thws Technical University of Applied Sciences Würzburg-Schweinfurt

FAKULTÄT INFORMATIK UND WIRTSCHAFTSINFORMATIK

LECTURE NOTES

Bias & Discrimination

How unequal treatment enters AI systems, and how to reason about it

MODUL
Introduction to the Ethics and Regulation of AI

SEMESTER
WS 26/27

VERSION
1.0

Prof. Dr. Markus Oermann
Instructor
markus.oermann@thws.de

Technische Hochschule Würzburg-Schweinfurt
Stand: July 23, 2026

Contents

In the previous unit we saw that responsibility cannot rest with the machine but must be distributed among human agents and built into development from the start; this unit takes up the value that proved hardest to specify and easiest to violate, namely fairness. AI systems draw distinctions for a living: a spam filter separates wanted from unwanted mail, a scoring model separates likely defaulters from safe borrowers. Drawing distinctions is not in itself wrong. The ethical and legal problem begins when a system treats people who are alike in the relevant respects unequally, systematically and to their detriment, along lines such as gender, ethnicity or disability. This unit asks how such bias enters AI systems, how we can classify it, how landmark cases expose it, and why no system will ever be perfectly fair.

01
KAPITEL

What is bias, and where does it come from?

1.1 A working definition

It helps to fix terms before piling up examples. Following Friedman and Nissenbaum, whose 1996 article opened the systematic study of the problem, we understand **bias** in a computer system as a *systematic and unfair discrimination against certain individuals or groups*. Three words carry the weight. The discrimination must be *systematic*, that is, a recurring tendency rather than a one-off error; it must be *unfair*, judged against some standard of what people are owed; and it must fall on *identifiable groups or individuals*, typically along socially salient lines. Note immediately that whether a given unequal treatment counts as *unfair* is not a technical fact read off the data. It is an ethical and political judgement, and different societies and legal orders draw the line differently. This is why bias cannot be fully outsourced to engineers alone: deciding which distinctions are acceptable is a normative task.

DEFINITION

Bias (Friedman & Nissenbaum 1996)

A systematic and unfair discrimination in a computer system against certain individuals or groups; “systematic” excludes random error, and “unfair” is a normative judgement, not a purely technical one.

1.2 The Friedman & Nissenbaum typology

The lasting contribution of Friedman and Nissenbaum is not the definition but the map. They distinguish three sources of bias according to *where in the life of a system* the bias originates, and the distinction matters because each source calls for a different remedy. **Preexisting bias**

has its roots outside the system, in social institutions, practices and attitudes, or in the personal views of those who design it; the computer merely absorbs and, often, amplifies a discrimination that predates it. **Technical bias** arises from the technology itself: from the limitations of hardware and software, from an algorithm ported into a context for which it was never built, from flawed randomisation, or from the crude digitisation of a human concept such as a legal rule. **Emergent bias** appears only later, in the context of actual use, when real users diverge from the "imagined user" the designers had in mind, or when new societal knowledge outdates the system's assumptions. The three are not mutually exclusive; a single system can suffer all three at once, and the interesting real cases usually do.

EXERCISE · TERMS

A hiring model that learns a historical preference for male candidates carries a **preexisting** bias; a face classifier that fails on under-sampled groups carries a **technical** bias; a search engine whose harmful rankings surface only through real users at scale carries an **emergent** bias.

The practical payoff of the typology is diagnostic. Because preexisting and technical biases are, in principle, present at the moment of development, a well-structured and ethically informed design process can usually detect and correct them before deployment. Emergent bias is harder, because by definition it does not exist yet when the system ships; it surfaces once the system meets a world the designers did not fully anticipate, and it surfaces especially when a system built for one context is scaled into another. The organisational answer is agile, responsive development with continuous feedback loops, so that the system can be readjusted as its effects become visible. That answer, however, is precisely the one that is hardest to apply where it matters most, in domains such as criminal justice or safety-critical control, where reliability requirements forbid casual live experimentation. Test the typology on real cases in the classifier below before we work through them one by one.

