

Introduction to Deep Learning

Session 7 - AI & Ethics, a short introduction



Part 1 – Context

1. Motivations
2. Philosophical principles
3. Taxonomy of biases

Part 2 – Criterias & Examples

1. Mathematical Formulation
2. Impossibility Theorem
3. COMPAS case

Discussion

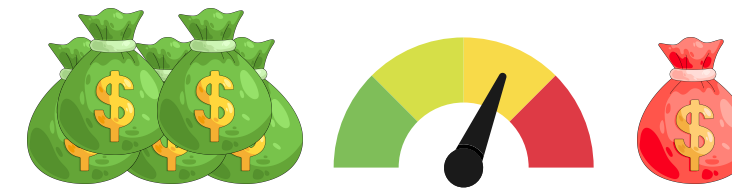
Justice

- Recidivism scores (COMPAS, 2016)



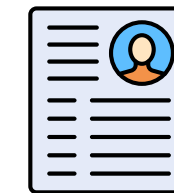
Credit

- Automated Credit Scoring



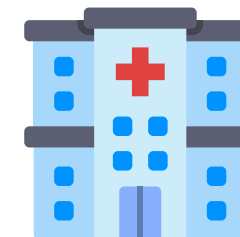
Employment

- Algorithmic CV screening (Amazon, 2014)



