Classification[5]

In this second model, we evaluate gender classification using a pre-trained **OpenCV model**.
This model is less complex than DeepFace and was trained on a different dataset (Adience). It serves as a useful **baseline** to compare performance and bias.

We apply it to the **adjusted UTKFace subset**, using the same evaluation procedure as before. This allows us to find out whether the observed bias patterns are specific to DeepFace or **persist across different model architectures**.

In [9]:

```python
# Load OpenCV Gender Classification Model, adjust to your paths
gender_net = cv2.dnn.readNetFromCaffe(
    r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\GenderClassification_Study\OpenCV_Model\deploy_gender.prototxt', # Adj
    r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\GenderClassification_Study\OpenCV_Model\gender_net.caffemodel'   # Adj
)

# Mapping of OpenCV classes
GENDER_LIST = ['Man', 'Woman']

#
def predict_gender_opencv(df, image_col='image_path'):
    predictions = []

    # Iterate over each image in the dataset with a progress bar (tqdm)
    for _, row in tqdm(df.iterrows(), total=len(df), desc='OpenCV Gender Prediction'):
        try:
            # Load the image
            image = cv2.imread(row[image_col])
            if image is None:
                predictions.append(None)
                continue

            # OpenCV expects 227x227 BGR-pictures with specific mean subtraction
            blob = cv2.dnn.blobFromImage(
                image,
                scalefactor=1.0,
                size=(227, 227),
                mean=(78.426, 87.768, 114.895), # BGR mean values from ImageNet - make sure that input format matches the
                swapRB=False, # No transformation to RGB
                crop=False
            )

            # Run prediction using the gender model
            gender_net.setInput(blob)
            preds = gender_net.forward()
            predictions.append(GENDER_LIST[preds[0].argmax()]) # "Man" or "Woman"

        except Exception:
            predictions.append(None)

    # Add predictions to DataFrame and drop failed entries
    df['predicted_gender'] = predictions
    return df.dropna(subset=['predicted_gender'])

# Apply model on copy of utk_adj
df_utk_adj_opencv = predict_gender_opencv(df_utk_adj.copy())

# Evaluation
evaluate_gender_classification(df_utk_adj_opencv, dataset_name='UTK adjusted (OpenCV)')
```

```
OpenCV Gender Prediction: 100%|██████████| 880/880 [00:10<00:00, 83.61it/s]
```

```
Evaluation Results for UTK adjusted (OpenCV)
Overall Accuracy: 0.85

Accuracy by Gender:
true_gender
Man      0.945
Woman    0.755
Name: correct, dtype: float64

Accuracy by Ethnicity:
true_race
Asian     0.814
Black     0.777
Indian    0.876
White     0.932
Name: correct, dtype: float64

Accuracy by Ethnicity × Gender:
true_race  true_gender
Asian      Man            0.927
           Woman          0.700
Black      Man            0.964
           Woman          0.591
Indian     Man            0.954
           Woman          0.800
White      Man            0.936
           Woman          0.927
Name: correct, dtype: float64

FNR & FPR by Ethnicity:
             FNR    FPR
true_race
Asian      0.300  0.073
Black      0.409  0.036
Indian     0.200  0.046
White      0.073  0.064
```

Model 2 (OpenCV): Improved Performance, Persistent Bias

The OpenCV model achieves **higher overall accuracy (85.0%)** than DeepFace on the same clean dataset and shows **less extreme disparities** across demographic groups.

Key Improvements Compared to DeepFace:

- **Women overall**: 75.5% (vs. 57.3% with DeepFace)
- **Black Women**: 59.1% (vs. 36.4%)
- **Indian Women**: 80% (vs. 44.5%)
- **White Women**: 92.7% (vs. 89.1%)

Interpretation:

- The **gender gap** is smaller, especially for **White and Indian women**
- **False Negative Rates** (misclassifying women) are lower across all groups
- Overall, the OpenCV model is **less biased**, especially in intersectional performance

---

But Bias Still Exists

Despite improvements, **representational unfairness remains**:

- Accuracy for **Black Women** is still lowest (59.1%)
- FNR for **Black Women** remains relatively high (40.9%)
- Model still performs **best for White individuals** across the board

This shows that:

> Even simpler models trained on different data show similar patterns —
> suggesting that **bias is a broader, structural issue**, not just specific to DeepFace.

Summary Table: DeepFace vs. OpenCV on Clean Subset

Metric	DeepFace	OpenCV
Accuracy (Women overall)	57.3%	75.5%
Accuracy (Black Women)	36.4%	59.1%

Metric	DeepFace	OpenCV
Accuracy (White Women)	89.1%	92.7%

While OpenCV clearly improves recognition for marginalized groups,
performance gaps persist — particularly for Women of Color.

Bias is reduced, but not resolved.

Reflection:

Would you use one of the models for the automated HR tool, that sorts applicant photos per gender?

Chapter Summary: Bias in Gender Classification

In this chapter, we examined how bias can manifest in **face-based gender classification models**, even when no explicit decision is made.

We compared two models across three datasets with varying quality and composition.
The results revealed a consistent pattern:

- Models perform **well for dominant groups** (especially White men)
- But **fail disproportionately** for **Women of Color**
- Even under controlled conditions, **representation bias persists**
- Some subgroups (e.g. Black women) are consistently **misclassified**

This is a clear example of **representational unfairness**. A model that **"works well on average"** but performs **poorly for specific identities** can still be deeply unfair.

Reflection:

Which types of bias occurred in this example?

Relevant types of bias

Several forms of bias were present in this gender classification task:[6]

- **Historical Bias**: The model reflects **societal stereotypes** and **structural inequalities** — for example, limited representation of Women of Color in online image datasets and binary gender norms
- **Representation Bias**: Some subgroups (e.g. Black women) are **underrepresented** or appear in lower quality in training and test data, leading to poor recognition performance
- **Evaluation Bias**: Fairness assessments **vary across datasets** used for evaluation
- **Aggregation Bias**: Looking only at **overall accuracy** hides performance gaps (differences only visible through intersectional analysis)
- **Measurement Bias**: Inconsistent labels or image quality

Looking Ahead: When Bias Reinforces Itself

Up to now, we've analyzed fairness issues based on **group-wise model performance**.
But what happens when these biases do **not just reflect**, but actually **shape future outcomes**?