1.3 A fuller taxonomy for machine learning

Friedman and Nissenbaum wrote before the deep-learning era, and modern machine learning has multiplied the entry points for bias across the development lifecycle. A recent systematic review by González-Sendino and colleagues reorganises the field around four families that map onto the stages of an ML project: **human bias** entering at requirements and data collection, **data bias** in the prepared training set, **learning bias** introduced by the model and its evaluation, and **deployment bias** that arises or intensifies in operation. The scheme is compatible with Friedman and Nissenbaum: human and data bias largely correspond to preexisting and technical sources, while deployment bias captures the emergent dimension, notably through the feedback loops we return to below.

This lifecycle view of bias, human bias (cognitive, historical, content production) and data bias (sample, features) entering early, learning bias (evaluation, aggregation) and deployment bias (deployment, feedback loop) entering later, follows González-Sendino, R., Serrano, E., Bajo, J. & Novais, P. (2024): A Review of Bias and Fairness in Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence* 9 (1), pp. 3-15, Fig. 4, https://doi.org/10.9781/ijimai.2023.11.001.

02
KAPITEL

Reading the cases

2.1 Algorithms of Oppression: search as a mirror of power

Safiya Umoja Noble's *Algorithms of Oppression* (2018) turned a private discomfort into a public argument. Searching for terms about Black women on Google, Noble found that a supposedly neutral engine returned degrading, sexualised and racist results and suggestions. Her point is not that an engineer typed in prejudice, but that a commercial ranking system, optimised for clicks and advertising revenue and fed by the aggregate behaviour of millions of users, faithfully reflects and then amplifies the racist and sexist associations already circulating in society. This is emergent bias in Friedman and Nissenbaum's sense, generated in live use, but it also carries a preexisting layer, since the raw material is social prejudice. Noble's deeper move is to reframe the problem as one of *power*: search engines are not mirrors held up to a neutral world but privately owned infrastructures whose defaults shape what counts as knowledge, and the groups least able to contest those defaults bear the cost.

2.2 Proxy bias: Apple Card and Gender Shades

A recurring defence against discrimination claims is that the system never used the protected attribute at all. The Apple Card episode of 2019 shows why that defence is weak. Applicants publicly reported that men received substantially higher credit limits than their spouses, despite shared assets and joint tax filings, and the issuer responded that gender, ethnicity, age and sexual orientation were never inputs to the model. Both statements can be true at once. The system did track and use behavioural data, such as shopping patterns, and such features correlate with gender because past inequality shaped them. Excluding the protected attribute therefore does not remove its influence; the model reconstructs it from **proxy** features. This is *proxy bias*, a technical bias in Friedman and Nissenbaum's vocabulary, because it stems from how human categories are (mis)encoded into features rather than from a designer's animus.

The most cited demonstration that "we did not use the attribute" is no defence comes from Joy Buolamwini and Timnit Gebru's *Gender Shades* study (2018). Auditing three commercial gender-classification systems, they found that overall accuracy figures masked a stark gap: the systems classified lighter-skinned men almost flawlessly, with error rates below one percent, while misclassifying darker-skinned women in up to 34.7 percent of cases. The bias here is technical, rooted in unrepresentative training and benchmark data that under-sampled darker-skinned faces and women, so the models were quietly optimised for the majority. Two lessons generalise. First, aggregate accuracy hides subgroup harm, and only *disaggregated* error rates reveal it. Second, the worst outcome sat at the *intersection* of two axes, darker skin and female gender, neither of which alone predicted the failure, which brings us to intersectionality.

QUICK CHECK

An issuer states that its credit model never uses gender, yet women systematically receive lower limits. What best explains this?