Healthcare

- Care prioritization, diagnostics



Content

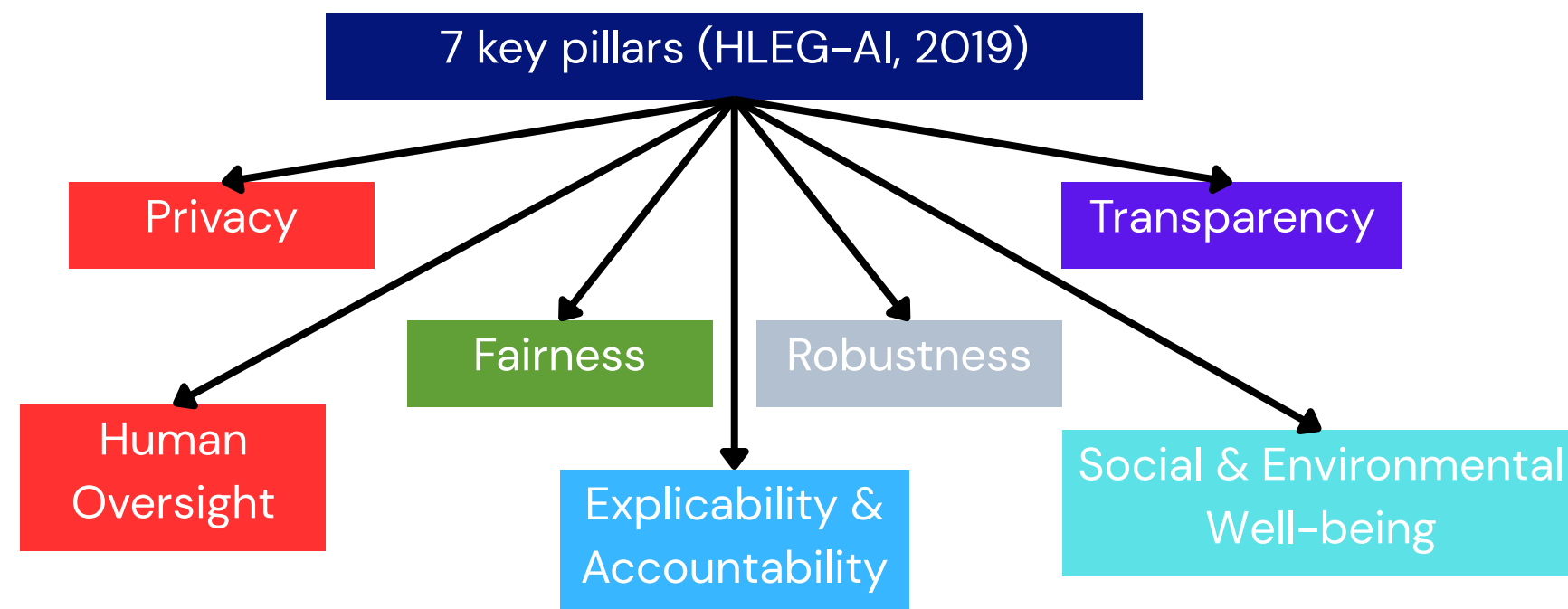
- Recommendations, moderation



AI Systems take decisions with a **real** impact on human beings

European Law

- **GDPR (2018)**
 - Right to explanation for automated decisions (art. 22)
- **HLEG-AI (2019)**
 - 7 key requirements including fairness, explicability, robustness
- **EU AI Act (2024)**
 - “High-risk” systems: mandatory audit and transparency



International Standards

- **IEEE Ethically Aligned Design (2019)**
- **OECD AI Principles (2019)**
- **NIST AI RMF (2023)**
 - Risk management framework

Art. 22 GDPR

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
 - (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
 - (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
 - (c) is based on the data subject's explicit consent.

Deontology – E. Kant (1800s)

- Certain actions are intrinsically good or bad, regardless of consequences.

In AI → absolute rules of non-discrimination, even at the cost of performance.



Consequentialism – J.S. Mill (1900s)

- The moral value of an action is judged by its effects on collective well-being.

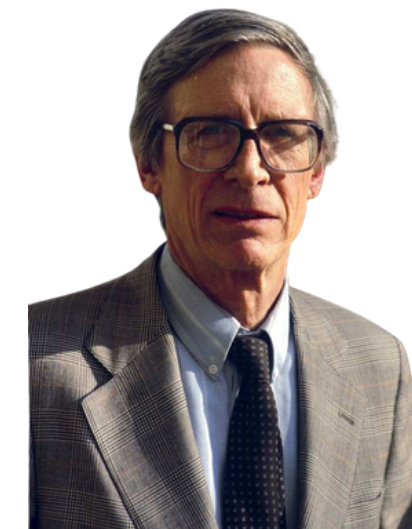
In AI → maximize aggregate utility. Tensions with individual rights.



Justice – Rawls (1971)

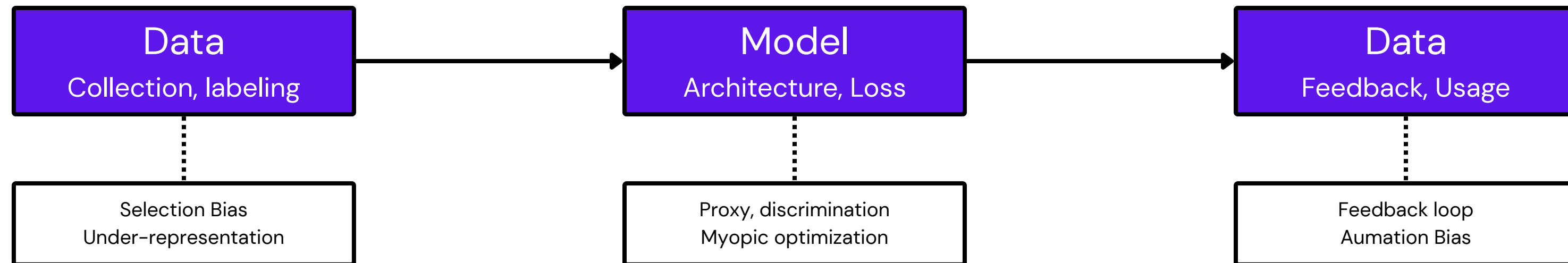
- Behind a “veil of ignorance”, what rules would you choose?

In AI → protect disadvantaged groups. *Difference Principle*.



Definition – Algorithmic bias

Systematic deviation of a model's outputs that disadvantages certain groups defined by sensitive attributes (e.g., race, gender, age, disability...).



