

Ethics and Fairness

Last modified: 2026-10-27

i Source

Adapted from Logan Sizemore’s Spring 2025 lecture on ethical problems in machine learning.

Every other chapter in these notes asks whether a model is *accurate*. This one asks whether it should have been built, whose errors it makes, and who bears them. Those questions are not separable from the technical ones: the fairness section, Section 3, is written in the confusion-matrix notation of Metrics, because the central result in algorithmic fairness is a statement about what a confusion matrix can and cannot do.

1 Professional codes

- The Association for Computing Machinery adopted a Code of Ethics and Professional Conduct in 1992 and substantially revised it in 2018, largely to address the reach that data-driven systems had acquired in between
- Its general principles are short enough to quote:
 - Contribute to society and to human well-being
 - Avoid harm
 - Be honest and trustworthy
 - Be fair and take action not to discriminate
 - Respect the work required to produce new ideas, inventions, creative works, and computing artifacts
 - Respect privacy
 - Honor confidentiality
- The full text is at [acm.org/code-of-ethics](https://www.acm.org/code-of-ethics)¹
- A code is not a decision procedure. It is a list of the considerations a practitioner is expected to have weighed, and its value is that it makes “nobody thought about it” an unusable defence

1.1 Why this became urgent

- Two developments turned professional ethics in computing from a background concern into an urgent one:
 - **Consequential automated decisions.** Models now participate in criminal sentencing, loan approval, hiring, policing and medical triage — domains where an error is not an inconvenience
 - **Mass data collection.** Developers assemble the corpora these models train on at a scale where, in practice, they do not seek individual consent

¹<https://www.acm.org/code-of-ethics>

- The obstacle specific to this field is a belief that technical work is ethically neutral — that a model is a tool, and responsibility attaches only to whoever deploys it. The cases in Section 2.2 are the argument against that view

i Watch

von Luxburg’s lecture covers the general debate on machine learning and society^a [1, Pt. 45].

^a<https://www.youtube.com/watch?v=5CH2qcZQrpK>

2 Bias

In everyday usage “bias” means prejudice; in statistics it means a systematic departure of an estimator from the quantity it estimates; in Theory it is one term of the bias–variance decomposition. The sense used here is a fourth one:

Bias, in this chapter, is a systematic error in a model’s predictions that disadvantages particular individuals or groups.

Keeping these apart matters, because a model can be statistically unbiased and unfair at once, and the bias–variance sense has nothing to do with either.

i Watch

Narayanan asks whether statistical bias is an adequate fairness criterion, from 6:55^a [2].

^a<https://www.youtube.com/watch?v=jIXIuYdnyyk&t=415s>

2.1 Sources of bias

Bias can enter at each stage between the world and a deployed model [3]. Following the stages in order is a useful way to look for it.

- **In the world the data describe.** *Historical bias*: the data faithfully record a world that was already unequal. A model trained on past hiring decisions learns the past, including whatever was wrong in the past
- **In who is recorded.** *Representation bias*: some groups appear in the training data in smaller numbers. Because the training objective minimizes error averaged over the whole set, accuracy on a minority group can be traded away for a small gain on the majority, and the optimizer will make that trade. *Sampling bias* is the wider version: the training distribution differs from the deployment distribution for reasons unrelated to group size, such as data collected in one hospital, one city, one season
- **In what is measured.** *Label bias*: the labels were produced by human judgement that was subjective, inconsistent, or itself prejudiced, so the model learns the labellers rather than the phenomenon. A close relative is a *proxy* label, where the quantity recorded is only correlated with the one we care about, as in the first case of Section 2.2
- **After deployment.** *Temporal bias*: the data were accurate and are now out of date. Language, prices, diagnostic criteria and populations all move
- Collecting more data from under-represented groups fixes only representation and sampling bias. Historical and label bias survive any amount of additional sampling from the same process, because the process itself is wrong

i Read

GeeksforGeeks explains bias in machine learning^a [4].