In the next section, we explore **Self-Fulfilling Predictions** and **Feedback Loops**:

> How do biased predictions (e.g. in policing or education) influence behavior, resource allocation, or social outcomes?
> And how can this lead to a **reinforcement of existing inequality**?

We'll see how machine learning models can **amplify the very problems they aim to solve.**

Quiz

1. True or False: Representational fairness evaluates whether model decisions are distributed fairly across groups.

1. ☐ True
2. ☐ False

2. What makes representational fairness different from decision fairness? *(Select one option)*

1. ☐ It applies only to classification tasks like loan approval
2. ☐ It focuses on equal outcomes rather than equal recognition
3. ☐ It relies solely on demographic parity
4. ☐ It asks whether people are fairly represented, not fairly treated

3. Which of the following best describes a key finding from the Gender identification evaluation? *(Select one option)*

1. ☐ Accuracy was highest for intersectional minority groups
2. ☐ Women were more often misclassified than men, especially Women of Color
3. ☐ DeepFace failed mostly due to low image quality
4. ☐ The model performed best when sensitive attributes were used as input

---

Sources:

1. Binns, 2018
2. Karkkainen & Joo, 2021
3. UTKFace Dataset, n.d.
4. Taigman et al., 2014
5. Levi & Hassner, 2015
6. Suresh & Guttag, 2021

In [ ]:

Appendix H: Notebook Page 6

Part 6: Self-Fulfilling Predictions & Feedback Loops

Machine learning systems do not just reflect the world — they can often shape it.

In this part of the notebook we try to understand how biased predictions can reinforce themselves. A simulation study of predictive policing is used as a concrete example.

Two important concepts are:[1]

Concept	Description
Self-Fulfilling Prediction	A prediction changes behavior, decisions, or resource allocation in a way that the predicted outcome becomes true — not because the prediction was accurate, but because it influenced the result.
Feedback Loop	A cyclical process where model outputs affect behavior or data collection, which in turn feeds back into the system and reinforces future outputs. The model is increasingly confident in its own predictions, not because reality has changed, but because it shaped what was measured.

How they interact

- The two often **occur together** and **reinforce** each other.
- Difference:
  - **Self-fulfilling prediction**: usually a single causal chain
  - **Feedback loop**: iterative, with multiple cycles over time
- Both can **amplify historical inequalities** and distort the system's objectivity.

Ethical Perspective[1,7]

Self-fulfilling predictions are ethically relevant because they **influence social reality** while maintaining an illusion of neutrality.

- They do not merely *describe* the world, they *shape* it.
- Predictions affect **how people are treated** and **how resources are allocated**.
- The line between **observation** and **intervention** becomes blurred.
  - Example: A prediction changes conditions in such a way that it becomes true — not due to accuracy, but due to its effect.
- In such cases, algorithms can create the very **evidence** that appears to confirm their correctness.

These systems often seem objective and data-driven but in reality, they can **reinforce structural inequalities**.

The feedback mechanisms are often **hidden** or **indirect**, making their impact **difficult to detect**.

> It looks like the model is predicting the future,
> but it is actually shaping the future it claims to foresee.

Example of a self-fulfilling prediction: A university predicts that a student is likely to drop out → denies financial aid → student actually drops out.

Can initiate a **feedback loop** if similar cases influence future predictions.

Example: Predictive Policing

Predictive policing systems forecast:

- *Where* crime might happen
- *Who* might be involved

These predictions are based on **historical crime data**. This data reflects not just crime, but **past police presence**, strategies, and reporting behavior.

Place-Based Systems (e.g. PredPol)[2]

Step	Description
1.	Model analyzes past crime data (location, time, type).
2.	Prediction: Area X has a high crime risk.

Step	Description
3.	Police increase patrols in Area X.
4.	More incidents are detected in Area X.
5.	New data used for retraining → confirms original prediction.

Self-fulfilling prediction: Step 3 — the prediction leads to increased patrols, which increases the likelihood of detecting crime, making the prediction come true.

Feedback loop: Steps 4–5 — new data from increased patrols feeds back into the system, reinforcing and repeating the prediction over time.

Person-Based Systems (e.g. Strategic Subject List)[2]

- Use data like:
  - Social networks
  - Police interaction history
  - Commercial data
- Output: **Risk score** → influences surveillance, stops, and arrests.
- High-risk individuals are more likely to be investigated → higher chance of recording minor offenses → risk score reinforced.

Ethical Issues[2,3]

Issue	Description
Illusion of objectivity	Predictions seem data-driven but actually shape the outcomes.
Social impact	Reinforces stigma, stress, inequality, and may increase police violence.
Lack of transparency	Neither the public nor law enforcement may understand how scores are calculated.
No accountability	People cannot know or challenge why they are classified as "high-risk".

Example: *Beware* system in Fresno — even police couldn't explain how threat scores were assigned.

Predictive Policing Simulation

This case study demonstrates how **self-fulfilling predictions** and **feedback loops** can emerge in predictive policing systems using historical crime data. We simulate the effects of increased police presence in high-risk areas and show how this influences future model outputs.

We use real-world data from the **Berlin Crime Atlas** published by the Berlin Police.[5] The dataset contains **frequency values** for various crime categories, normalized by population to enable fair comparison between subdistricts. For this simulation, we use a processed version that:

- Focuses on the years **2018-2023** to reflect recent trends
- Contains only **subdistrict-level data**, avoiding double-counting from higher-level aggreagations and allowing a detailed **spatial analysis** of crime patters

---

First we import the necessary libraries. Use pip install for packages you do not have.