The claim must be false; gender is secretly an input.
Correlated behavioural features act as proxies for gender, so excluding the attribute does not remove its effect.
Credit models are always accurate, so the pattern reflects reality.
Removing the gender field guarantees a fair model.

2.3 Intersectionality: why one axis at a time is not enough

The concept that names the *Gender Shades* pattern comes from law, not computer science. The legal scholar Kimberlé Crenshaw introduced **intersectionality** to describe harms that arise from the *combination* of several dimensions of discrimination, harms that are invisible if each dimension is examined separately. Her paradigm case is *DeGraffenreid v. General Motors* (1976). Black women sued General Motors for discrimination after layoffs; the court dismissed the claim, reasoning that the company did not discriminate against women, since it employed white women, and did not discriminate against Black people, since it employed Black men. The single-axis analysis missed the group that actually bore the harm, Black *women*, who were hired into neither track. Intersectionality is the antidote: it insists that discrimination against a compound group cannot be decomposed into independent tests along each axis, because the disadvantage lives precisely in the intersection. For AI fairness this is a design instruction. Auditing a model for gender parity and, separately, for racial parity can certify a system that still fails badly for a specific intersecting subgroup, exactly as *Gender Shades* found.

DEFINITION

Intersectionality (Crenshaw)

Discrimination that results from the combination of several characteristics (such as being both Black and a woman) and that a single-axis analysis, testing each characteristic in isolation, systematically fails to detect.

2.4 Error-rate bias and feedback loops: COMPAS

The COMPAS case is where fairness stops being intuitive and becomes mathematically thorny. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a risk-assessment tool used in US courts to estimate a defendant's likelihood of reoffending, informing bail and sentencing decisions.

COMPAS produces a raw recidivism score on a **1-to-10 scale**, which Northpointe then buckets into three risk categories: Low (1–4), Medium (5–7) and High (8–10). In practice, judges see this categorical label rather than the raw number, so "higher risk" in what follows means Medium or High. As with the explorer above, the raw score itself is neutral; the consequential decision, and the one open to challenge, is where the cut-offs between Low, Medium and High are drawn.

In 2016 the investigative outlet ProPublica analysed COMPAS's outputs and found a disturbing asymmetry in *how the tool failed*. Among defendants who did **not** go on to reoffend, Black defendants were labelled higher-risk far more often than white defendants; among those who **did** reoffend, white defendants were labelled lower-risk far more often. The errors, in other words, were distributed unequally, and the costs of those errors, wrongful detention on one side, fell disproportionately on Black defendants.

> *View the original chart, "Prediction Fails Differently for Black Defendants", ProPublica (2016), analysis of Broward County data: among defendants who did not reoffend, 44.9% of Black defendants but only 23.5% of white defendants were labelled higher risk; the tool made the opposite mistake for white defendants.*

This is **error-rate bias**: a system can be equally *accurate* overall for two groups yet distribute its false positives and false negatives very differently between them. The harm is compounded by a **feedback loop** peculiar to such deployments. Only defendants who are released generate new data about whether they reoffend, so the system learns chiefly from its own release decisions; false positives, people wrongly flagged as high-risk and therefore detained, rarely get the chance to prove the prediction wrong. The same structure appears in credit scoring, where only applicants granted a loan can be observed defaulting or repaying, so "false negatives", creditworthy people wrongly refused, never enter the correction loop. Over time such loops entrench existing disadvantage and manufacture *self-fulfilling prophecies*: the system's prediction shapes the outcome that then appears to confirm the prediction.