^a<https://www.geeksforgeeks.org/machine-learning/bias-in-machine-learning-identifying-mitigating-and-preventing-discrimination/>

2.2 Documented cases

These are worth reading in full, and each names a different failure mode.

- **A health-care risk score** (2019). A commercial algorithm used across United States health systems ranked patients for extra care programs, using each patient's future health *costs* as a stand-in for their health *needs*. Obermeyer et al. [5] found that Black patients given the same risk score were sicker than White patients, because less money is spent on Black patients at the same level of need. They estimated that removing the bias would more than double the share of Black patients flagged for extra care. They also showed that changing the target away from cost reduces the bias. This is a case of label bias through a proxy: the model predicted its target well, and the target was the wrong one
- **COMPAS** (2016). A commercial recidivism-risk score used in United States pretrial and sentencing decisions. ProPublica's analysis found that among defendants who did not go on to reoffend, Black defendants were about twice as likely as white defendants to have been labeled high risk. The vendor's response was that the score was equally *calibrated* across groups — and it was. Both claims were true at once, and the next section shows how (ProPublica's methodology²)
- **Commercial face analysis** (2018). Buolamwini and Gebru [6] audited three commercial gender-classification systems on faces grouped by skin tone and gender. The error rate was at most 0.8% for lighter-skinned men and up to 34.7% for darker-skinned women. The authors also found the benchmark datasets used to report accuracy to be overwhelmingly lighter-skinned, so a single headline accuracy could hide the gap. This is representation bias, made visible by evaluating on subgroups rather than on the whole set
- **Amazon's recruiting tool** (2018). Trained on ten years of the company's own résumés, the model learned to penalize résumés containing the word “women's”. Amazon stopped using it. This is a clear case of historical bias: the model was an accurate summary of the training data (Reuters³)

i Watch

- Buolamwini talks about bias in face recognition^a [7].
- von Luxburg's lecture covers unfairness in machine learning^b [1, Pt. 46].

^ahttps://www.youtube.com/watch?v=UG_X_7g63rY

^b<https://www.youtube.com/watch?v=QgsRvQYjgKo>

2.3 Teaching these cases inside a machine learning course

- Harvard's Embedded EthiCS⁴ program is run jointly by Harvard's computer science and philosophy faculties. It puts philosophers into ordinary computer science courses, each of which then teaches a short module on the ethics of its own material, rather than leaving ethics to a separate course [8]
- The module Harvard lists for the Spring 2023 offering of its machine learning course, CS 181, is on bias in machine learning design [9]. It starts from a case of racial bias in a healthcare algorithm, and asks how seemingly innocuous technical decisions and straightforward data labels can produce biased outcomes. Students find the choice points in the design and propose alternatives, such as different

²<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

³<https://www.reuters.com/article/world/insight-amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/>

⁴<https://embeddethics.seas.harvard.edu/>

predictor variables, and then ask how a system can be improved once its bias is discovered. A second case, PredPol, asks students to identify the potential impacts of a predictive-policing system

- In the healthcare case, the algorithm’s label was a patient’s health costs, standing in for their health needs [5]. Less money had been spent on Black patients with the same level of need, so the algorithm judged them healthier than equally sick White patients. No labeller assigned that label: costs are recorded, not judged. The unequal spending they record is historical bias, let in by the choice of what the label measures
- The module’s page⁵ gives its reading, its class agenda and a sample activity, and describes the problem set that follows it

3 Fairness is not one quantity

Split the population into groups g — by race, sex, age, whatever the protected attribute is — and compute a separate confusion matrix per group, using the notation of Metrics. Four definitions of fairness are common, and each sounds like the obvious one.

i Watch

Narayanan shows where each fairness metric comes from, from 17:05^a [2].