- **Disparate treatment** – *explicit* use of a sensitive attribute as input
- **Disparate impact** – *indirect* discrimination via proxies (e.g., zip code, first name)

Notations

$X \in \mathcal{X}$ feature vector

$Y \in \{0, 1\}$ true target variable

$A \in \{0, 1\}$ sensitive attribute

$\hat{Y} = h(X)$ prediction

$p_a = P(Y = 1 | A = a)$ base rate

$TPR_a = P(\hat{Y} = 1 | Y = 1, A = a)$ True Positive Rate

$FPR_a = P(\hat{Y} = 1 | Y = 0, A = a)$ False Positive Rate

$PPV_a = P(Y = 1 | \hat{Y} = 1, A = a)$ Positive Predictive Value

3 families of criteria according to the dependance between A , \hat{Y} and Y

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

Independence

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

Separation

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

Sufficiency

Definition - Statistical Parity

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

The probability of a positive decision is identical in each group.

Advantages

- Easy to measure and impose as a constraint
- Corresponds to fair representation

Limitations

- Ignores accuracy
 - a random classifier can satisfy it
- Can degrade both groups simultaneously

Ex: If 30% of male candidates are selected, demographic parity also requires 30% of female candidates to be selected regardless of qualifications.

Equal Opportunity (Hardt et al. 2016)

$$TPR_0 = TPR_1$$

Among truly positive individuals, the identification rate is equal.

Equalised Odds (Hardt et al. 2016)

$$TPR_0 = TPR_1 \quad \text{and} \quad FPR_0 = FPR_1$$

The classifier makes the same types of errors in both groups.

Ex: Among *truly qualified* candidates, the proportion selected is the same regardless of gender (*equal opportunity*). *Equalized odds* additionally requires the same false positive rate.

Predictive Parity

$$PPV_0 = PPV_1$$

The positive predictive value (PPV, or precision) is equal in both groups.

Calibration

For any score $s \in [0, 1]$:

$$P(Y = 1 | \hat{P} = s, A = 0) = P(Y = 1 | \hat{P} = s, A = 1) = s$$

The score has the same probabilistic meaning for all groups.

Chouldechova (2017), Kleinberg, Mullainathan & Raghavan (2016)

Unless $p_0 = p_1$, no non-trivial classifier can simultaneously satisfy:

$$TPR_0 = TPR_1$$

$$FPR_0 = FPR_1$$

$$PPV_0 = PPV_1$$

Proof Sketch (Chouldechova). We have the exact identity:

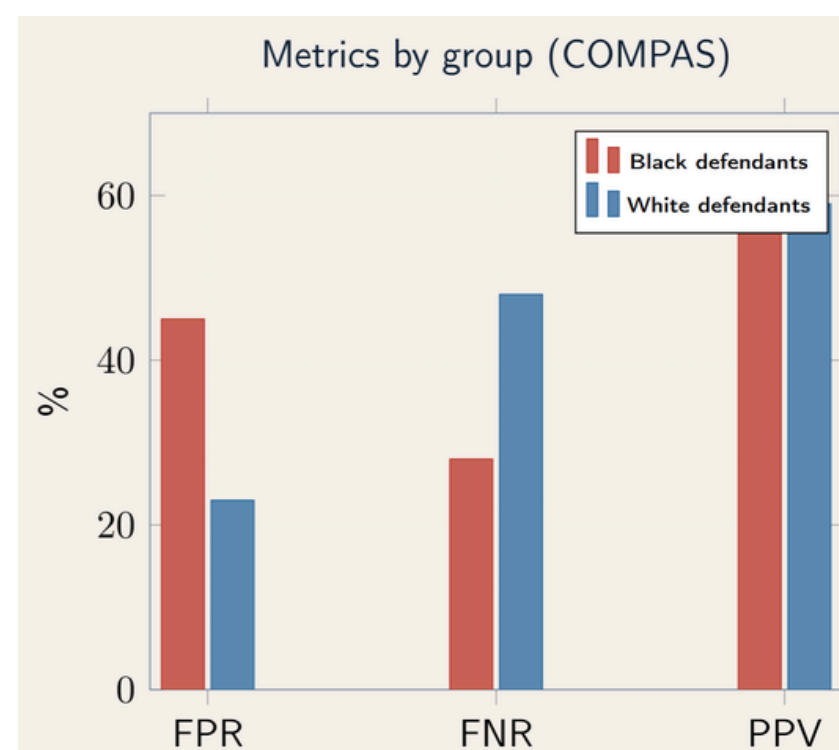
$$\frac{1 - PPV_a}{PPV_a} = \frac{1 - p_a}{p_a} \cdot \frac{FPR_a}{TPR_a}$$