> *View the original chart in Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016, October 17). "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear."* The Washington Post*: the COMPAS risk-score distribution by race shows a far higher share of Black defendants falling into the medium and high-risk categories, so even Black defendants who do not reoffend are more likely to be scored high risk.*

2.5 The impossibility result

Northpointe, the maker of COMPAS, did not concede the point. It replied that its tool was **calibrated**: at any given risk score, the actual reoffending rate was the same for both groups, so a score of seven meant the same thing regardless of race. **Remarkably, both sides were right.** ProPublica's complaint about unequal false-positive rates and Northpointe's claim of calibration are both mathematically correct statements about the same tool. Alexandra Chouldechova (2017) explained why they must both be true and cannot be reconciled. When the *base rates* of the outcome differ between two groups, as they do here because of upstream, socially produced differences in arrest and conviction, it is mathematically impossible for a classifier to satisfy calibration, an equal false-positive rate, and an equal false-negative rate all at once. Satisfying any two forces a violation of the third. **The consequence is profound: there is no purely technical fix that makes such a system fair on every reasonable definition simultaneously. Choosing *which* fairness criterion to honour, and therefore which kind of error to tolerate and impose on whom, is an irreducibly political and ethical decision that no algorithm can make for us.**

QUICK CHECK

What does Chouldechova's impossibility result establish?

A well-designed model can satisfy every fairness metric at once.
When base rates differ between groups, calibration and equal false-positive and false-negative rates cannot all hold simultaneously.
Calibration is always the correct fairness criterion.
Bias can always be removed by deleting the protected attribute.

The result is easier to believe once you have tried and failed to beat it yourself. The explorer below models a COMPAS-like tool that scores every defendant on a 10-point risk scale; the tool itself is fixed, so everything comes down to *where you draw the line*, that is, which score on that scale counts as "high risk". You control that decision threshold and the gap between the two groups' base rates, and the three fairness metrics update live; try to find a threshold that makes all three equal at once while the base rates differ.

DEEP DIVE **Deep dive: base rates, and why they are not neutral**

It is tempting to treat the differing base rates in the COMPAS data as a brute fact that the algorithm merely reflects. That reading is too quick. Arrest and reconviction rates are themselves the product of policing patterns, prosecutorial choices and historical discrimination, so the "ground truth" the model is trained to predict is not an innocent measurement of criminality but a record of how the criminal-justice system has treated groups differently. Preexisting bias thus enters through the very definition of the target variable, before any question of metrics arises. This is why the impossibility result, though genuinely mathematical, does not licence a shrug: the choice of what to predict, on what data, is prior to and more consequential than the choice among fairness metrics, and it is a choice for which humans remain answerable.

03
KAPITEL

A structured test for unequal treatment

3.1 Borrowing the equal-treatment schema

Faced with a concrete system, how should one decide whether an unequal treatment is an acceptable distinction or an unfair bias? A workable method can be borrowed from the way constitutional equal-treatment guarantees are examined, adapted here as an ethical analysis rather than a legal opinion. The analysis proceeds in two stages.

1. First, one asks whether there is a *systematic, repeated unequal treatment of individuals or groups who are alike in the characteristics that matter for the context.* The unequal treatment can be **direct**, tied openly to a protected characteristic; **indirect**, tied to a proxy that correlates with it, as in the Apple Card; or **intersectional**, tied to a combination of characteristics, as in *DeGraffenreid*.
2. Second, if such unequal treatment exists, one asks whether there is an *objective, legitimate reason* for it that is suitable and proportionate to its consequences. The heavier the burden imposed, and the less the affected group or individuals can influence the differentiating factor, the stronger the justification must be. In these cases we end in an weighning of arguments.