^a<https://www.youtube.com/watch?v=jIXIuYdnyyk&t=1025s>

- **Demographic parity.** The model flags the same fraction of each group:

$$P(\hat{y} = T \mid g) \text{ is equal for all } g$$

- **Equalized odds.** The model makes the same *kinds* of mistakes in each group — equal false positive rates and equal false negative rates:

$$\frac{F_p}{F_p + T_n} \quad \text{and} \quad \frac{F_n}{T_p + F_n} \text{ are equal for all } g$$

- **Calibration.** A score means the same thing in each group. Among everyone assigned risk s , the same fraction actually offends, regardless of group:

$$P(y = T \mid \hat{s} = s, g) = s \text{ for all } g$$

- **Predictive parity.** Among everyone the model flags, the same fraction actually offends, regardless of group — equal positive predictive value, or PPV for short:

$$\frac{T_p}{T_p + F_p} \text{ is equal for all } g$$

- Calibration and predictive parity are not the same criterion, and the difference matters in Section 3.1. Calibration is a condition at every score value; predictive parity is a condition at one of them, the threshold at which we decide. Neither implies the other: a score can be calibrated in both groups while the fraction of flagged people who actually offend differs between them, because calibration constrains each score value on its own and says nothing about how a group’s mass is spread across those values. So the impossibility result has to be proved separately for each criterion, and it was.

i Read

GeeksforGeeks explains demographic parity and equalized odds^a [10].

^a<https://www.geeksforgeeks.org/artificial-intelligence/fairness-metrics-demographic-parity-equalized-odds/>

⁵<https://embeddedethics.seas.harvard.edu/cs-181-spring-2023/>

3.1 The impossibility result

- Whenever the base rate $P(y = T \mid g)$ genuinely differs between groups, and the model makes errors of both kinds — a false positive rate (FPR) and a false negative rate (FNR), with $\text{FPR} > 0$ and $\text{FNR} < 1$ — **predictive parity and equalized odds cannot both hold**, and so neither can calibration and equalized odds. This is a theorem, not an engineering shortfall — Chouldechova [11] and Kleinberg et al. [12] proved it independently
- The sketch is short enough to give. Write p for a group’s base rate, and let FPR and FNR be its two error rates. The positive predictive value — the quantity predictive parity equalizes — is determined by those three numbers:

$$\text{PPV} = \frac{p(1 - \text{FNR})}{p(1 - \text{FNR}) + (1 - p)\text{FPR}}$$

- Fix FPR and FNR to be equal across groups, as equalized odds demands, and PPV is then a function of p alone. That function is strictly increasing exactly when $\text{FPR} > 0$ and $\text{FNR} < 1$, which the theorem’s hypothesis guarantees: different p , different PPV — so predictive parity fails. Force predictive parity instead and the error rates must differ. There is no third option
- The hypothesis is needed, and is not only a technicality. Take $\text{FPR} = 0$ with $\text{FNR} = 0.5$ in both groups: the model is far from perfect, it misses half the positives, and yet $\text{PPV} = 1$ in both groups whatever the base rate, so predictive parity and equalized odds hold together. A model that never raises a false alarm has predictive parity trivially, because everything it flags is a true positive
- This is exactly the COMPAS dispute. Northpointe reported calibration; ProPublica reported unequal false positive rates. Neither was mistaken and neither was answering the other, because the base rates of re-arrest in the two groups were different and no scoring rule could have satisfied both criteria
- In practice, **“is this model fair?” is not a clearly defined question** until we name a fairness criterion. Choosing among them is a decision about which harm is worse — wrongly flagging someone, or wrongly clearing them — and that decision belongs to the people affected and their representatives, not to whoever is fitting the model

i Watch

- Narayanan explains why group fairness criteria conflict, from 10:27^a [2].
- von Luxburg’s lectures cover formal^b and algorithmic^c approaches to fairness [1, Pt. 47] and 48.