In [2]:
```
import pandas as pd
import geopandas as gpd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

Load and Merge Data

- Load Berlin subdistrict shapefile (`LOR_2023-01-01_BZR_EPSG_25833_nur_ID.shp`) for geographic boundaries.[4]
- Load Berlin crime data from 2018–2023 (`HZ_2018-2023.csv`).[5]
- Merge both datasets to associate crime values with spatial regions.

In [3]:
```
# Loading the shapefile - adjust to your save path
shapefile_path = r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\PredictivePolicing_Study\LOR 2.0\LOR_2023-01-01_BZR_EPSG_
geo = gpd.read_file(shapefile_path)

# Loading crime data (2018-2023) - adjust to your save path
crime_data_path = r'C:\Users\lukas\Desktop\Fairness_Notebook\Data\PredictivePolicing_Study\HZ_2018-2023.csv'
crime_df = pd.read_csv(crime_data_path)

# Merge data
geo['BZR_ID'] = geo['BZR_ID'].astype(str).str.zfill(6)
crime_df['LOR_Schluessel'] = crime_df['LOR_Schluessel'].astype(str).str.zfill(6)
berlin_crime_df = geo.merge(crime_df, left_on='BZR_ID', right_on='LOR_Schluessel')
```

Select and Standardize Features

- Select a wide range of crime categories across six years.
- Standardize all numerical features to make clustering meaningful.
  - This ensures that categories with larger raw values do not dominate.

The crime types are organized into:

- **Top-level total** ('Straftaten insgesamt')
- **Intermediate categories** (e.g. 'Diebstahl insgesamt')
- **Specific offenses** (e.g. 'Fahrraddiebstahl')

To make clustering informative but not redundant, the following **selection strategy** was used:

- **Exclude** overall total and 'Kieztaten' to avoid redundant information.
- **Keep intermediate categories** (e.g. 'Sachbeschädigung') because the specific subcategories have similar spatial pattern as their parent category (e.g. 'Sachbeschädigung durch Graffiti').
- Exception: **Theft types** - here we kept the **four subcategories**, because they show different spatial patterns:
  - 'Fahrraddiebstahl' (more in central districts)
  - 'Wohnraumeinbruch' (more in peripheral districts)
  - 'Diebstahl an/aus Kfz' (more in central districts)
  - 'Diebstahl Kraftwagen' (more in peripheral districts)

In [4]:
```
# Select relevant features for clustering
features = ['Raub_2018', 'Raub_2019', 'Raub_2020', 'Raub_2021', 'Raub_2022', 'Raub_2023',
            'Koerperverletzungen_2018', 'Koerperverletzungen_2019', 'Koerperverletzungen_2020',
            'Koerperverletzungen_2021', 'Koerperverletzungen_2022', 'Koerperverletzungen_2023',
            'Diebstahl_Kraftwagen_2018', 'Diebstahl_Kraftwagen_2019', 'Diebstahl_Kraftwagen_2020',
            'Diebstahl_Kraftwagen_2021', 'Diebstahl_Kraftwagen_2022', 'Diebstahl_Kraftwagen_2023',
            'Diebstahl_an/aus_Kfz_2018', 'Diebstahl_an/aus_Kfz_2019', 'Diebstahl_an/aus_Kfz_2020',
            'Diebstahl_an/aus_Kfz_2021', 'Diebstahl_an/aus_Kfz_2022', 'Diebstahl_an/aus_Kfz_2023',
            'Fahrraddiebstahl_2018', 'Fahrraddiebstahl_2019', 'Fahrraddiebstahl_2020', 'Fahrraddiebstahl_2021',
            'Fahrraddiebstahl_2022', 'Fahrraddiebstahl_2023', 'Wohnraumeinbruch_2018', 'Wohnraumeinbruch_2019',
            'Wohnraumeinbruch_2020', 'Wohnraumeinbruch_2021', 'Wohnraumeinbruch_2022', 'Wohnraumeinbruch_2023',
            'Branddelikte_2018', 'Branddelikte_2019', 'Branddelikte_2020', 'Branddelikte_2021', 'Branddelikte_2022',
            'Branddelikte_2023', 'Sachbeschaedigung_2018', 'Sachbeschaedigung_2019', 'Sachbeschaedigung_2020',
            'Sachbeschaedigung_2021', 'Sachbeschaedigung_2022', 'Sachbeschaedigung_2023', 'Rauschgiftdelikte_2018',
            'Rauschgiftdelikte_2019', 'Rauschgiftdelikte_2020', 'Rauschgiftdelikte_2021', 'Rauschgiftdelikte_2022',
            'Rauschgiftdelikte_2023']

for feature in features:
    berlin_crime_df[feature] = berlin_crime_df[feature].replace(',', '', regex=True).astype(float)

# Standardization
scaler = StandardScaler()
X_scaled = scaler.fit_transform(berlin_crime_df[features])
```

Initial Clustering

- Use KMeans to identify **4 crime pattern clusters** across all subdistricts.
- Sort clusters based on total crime volume.
- Assign each cluster a color:
  - 🔴 Red = highest total crime
  - 🟠 Orange = high
  - 🟡 Yellow = medium
  - 🟢 Green = low

In [5]:
```python
# KMeans-Clustering
kmeans = KMeans(n_clusters=4, random_state=42, n_init=20)
berlin_crime_df['Cluster'] = kmeans.fit_predict(X_scaled)

# Sort cluster and assign colors
cluster_means = berlin_crime_df.groupby('Cluster')[features].mean()
cluster_means['Total_Crime'] = cluster_means.sum(axis=1)

cluster_order = cluster_means.sort_values('Total_Crime', ascending=False).index.tolist()
colors = ['red', 'orange', 'yellow', 'green']
color_map = dict(zip(cluster_order, colors))
berlin_crime_df['Color'] = berlin_crime_df['Cluster'].map(color_map)
```

Visualize Initial Risk Clusters

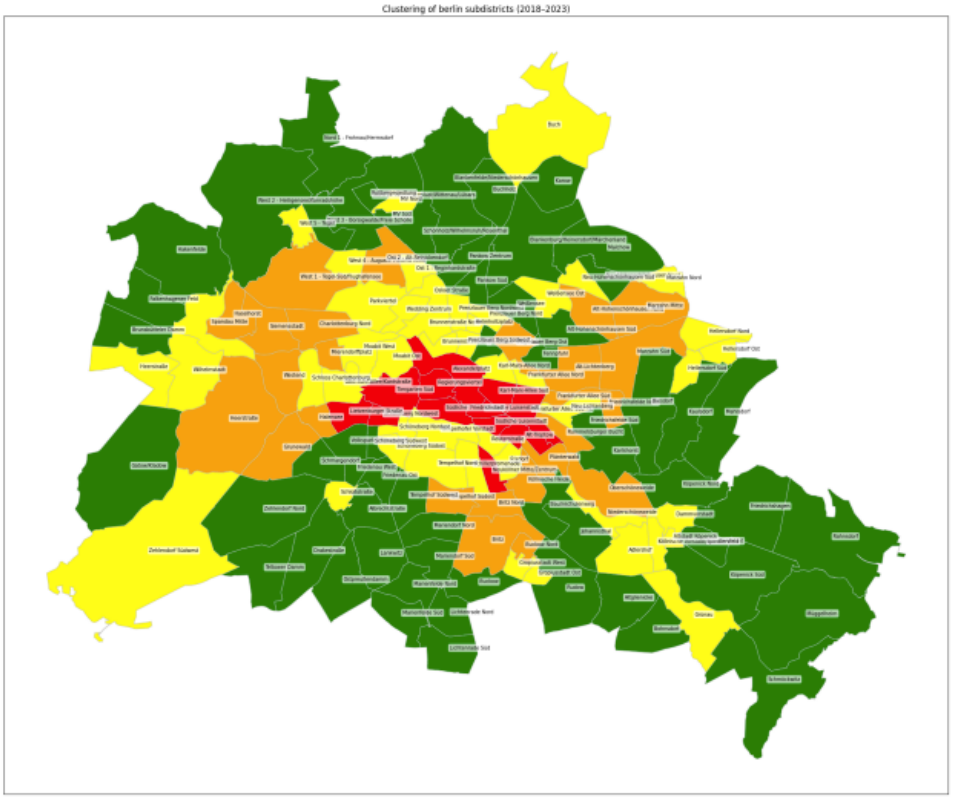
- Display all subdistricts colored by their initial crime cluster.
- These clusters represent the model's **first prediction** of risk levels.

> In real-world predictive policing, such predictions could guide increased patrols.

Reflection:

Where do you expect the high-risk clusters to be located - in central or peripheral districts?

In [6]:
```python
# Visualize the clusters (2018-2023)
fig, ax = plt.subplots(1, 1, figsize=(25, 20))
berlin_crime_df.plot(color=berlin_crime_df['Color'], linewidth=0.8, ax=ax, edgecolor='0.8')