CASE STUDY

Case: a scholarship-screening model

A university deploys a model that scores applicants for a merit scholarship using academic record, extracurricular activities and a "leadership potential" score derived from an interview transcript. After a year, staff notice that first-generation students and applicants from certain regions are systematically scored lower, though the model never uses socio-economic status or origin as an input. Is this bias, and how would you analyse it?
Solution. Apply the two-stage test. At the first stage there is a **systematic, repeated unequal treatment** of a group alike in the relevant respect (academic merit): first-generation and certain regional applicants score lower. Because the disadvantage runs through the "leadership potential" and extracurricular features rather than through socio-economic status directly, this is **indirect** (proxy) bias, most likely a *preexisting* bias absorbed from historical data about who displayed the valued traits, encoded through *technical* proxy features. It may also be **intersectional** if, say, first-generation *women* fare worst, which a single-axis audit would miss; the remedy is disaggregated evaluation across intersecting subgroups. At the second stage one asks for an **objective, legitimate reason** that is suitable and proportionate. Predictive accuracy alone will not do if the features encode disadvantage rather than merit, and the burden is heavy (denial of funding) while the differentiating factor lies largely outside the applicants' control. The responsible response is to interrogate the proxies, monitor error rates per subgroup, and route decisions through human review, echoing the value-driven design of the previous unit. The point is not that all differentiation is forbidden, but that this one lacks a proportionate justification.

04
KAPITEL

Four takeaways, and the limits of a technical fix

Pulling the threads together yields four conclusions that should outlast the individual cases. First, drawing distinctions is a *basic function* of many AI systems, so the question is never whether a system discriminates in the neutral sense but whether it does so unfairly. Second, *which* forms of unequal treatment count as unfair bias is an ethical and political decision, not a fact that engineering can settle. Third, and following directly, the question of *who has the power* to make those design and definitional choices, and to contest them, is of the first importance, which is Noble's and Crenshaw's central concern. Fourth, and soberingly, *there will never be an AI system entirely free of bias*: the impossibility result forecloses a perfectly fair classifier whenever base rates differ, and social reality guarantees that they will.

None of this counsels resignation. Bias can be reduced, and the review literature converges on a combination of measures rather than a single trick. Diversity along as many dimensions as possible within the responsible development team helps surface blind spots, and treating intersectionality as a working value, not merely a concept, sharpens the team's ability to spot compound harms. Technical mitigation is most effective when applied *before* training, by curating representative data, and is weakest when bolted on afterwards, and all of it must be embedded in data-governance and documentation practices that make external auditing possible. These are the same procedural, responsibility-distributing commitments the previous unit derived from the Collingridge dilemma, now aimed squarely at fairness.

QUICK CHECK

Which statement best reflects this unit's conclusion about bias mitigation?

A single debiasing algorithm can make any model fair.
Deleting the protected attribute is sufficient to prevent discrimination.
No system is perfectly bias-free; mitigation requires a combination of representative data, disaggregated evaluation, diverse teams and governance, and deciding which fairness criterion to honour is a normative choice.
Bias is purely a technical problem for engineers to solve.

05
KAPITEL

From fairness to data

Bias, we have seen, often enters through the data: through unrepresentative samples, through proxy features that reconstruct protected attributes, and through target variables that encode past discrimination. That makes the governance of personal data the natural next subject, and the following unit turns to the foundations of European data-protection law and the GDPR, which set the legal ground rules for how the data behind these systems may be collected and used in the first place.

06
KAPITEL

References

6.1 Literature

- Buolamwini, J. & Gebru, T. (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81 (Conference on Fairness, Accountability and Transparency), pp. 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html
- Chouldechova, A. (2017): Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2), pp. 153-163. https://doi.org/10.1089/big.2016.0047
- Crenshaw, K. (1989): Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* 1989 (1), pp. 139-167.
- Friedman, B. & Nissenbaum, H. (1996): Bias in Computer Systems. *ACM Transactions on Information Systems* 14 (3), pp. 330-347. https://doi.org/10.1145/230538.230561
- González-Sendino, R., Serrano, E., Bajo, J. & Novais, P. (2024): A Review of Bias and Fairness in Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence* 9 (1), pp. 3-15. https://doi.org/10.9781/ijimai.2023.11.001
- Noble, S. U. (2018): *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York.

6.2 Investigative reporting

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

6.3 Case law

- *DeGraffenreid v. General Motors Assembly Division*, 413 F. Supp. 142 (E.D. Mo. 1976).