If $PPV_0 = PPV_1$ and $p_0 \neq p_1$, the ratio $\frac{FPR_a}{TPR_a}$ differs between groups \rightarrow simultaneous equalisation of FPR and TPR is impossible

CCL: Choosing a fairness criterion is a **political and ethical choice** (not a technical calculation) \rightarrow **it reflects the values one decides to prioritize.**

Context

- Recidivism scores (1-10), Broward County (FL), used for bail release decisions
- 2016: ProPublica reveals racial bias (Angwin et al.)
- Base rates: $p_{Black} \approx 51\%$, $p_{White} \approx 39\%$



Source: Angwin et al. (2016).

The debate

- ProPublica: $FPR_{Black} \approx 45\%$ vs 23%
→ violation of equalized odds

Taxonomy (Barocas, Hardt & Narayanan, 2023)

Three intervention stages in the ML pipeline

1. Pre-processing (on data)

- Resampling
- Group reweighting
- Fair embeddings (adversarial)

2. In-processing (during training)

- Constraints in the loss
- Adversarial debiasing
 - (Zhang et al., 2018)
- Differential regularization

3. Post-processing (on predictions)

- Group-specific thresholds
- Exact eq. odds solution (Hardt 2016)
- Specific calibration

Fairness-performance trade-off

Any fairness constraint **restricts the hypothesis space**, inducing a loss in AUC or overall accuracy.

→ An inevitable trade-off.

Open-source libraries

- **AI Fairness 360 (IBM)**
 - 70+metrics, 11 debiasing algorithms
- **Fairlearn (Microsoft)**
 - post-training constraints, dashboard
- **What-If Tool (Google)**
 - visual exploration
- **Aequitas (UChicago)**
 - multi-group audit

Open questions

- Counterfactual *causal* fairness (Pearl)?
 - a decision is fair towards an individual if the outcome is the same in reality as it would be in a *counterfactual* world, in which the individual belongs to a different demographic
- Fairness in LLMs?
- Who is responsible for algorithmic decisions?



1. Algorithmic decisions have **real** and **regulated societal impact**
2. Three families
 - a. **Independence**
 - b. **Separation**
 - c. **Sufficiency**
3. The **impossibility theorem** guarantees their **general incompatibility**
4. Choosing a criterion is a **political** and **ethical** choice
5. Any **mitigation** method implies a **trade-off**

Key references: Chouldechova (2017), Hardt et al. (2016), Dwork et al. (2012), Barocas & Hardt (2023), EU AI Act (2024)

- **Angwin et al.** (2016). Machine Bias. *ProPublica*.
- **Chouldechova, A.** (2017). Fair prediction with disparate impact. *Big Data*, 5(2).
- **Dwork, C. et al.** (2012). Fairness through awareness. *ITCS 2012*.
- **Hardt, M., Price, E., Srebro, N.** (2016). Equality of opportunity in supervised learning. *NeurIPS*.
- **Kleinberg, J., Mullainathan, S., Raghavan, M.** (2016/2017). Inherent trade-offs in the fair determination of risk scores. *ITCS*.
- **Barocas, S., Hardt, M., Narayanan, A.** (2023). *Fairness and Machine Learning*. MIT Press. fairmlbook.org
- **Zhang, B. H. et al.** (2018). Mitigating unwanted biases with adversarial learning. *AIES*.
- **HLEG-AI** (2019). *Ethics Guidelines for Trustworthy AI*. European Commission.
- **EU AI Act** (2024). Regulation (EU) 2024/1689.
- **NIST AI RMF** (2023). Artificial Intelligence Risk Management Framework.