# Show subdistrict names
for idx, row in berlin_crime_df.iterrows():
    centroid = row['geometry'].centroid
    ax.annotate(
        row['Bezeichnung'],
        xy=(centroid.x, centroid.y),
        xytext=(2, 2), # Slightly of center for less overlapping
        textcoords='offset points',
        fontsize=7,
        color='black',
        bbox=dict(boxstyle='round,pad=0.3', edgecolor='none', facecolor='white', alpha=0.7),
    )

# Remove axis labels and ticks
ax.set_xlabel('')
ax.set_ylabel('')
ax.set_xticks([])
ax.set_yticks([])

plt.title('Clustering of berlin subdistricts (2018-2023)')
plt.show()
```



Interpretation

As expected, the **highest-risk clusters** are concentrated in **central districts** like Kreuzberg, Mitte, and Friedrichshain.

These areas have higher rates of offenses such as **bicycle theft, drug-related crime, assault**, and **robbery**.

In contrast, **lower-risk clusters** are mostly located in **peripheral areas**, where offenses like **car theft** and **burglary** are more common but at lower overall volumes.

Simulate a Feedback Loop

We simulate a 5-year scenario where:

- Police allocate more resources to **high-risk clusters**.
- This leads to more crimes being discovered in these areas — **not necessarily more committed**, but more *recorded*.
- We model this as:
  - +10% annual increase in cluster 0 (highest)
  - +5% annual increase in cluster 1 (second highest)

This illustrates a **self-fulfilling prediction** that turns into a **feedback loop**.

In [7]:
```
# Copy
features_simulated = features.copy()
berlin_crime_simulated = berlin_crime_df.copy()

# Identify clusters
hotspot_cluster = cluster_order[0]  # Cluster with highest crime rate
secondary_cluster = cluster_order[1]

# Select hotspot and secondary areas
hotspot = berlin_crime_simulated['Cluster'] == hotspot_cluster
secondary = berlin_crime_simulated['Cluster'] == secondary_cluster
```

```
# 5 years simulation (10% in cluster 0, 5% in cluster 1)
for year in range(1, 6):  # 5 years
    berlin_crime_simulated.loc[hotspot, features_simulated] *= 1.1  # 10% increase
    berlin_crime_simulated.loc[secondary, features_simulated] *= 1.05  # 5% increase
```

Re-Clustering After Biased Data Accumulation

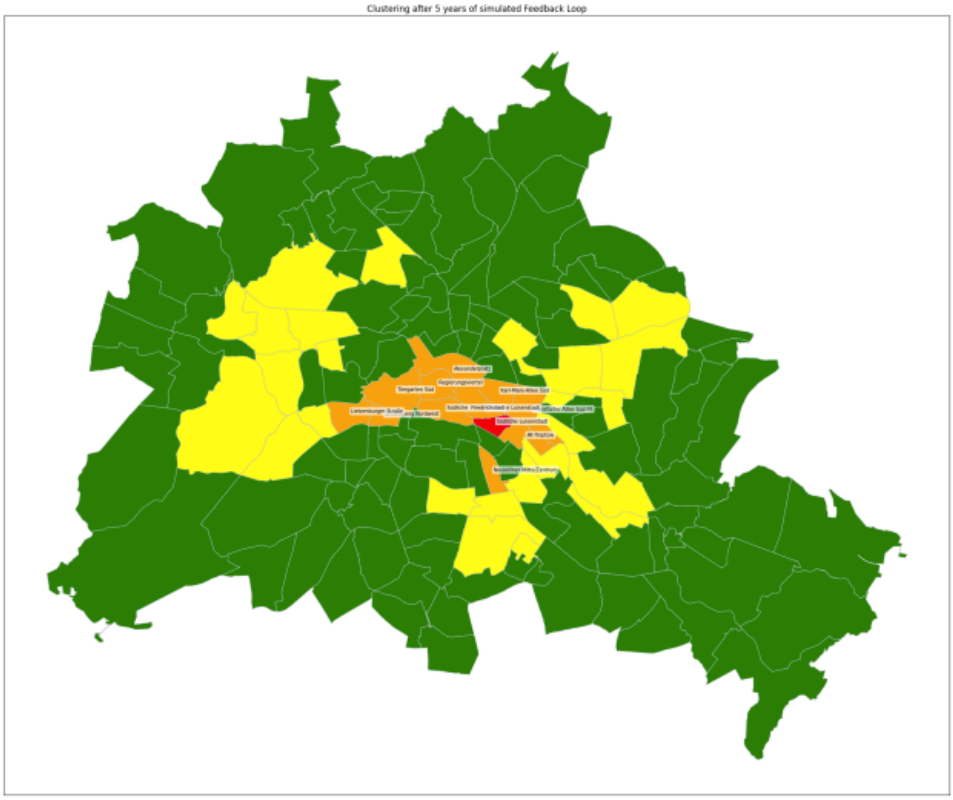
- Recalculate crime totals based on the adjusted numbers.
- Rerun clustering with the new data.
- Visualize new cluster assignments to assess the impact of biased reinforcement.

Reflection:

Do you expect any differences to the initial clustering?

In [11]:

```
# Scaling and clustering with simulated values
X_scaled_simulated = scaler.fit_transform(berlin_crime_simulated[features_simulated])
berlin_crime_simulated['Cluster_Sim'] = kmeans.fit_predict(X_scaled_simulated)

# Sort and assign colors
cluster_means_simulated = berlin_crime_simulated.groupby('Cluster_Sim')[features_simulated].mean()
cluster_means_simulated['Total_Crime'] = cluster_means_simulated.sum(axis=1)
cluster_order_simulated = cluster_means_simulated.sort_values('Total_Crime', ascending=False).index.tolist()
color_map_simulated = dict(zip(cluster_order_simulated, colors))
berlin_crime_simulated['Color_Sim'] = berlin_crime_simulated['Cluster_Sim'].map(color_map_simulated)

# Visualization
fig, ax = plt.subplots(1, 1, figsize=(25, 20))
berlin_crime_simulated.plot(color=berlin_crime_simulated['Color_Sim'], linewidth=0.8, ax=ax, edgecolor='0.8')

# Only annotate subdistricts in the two highest crime clusters
high_risk_clusters = cluster_order_simulated[:2]

# Add subdistrict names
for idx, row in berlin_crime_simulated.iterrows():
    if row['Cluster_Sim'] in high_risk_clusters:
        centroid = row['geometry'].centroid
        ax.annotate(
            row['Bezeichnung'],
            xy=(centroid.x, centroid.y),
            xytext=(2, 2),
            textcoords='offset points',
            fontsize=7,
            color='black',
            bbox=dict(boxstyle='round,pad=0.3', edgecolor='none', facecolor='white', alpha=0.7),
    )