^a<https://www.youtube.com/watch?v=jIXIuYdnyyk&t=627s>

^b<https://www.youtube.com/watch?v=QWfJj0OxKso>

^c<https://www.youtube.com/watch?v=mi9kCpFNQNw>

4 Misinformation and privacy

- **Generation at scale.** A model that produces fluent text produces fluent false text just as readily, at a cost per item low enough to change what kinds of influence campaign are economical. The constraint on disinformation was never plausibility; it was labour
- **Memorization.** Large models memorize parts of their training corpora and can be induced to emit them. Nasr et al. [13] recovered megabytes of verbatim training data from production models, including personally identifiable information, using prompts no more sophisticated than asking the model to repeat a word indefinitely
- Memorization means an attacker can try to extract the training corpus from the deployed system. “We only trained on it” does not protect the data

Watch

Dwork gives the formal definition of differential privacy, from 17:46^a, and describes its use in the 2020 US Census, from 43:55^b [14].

^a<https://www.youtube.com/watch?v=Ai2O0xZBH8k&t=1066s>

^b<https://www.youtube.com/watch?v=Ai2O0xZBH8k&t=2635s>

Read

GeeksforGeeks explains differential privacy^a [15] in deep learning.

^a<https://www.geeksforgeeks.org/deep-learning/differential-privacy-and-deep-learning/>

5 Principles for practice

UNESCO's *Recommendation on the Ethics of Artificial Intelligence* [16], adopted by all 193 of its member states in November 2021, is the most widely endorsed statement of this kind. Its ten principles fall into three groups.

- *What a system must not do*: it must be proportionate to its purpose and do no harm, be safe and secure, protect privacy and personal data, and be fair and non-discriminatory
- *What people must be able to see and control*: the system's workings are transparent and explainable, humans oversee it and keep the final determination, and the people affected have a place in how it is governed
- *What its makers and owners owe*: responsibility and accountability for its effects, attention to sustainability, and public awareness and literacy about what such systems are

The full text is at [unesco.org](https://www.unesco.org)⁶.

Read alongside the impossibility result of Section 3.1, we should treat the list as a set of questions to answer rather than a checklist to satisfy. “Fairness and non-discrimination” does not tell you which fairness criterion to adopt; it tells you that you owe an answer.

Read

GeeksforGeeks explains AI ethics^a [17].

^a<https://www.geeksforgeeks.org/artificial-intelligence/ai-ethics/>

6 Questions for discussion

- Take a dataset you have worked with in this course. For each of the sources of bias in Section 2.1, decide whether you could detect it from the dataset alone, or only by knowing how the data was collected. Which of them leave no trace at all in the data they produced?
- You are asked to build a model that decides who receives a scarce resource — a transplant, a scholarship, a loan. Which of the four fairness criteria would you adopt, and what would you have to believe about the problem for that choice to be the right one?
- Under what circumstances is the correct engineering recommendation that the system should not be built?

⁶<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

- Can the developers of a system be morally responsible for its outcomes when their design choices are not the only cause of those outcomes? The Embedded EthiCS module of Section 2.3 poses this question [9], Key Philosophical Questions

7 Further reading

- Chouldechova [11] — the impossibility result in the COMPAS setting, free on arXiv:1703.00056⁷
- Kleinberg et al. [12] — the same result, proved independently and more generally. Both appeared in 2017, Chouldechova’s as a journal article and this one in conference proceedings. The preprints run the other way round: this one is dated 19 September 2016 and Chouldechova’s 28 February 2017. The proceedings are open access, and the preprint is on arXiv:1609.05807⁸
- Barocas et al. [3] — the standard textbook, available free from fairmlbook.org⁹
- Matloff [18] — a UC Davis course on the fairness of machine learning, built around the COMPAS case. Its reading list runs from the *ProPublica* investigation and Simpson’s paradox to fairness measures and remedies, and its homework, quizzes and discussion-section notebooks work in R on the COMPAS, German credit and other datasets from Scutari’s *fairml* package