# Remove axis labels and ticks
ax.set_xlabel('')
ax.set_ylabel('')
ax.set_xticks([])
ax.set_yticks([])

plt.title('Clustering after 5 years of simulated Feedback Loop')
plt.show()
```



Interpretation

The new clustering map after five years of simulated predictive policing shows a more **centralized pattern** of hotspot detection. The previously high-risk district **südliche Luisenstadt** in Kreuzberg remains red, while sourrounding areas in Friedrichshain and Mitte shift from red/orange to orange/yellow. Peripheral areas are now all green.

This does not reflect an **actual decline** in crime elsewhere. It is the result of **biased data collection**. The intensified focus on high-risk areas leads to more detected incidents, which reinforces their classification. Over time, this creates a **feedback loop** that amplifies the model's inital assumption.

What began as a prediction turns into a **self-fulfilling prediction**. The model learns that the same areas stay dangerous, not because more crimes are committed, but because more crimes are found there.

Ethical Implications[2]

- What seems like data-driven **objectivity** can be the **result of the model influencing its own input**.
- This creates **self-fulfilling predictions**: after initial clustering → more police presence in hotspots → more detected crime → stronger risk signals → even more presence.
- The resulting **feedback loop** does not reflect actual crime, but a distorted version of it.
- This can **amplify social inequality** and lead to **over-policing** of already marginalized communities.
- As shown by Ensign et al. (2017), these feedback effects cannot be solved by adding more data — only by **correcting for model bias**, e.g. by down-weighting discovered incidents.[6]
- Without correction, such systems risk misleading decision-makers while appearing **neutral**.
- **Transparency, accountability**, and **critical evaluation** are essential in the design and evaluation of predictive models.

This example has shown how predictive models can unintentionally shape the very data they rely on. But feedback loops and self-fulfilling predictions are only one example of **many fairness challenges** in machine learning. In the next section, we will explore why **achieving fairness is so difficult** in practice and what strategies can help us move toward more responsible ML systems.

Quiz

1. True or False: Feedback loops in predictive systems can cause the model's predictions to become less accurate over time.

1. ☐ True
2. ☐ False

2. What is a key risk of predictive policing systems influenced by feedback loops? *(Select one option)*

1. ☐ Misallocation of resources and distorted crime patterns
2. ☐ Data privacy violations only
3. ☐ Focus is too wide
4. ☐ Better crime prediction in low-risk areas

3. Which of these interventions could help mitigate feedback loops in predictive systems? *(Select one option)*

1. ☐ Increasing model complexity
2. ☐ Down-weighting discovered incidents during training
3. ☐ Ignoring high-risk areas in future predictions
4. ☐ Collecting more data of the same type

Sources:

1. Barocas et al., 2023
2. Robinson & Koepke, 2016
3. Lum & Isaac, 2016
4. Amt für Statistik Berlin-Brandenburg, 2021
5. Polizei Berlin, 2023
6. Ensign et al., 2017
7. Osoba & Welser, 2017

In [ ]:

Appendix I: Notebook Page 7

Part 7: Challenges and Recommendations for Fairness in Machine Learning

Why Improving Fairness is Hard

Fairness in machine learning is not a one-time technical fix.
It is a **continuous, context-sensitive process** that must balance competing values and account for the real-world effects of algorithmic decisions.

This section highlights **five major challenges** in achieving fairness — each one illustrated through the three case studies:
student dropout prediction, **gender classification**, and **predictive policing**.

---

1. Conceptual limitations of fairness definitions

Fairness in machine learning is often defined using mathematical metrics like **Equalized Odds** or **Calibration**.
These definitions are useful diagnostic tools but they come with **mathematical and conceptual limitations**.

Correlation vs. Causality[16]

Observational fairness metrics are based on **statistical correlations**, not **causal relationships**. Machine learning models trained on observational data often learn patterns that **reflect existing inequalities**, rather than their causes.

Example: Predictive Policing

A model might learn that certain neighborhoods have higher crime rates — not because they are more dangerous, but because they are more heavily policed. Treating such correlations as causal can reinforce biased enforcement strategies and fail to address the actual causes of crime.

> Fairness interventions that ignore causal mechanisms risk producing **harmful or misleading outcomes**.

Whenever possible, **causal reasoning** should complement statistical fairness metrics — even though it requires strong **domain knowledge** and may not always be feasible.

Incompatibility of metrics

As already discussed in previous sections, some fairness metrics are **mathematically incompatible**. It is often impossible to satisfy both simultaneously unless the model is perfect or group base rates are identical.[1,2]

Example: Student Dropout Prediction:

Enforcing **Separation** via threshold adjustment led to a loss in **Sufficiency**.
This meant that predicted risk scores no longer had a consistent meaning across groups.

Inframarginality[3]

Many fairness metrics, such as Statistical Parity or Predictive Equality, operate on **infra-marginal statistics**. They summarize model performance **across the whole distribution**, rather than focusing on **decisions near the threshold**, where real-world consequences are greatest.

Example: Diabetes Screening (Corbett-Davies et al., 2023)
Suppose a single threshold (e.g., 1.5% risk) is used to screen for diabetes.
Groups with **higher baseline risk** (e.g. Asian Americans) will be screened more often — maximizing **overall health outcomes**.
But this **violates Statistical Parity**, because not all groups are screened at the same rate.

Now, if we **adjust the thresholds** to **enforce parity**, we must:

- Raise the threshold for high-risk groups → they lose access to needed care.
- Lower the threshold for low-risk groups → they receive unnecessary tests.

The result: every group is worse off.

This fairness intervention leads to a situation where a different policy (e.g. keeping a uniform threshold) would produce better outcomes **for all**.

Pareto-dominated outcomes[3]

As Corbett-Davies et al. (2023) argue, fairness interventions that **optimize infra-marginal metrics** (like Statistical Parity) can lead to **pareto-dominated policies**.

A policy is pareto-dominated when an **alternative exists that would improve outcomes for all groups** — but is rejected due to the fairness constraint.

See example above: Diabetes Screening
Enforcing statistical parity led to **worse outcomes for every group**, even though the metric was satisfied.
The fairness intervention was **pareto-inefficient**: all groups could have been better off under the original threshold.

General insight:
This shows that **fairness metrics detached from real-world utility** can be misleading.

> **Satisfying a metric ≠ Fair outcome**

Fairness should not be judged solely by mathematical parity but by **whether interventions improve outcomes for the intended groups** — especially at the decision margin.

Key Insight:

- Many standard fairness metrics rely on simplified assumptions and ignore what happens at the decision boundary.
- This can lead to **pareto-dominated outcomes**, where **no group benefits**, even if fairness metrics are satisfied.
- Therefore, **fairness interventions should not be judged solely by mathematical definitions**, but by whether they **actually improve outcomes for the people involved**.

This calls for a **consequentialist perspective** on fairness:[3]
Instead of optimizing metrics in isolation, we must ask:

> *"What are the real-world effects of this intervention – and who benefits or is harmed?"*

Only by evaluating fairness in terms of **context, utility, and impact** can machine learning systems be designed to support equitable outcomes.

---

2. Context dependence of fairness metrics

Fairness definitions are **not universal**, they depend on **application domain**, **harms**, and **social context**. What is a fair decision in one setting may be inappropriate in another.[4,5]

Example: Student Dropout Prediction

Choosing between fairness metrics (e.g. Equal Opportunity vs. Predictive Parity) required ethical reflection about the **consequences** for students.

Example: Gender Classification

Standard metrics like Statistical Parity fail here.
Representational harms need other forms of assessment, like intersectional accuracy, and take visibility into account.

Example: Predictive Policing

Even if a model minimizes overall error, it may reinforce **historical over-policing**.
Fairness here must account for political and social history. Evaluating clustering requires a different procedure and cannot be captured by standard classification fairness metrics.

---

3. Representation and measurement issues

Bias often stems from **how the world is represented** in data and labels — not just how the model learns.[6]

Example: Student Dropout Prediction

The label "dropout" may reflect financial or structural **disadvantages** — not actual academic ability.
Models risk learning that "disadvantage = failure".

Example: Gender Classification

Gender is not always binary or observable. Labels in datasets like FairFace are often **annotator-assigned**, not **self-identified**.
This raises issues of **label validity** and representational fairness.[2,7]

Example: Predictive Policing

Training on historical police data reflects where police were active, not necessarily where crime occurred.
Models predict **policing behavior**, not **crime risk**, reinforcing **feedback loops**.

---

4. Lack of standards, transparency, and accountability

There are **no consistent standards** for evaluating fairness.[7,8]
Without transparency or clear responsibilities, fairness becomes difficult to assess or enforce.

Example: Student Dropout Prediction

There is **no agreement** on which metric to use.
Without standards, fairness choices become arbitrary and students may not understand or contest the decision.

Example: Gender Classification

Systems are often **black boxes**. Candidates might be sorted based on gender without knowing it.
This undermines autonomy and prevents contestation.

Example: Predictive Policing

Police officers at times don't even know how risk predictions are made.[9]
Lack of **explainability** and **accountability** limits meaningful oversight.

Tools like **Model Cards**[10] and **Datasheets for Datasets**[11] help improve transparency but are rarely adopted unless legally required.

---

5. Sociotechnical and institutional constraints

Fairness is not just a technical issue — it is embedded in **sociotechnical systems** shaped by organizations, stakeholders, and processes.[5]

> A **sociotechnical system** refers to the interplay between **technical components** (e.g. models, data, software) and the **social structures** around them — including institutions, regulations and user behavior.

In such systems, technical decisions are never neutral. They are shaped by and impact human contexts.

Example: Predictive Policing

Even a technically fair model still operates within **biased policing practices**.
ML systems can **legitimize unjust structures** if context is ignored.[12]

Example: Student Dropout Prediction

Universities may prioritize **cost-efficiency** over full support for all high-risk students.
Fairness interventions can clash with **institutional goals**.

Example: Gender Classification

HR systems often rely on **third-party black-box models**.
Institutions may have no control over data or modeling decisions.

The Five Abstraction Traps[5]

Fairness problems often result from **abstracting away social context**, to build generalizable, modular systems.
Selbst et al. (2019) describe five "abstraction traps" that lead to failed fairness efforts.

Trap	Description	Example
Framing Trap	Treating fairness as a modeling problem only	Ignoring why students drop out in the first place
Portability Trap	Assuming fairness solutions are generalizable	Using standard metrics for gender classification
Formalism Trap	Reducing fairness to math	Overemphasis on metric optimization (e.g. Equalized Odds)
Ripple Effect Trap	Ignoring how ML systems change behavior	Predictive policing influences policing patterns
Solutionism Trap	Assuming ML is always the best solution	Automating gender recognition without clear benefit

Using the Traps as Reflective Questions

Selbst et al. suggest that these traps can be **turned into a critical checklist** by reversing their order and framing them as questions.
This helps structure ethical reflection in the development or assessment process.

> **Start with the most fundamental question and work upward:**

1. **Solutionism** → *Should we build this system at all? Is machine learning the right tool here?*
2. **Ripple Effect** → *How will the system change the environment it operates in?*
3. **Formalism** → *Which notions of fairness are relevant – and are they contested?*
4. **Portability** → *Does the fairness approach actually fit the specific context?*
5. **Framing** → *Are we addressing the right problem – or abstracting away key social dynamics?*

> Applying fairness is not just about choosing a metric. It's about **asking the right questions**, in the right order.

Using these reversed abstraction traps as a guide helps ensure that fairness efforts are not only technical, but also socially meaningful.

Practical Implications and Tools

Improving fairness in machine learning requires more than technical interventions.
It involves institutional, procedural, and ethical considerations and faces **real obstacles** in practice.

Despite growing awareness, many fairness recommendations are **not widely implemented**, unless supported by **regulatory enforcement**.

> Some companies adopt ethical guidelines mainly to **avoid binding oversight**, not out of intrinsic commitment.[13]

In the following section, we highlight practical implications that can help support fairness.

Fairness as a Process, Not a Fix

Fairness is **not a static goal** or a property of the algorithm alone.
It is shaped by how the system is **designed, deployed, monitored** and situated within its **social context**.[5]

This process-oriented view emphasizes:

- **Continuous reflection** across the ML lifecycle
- The need to assess **real-world consequences**
- The importance of **stakeholder inclusion** and **interdisciplinary input**

As discussed earlier, fairness metrics can be misleading when detached from impact.

> A **consequentialist perspective** asks: *Who benefits? Who is harmed? What actually changes for people involved?*

The shift from solution-based to **process-based approaches** is supported by many authors. It shows that fairness is not a final state to be achieved, but more a commitment to **continuous reflection**.[3,5,11,14]

Fairness across the ML lifecycle

Fairness interventions can happen at different stages:[15]

- **Pre-processing:** Modify data before training (e.g. resampling, reweighting)
- **In-processing:** Adjust algorithms (e.g. fairness constraints, adversarial learning)
- **Post-processing:** Modify outputs after training (e.g. Fairlearn's ThresholdOptimizer)

Each method requires choosing which type of fairness matters most and being aware of **contextual consequences**.

Transparency, explainability, and accountability

Improving fairness in machine learning is about making system **understandable, accessible, and contestable**.

A key factor for this is **communication**:[6,7,16]

> Fair ML systems must enable both **developers** and **affected individuals** to understand how decisions are made and to respond when they are unfair.

Tools for transparency and documentation:

- **Model Cards**:[10] Document the model's purpose, assumptions, limitations, and performance across demographic groups.
  → Should include **intended use**, **intersectional evaluation**, and **risk of misuse**.

- **Datasheets for Datasets**:[11] Describe dataset origin, structure, collection methods, and potential biases.
  → Help evaluate whether a dataset is suitable for a given task or context.

These tools increase visibility but they only help **if the information is accessible and communicated clearly**.

Explainability[7]

Even when models are documented, they may remain opaque.
Explainability means making the system's behavior **understandable for humans**, especially those affected by its decisions.

- For end users: explanations should be **simple and meaningful**
- For domain experts: tools like **LIME** or **SHAP** can offer technical insights

However, explainability often comes with trade-offs:

- High-performing models may be hard to explain
- Privacy concerns may limit what can be disclosed[17]

Accountability

Transparency and explainability are **necessary, but not sufficient**.

> Without **clear responsibility**, even a well-documented system lacks fairness.