References

- [1] U. von Luxburg, “Statistical machine learning.” University of Tübingen, lecture videos on YouTube, 2020. Available: <https://www.youtube.com/playlist?list=PL05umP7R6ij2XCvrRzLokX6EoHWaGA2cC>
- [2] A. Narayanan, “21 fairness definitions and their politics.” Tutorial at the Conference on Fairness, Accountability, and Transparency (FAT*); video on YouTube, 2018. Available: <https://www.youtube.com/watch?v=jIXIuYdnyyk>
- [3] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023. Available: <https://fairmlbook.org/>
- [4] GeeksforGeeks, “Bias in machine learning: Identifying, mitigating, and preventing discrimination.” GeeksforGeeks, tutorial article, 2025. Available: <https://www.geeksforgeeks.org/machine-learning/bias-in-machine-learning-identifying-mitigating-and-preventing-discrimination/>
- [5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342¹⁰.
- [6] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proceedings of the 1st conference on fairness, accountability and transparency*, in Proceedings of machine learning research, vol. 81. 2018, pp. 77–91. Available: <https://proceedings.mlr.press/v81/buolamwini18a.html>
- [7] J. Buolamwini, “How I’m fighting bias in algorithms.” TED, video on YouTube, 2017. Available: https://www.youtube.com/watch?v=UG_X_7g63rY
- [8] Embedded EthiCS, “About embedded EthiCS.” Harvard University, program webpage, n.d. Accessed: Oct. 05, 2026. [Online]. Available: <https://embeddedethics.seas.harvard.edu/about/>
- [9] M. Pope, “Bias in machine learning design: Embedded EthiCS module for CS 181, machine learning, spring 2023.” Embedded EthiCS, Harvard University, module webpage; the page dates the module AY 2021–2022, 2023. Accessed: Oct. 05, 2026. [Online]. Available: <https://embeddedethics.seas.harvard.edu/cs-181-spring-2023/>
- [10] GeeksforGeeks, “Fairness metrics - demographic parity, equalized odds.” GeeksforGeeks, tutorial article, 2025. Available: <https://www.geeksforgeeks.org/artificial-intelligence/fairness-metrics-demographic-parity-equalized-odds/>
- [11] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big Data*, vol. 5, no. 2, pp. 153–163, 2017, doi: 10.1089/big.2016.0047¹¹.

⁷<https://arxiv.org/abs/1703.00056>

⁸<https://arxiv.org/abs/1609.05807>

⁹<https://fairmlbook.org/>

¹⁰<https://doi.org/10.1126/science.aax2342>

¹¹<https://doi.org/10.1089/big.2016.0047>

- [12] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” in *8th innovations in theoretical computer science conference (ITCS 2017)*, in Leibniz International Proceedings in Informatics (LIPIcs), vol. 67. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43¹².
- [13] M. Nasr *et al.*, “Scalable extraction of training data from (production) language models.” arXiv:2311.17035, 2023. Available: <https://arxiv.org/abs/2311.17035>
- [14] C. Dwork, “Differential privacy and the US census.” Institute for Pure and Applied Mathematics, UCLA; video on YouTube, 2022. Available: <https://www.youtube.com/watch?v=Ai2O0xZBH8k>
- [15] GeeksforGeeks, “Differential privacy and deep learning.” GeeksforGeeks, tutorial article, 2023. Available: <https://www.geeksforgeeks.org/deep-learning/differential-privacy-and-deep-learning/>
- [16] UNESCO, “Recommendation on the ethics of artificial intelligence.” 2021. Available: <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>
- [17] GeeksforGeeks, “AI ethics.” GeeksforGeeks, tutorial article, 2025. Available: <https://www.geeksforgeeks.org/artificial-intelligence/ai-ethics/>
- [18] N. Matloff, “ECS 189G, Spring 2023: A Course on (Social) Fairness of Machine Learning Algorithms.” Course materials, University of California, Davis; GitHub repository, 2023. Accessed: Oct. 27, 2026. [Online]. Available: <https://github.com/ucdavis/FairMLCourse>

¹²<https://doi.org/10.4230/LIPIcs.ITCS.2017.43>