- Who decided which features to use?
- Who defined the target variable?
- Who can be contacted to contest a decision?

Accountability requires **defined roles**, **oversight mechanisms**, and (ideally) the ability to **appeal or challenge** automated decisions.

In all three case studies, affected individuals had little understanding of the system and no way to contest outcomes:

- **Students** receiving dropout risk labels without explanation
- **Job applicants** being filtered by black-box gender classifiers
- **Citizens** being policed based on opaque hotspot predictions

> Fairness depends on more than open data or transparent models.

It requires **communication structures** that enable understanding, trust, and recourse for **all stakeholders**, especially those affected by the system.[4,16]

Tools and Institutional Practices

Several tools and practices can support fairness. However, these tools should not be used in isolation. They must be applied with an understanding of the social context and stakeholder needs.

Example Toolkits:

- **AIF360 (IBM)**:[18] Metrics and pre/in/post-processing methods for fairness evaluation
- **Fairlearn (Microsoft)**:[19] Model diagnostics, fairness-accuracy trade-offs, and threshold tuning
- **SageMaker Clarify (Amazon)**:[20] Bias detection and explainability tools

Institutional Practices:

Fairness in machine learning is not only a question of metrics or tools, it depends on **how organizations develop and govern these systems**.

To move from abstract principles to meaningful practice, institutions can adopt several strategies:

- Establish **fairness review boards**, ethics committees, or internal audit processes
  → These structures ensure that fairness decisions are **not left to individual developers**, but handled collectively and transparently.[21]
- Provide **training on fairness, bias awareness, and societal impacts**[16]
  → Teams need to understand how their decisions affect real people — beyond just technical performance.
- Promote **diverse team composition**[22,23]
  → Homogeneous teams often fail to anticipate harms experienced by marginalized groups.
  Example: Lee (2018) shows that the lack of diversity in the tech industry can lead to serious blind spots. She describes a photo-tagging algorithm that labeled Black individuals as "gorillas". This failure that could likely have been avoided with greater team diversity.
- Enable **interdisciplinary collaboration**[6]
  → Fairness challenges require knowledge from law, sociology, ethics, and domain-specific fields.
  → Including these perspectives helps detect risks and design systems that are better aligned with societal values.
- Involve **affected communities and domain experts** early in the process[16,24]
  → Fairness cannot be defined solely by engineers.

→ Participatory design improves contextual understanding and **legitimizes fairness goals**.

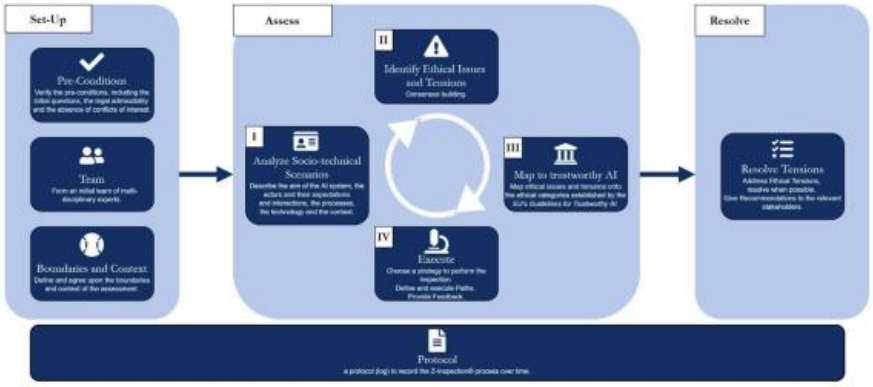
As the case studies illustrate, many harms arise not from malicious intent, but from **limited perspectives and missing expertise**.

> Addressing fairness requires institutions to invest in **inclusive practices, interdisciplinary thinking, and stakeholder engagement** — not just tools.

Z-Inspection®: A practical fairness framework[13,25,26]

Z-Inspection® offers a **holistic and interdisciplinary framework** to evaluate fairness in high-impact AI systems.

Phase	Description
Set-up	Define scope and form independent, diverse expert team
Assess	Develop sociotechnical scenarios, identify ethical tensions, map to EU AI guidelines
Resolve	Generate concrete recommendations — including technical fixes or stopping deployment



Example:
An AI system for detecting cardiac arrest in emergency calls showed reduced performance for non-native speakers. Fairness concerns were not visible through metrics alone — they emerged through contextual assessment.

> Z-Inspection® emphasizes that **fairness is embedded in systems, not just models**.

It requires **interdisciplinary collaboration**, **continuous monitoring**, and **ethical reflection** throughout the lifecycle.

Summary

- **Fairness is not a fixed goal** but a **continuous process** shaped by context and consequences.
- Many standard fairness metrics have **mathematical and conceptual limitations**.
- Each of the three case studies shows different challenges — and why **technical fixes alone aren't enough**.
- Addressing fairness means looking at the complete sociotechnical system and not just numbers.

Looking Ahead

Hopefully this notebook has provided you with a solid foundation for understanding the **ethical challenges of fairness in machine learning**. The goal was not to present exhaustive solutions, but to introduce core concepts, raise awareness of common pitfalls, and encourage critical reflection. By combining theory, code, and case studies, the notebook offers an **entry point** into a complex and evolving field. Throughout the three integrated studies — on **student dropout prediction, gender classification**, and **predictive policing** – you have seen that fairness:

- cannot be reduced to a single metric or technical fix
- requires attention to representation, context, and long-term effects
- must be assessed within the **broader sociotechnical environment** in which systems operate

The notebook is designed to help you build a **conceptual framework** that you can expand over time — as technologies evolve, debates progress, and regulatory frameworks emerge.

> **Note**: Legal and regulatory aspects (such as the EU AI Act) were not addressed in this notebook, but they are increasingly influencing real-world practice.

If you are interested in learning more, here are some recommendations:

Source	Description	Link
Barocas et al. (2023)	Comprehensive overview of fairness, bias, and algorithmic decision-making	Fair ML Book
Verma & Rubin (2018)	Summary of different fairness metrics	Fairness Metrics
Zicari et al. (2021) & Boonstra et al. (2024)	Real-world applications of the Z-Inspection® framework — shows what fairness assessments look like in practice	AI in healthcare & AI in nature monitoring
Angwin et al. (2016)	Original investigation of the COMPAS case	COMPAS
Corbett-Davies (2023)	Highlights key challenges like pareto domination in fairness metrics	Challenges in Fair ML
Kaggle: Intro to AI Ethics	Practical exercises, interactive tutorials, and further links such as Google's interactive explainer	Kaggle course
Mehrabi et al. (2021)	A more compact overview of fairness and bias in ML (compared to Barocas et al.)	Survey on Bias and Fairness
Lum & Isaac (2016)	Introduction to predictive policing	Intro Predictive Policing
Robinson & Koepke (2016)	Builds on insights from Lum & Isaac	More Predictive Policing

Other references used in this notebook are listed in the user guide and cited throughout the sections.
For a deeper exploration, you can also access the master's thesis on which this notebook is based: *GitHub*.

Fairness in machine learning cannot be fully solved but it can be better **understood**, more **transparently discussed**, and more **responsibly addressed**. We hope this notebook helped you take the first step in that direction.

Quiz

1. True or False: Fairness metrics can always tell us whether a system is fair in practice.

1. ☐ True
2. ☐ False

2. What does inframarginality describe? *(Select one option)*

1. ☐ Errors that occur due to biased sampling
2. ☐ Ignoring fairness concerns at the decision boundary
3. ☐ When a fairness metric applies only to large groups
4. ☐ When predictions are perfectly calibrated across all subgroups

3. What is a pareto-dominated outcome? *(Select one option)*

1. ☐ A system that performs equally across all metrics
2. ☐ An outcome that benefits one group at the expense of another
3. ☐ An outcome where a better alternative exists for all groups
4. ☐ A fairness intervention that leads to identical thresholds for all groups

---

Sources:

1. Chouldechova, 2017
2. Kleinberg et al., 2016
3. Corbett-Davies et al., 2023
4. Osoba & Welser, 2017
5. Selbst et al., 2019
6. Suresh & Guttag, 2021
7. Ntoutsi et al., 2020
8. Olteanu et al., 2019
9. Robinson & Koepke, 2016
10. Mitchell et al., 2019
11. Gebru et al., 2018
12. Lum & Isaac, 2016
13. Boonstra et al., 2024
14. Binns, 2018
15. Calegari et al., 2023
16. Barocas et al., 2023
17. Schmidt et al., 2024

18. LF AI & Data Foundation, n.d.
19. Fairlearn Organization, n.d.
20. Amazon Web Services, n.d.
21. Raji et al., 2020
22. Lee, 2018
23. Yapo & Weiss, 2018
24. Howard & Borenstein, 2018
25. Zicari et al., 2021
26. Zicari et al., 2022
27. Scheuerman et al., 2020

In [ ]: