

Socio-Economic Parity in AI Systems

Eirini Ntoutsi

Based primarily on Roy et al., ACM FaccT 2025

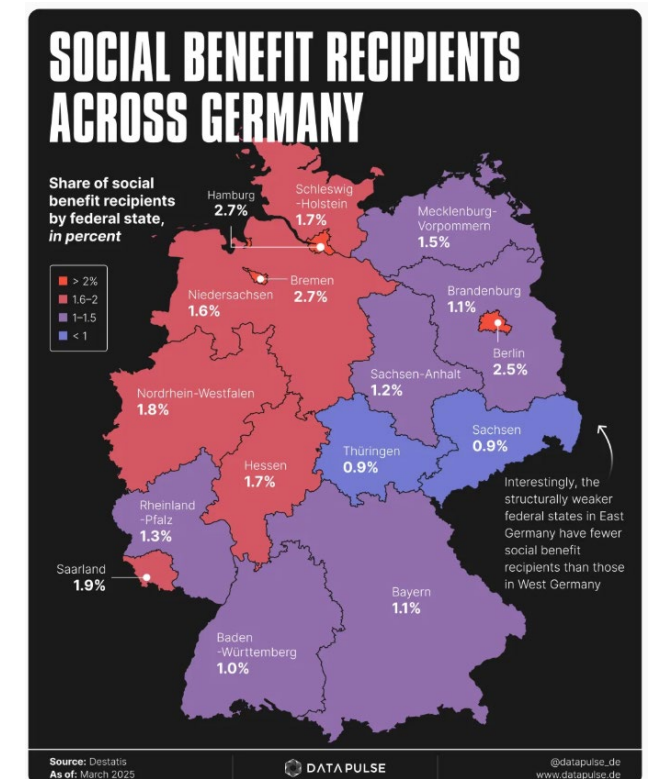
Agenda

- **Part 1:** Context and foundations of fairness-aware learning
- **Part 2:** Socio-economic fairness in AI systems

Motivation

AI systems increasingly shape high-stake domains

- **Finance**: credit scoring, loan approval
- **Employment**: hiring, promotion
- **Education**: admissions
- **Justice**: risk assessment (e.g., recidivism prediction)
- **Public services**: welfare allocation



Fairness matters!

Unfair AI systems can cause real-world harm

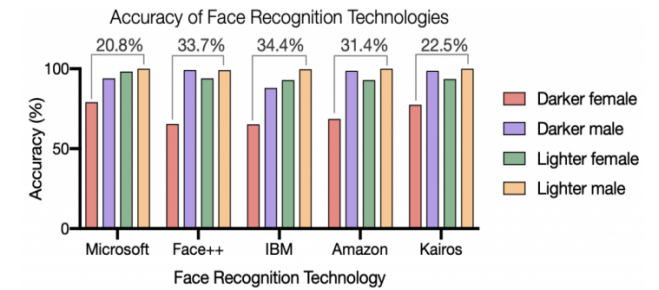
- **Allocative harms**

- When decision-making systems in criminal justice, health care, etc. are discriminatory, they create allocative harms, which are caused when a system withholds certain groups an opportunity or a resource.

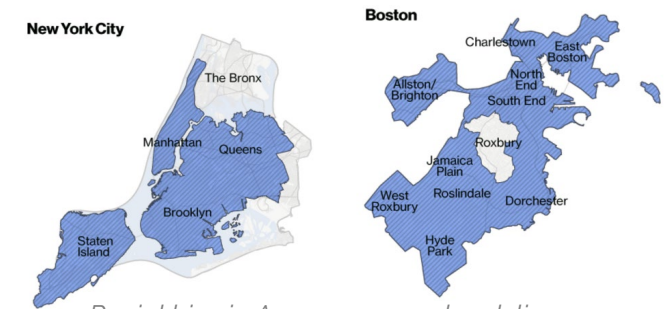
- **Representational harms**

- When systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc., they create stereotype perpetuation and cultural denigration.

K. Crawford (2017). *The Trouble with Bias*, NIPS 2017 Keynote



Gender bias in face recognition technologies.
The Gender Shades



Racial bias in Amazon same day delivery service provision. *Bloomberg*

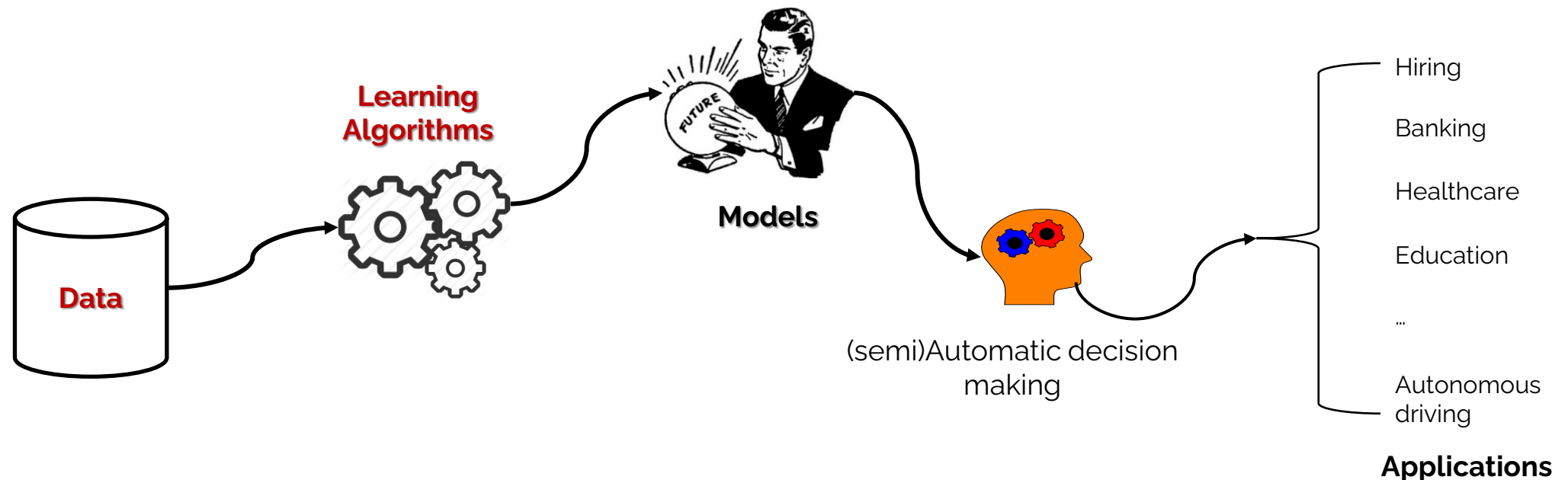


AI-generated images when asked to draw photorealistic scientists. *Lisa M.P. Munoz, SciComm*

Why can AI systems discriminate?

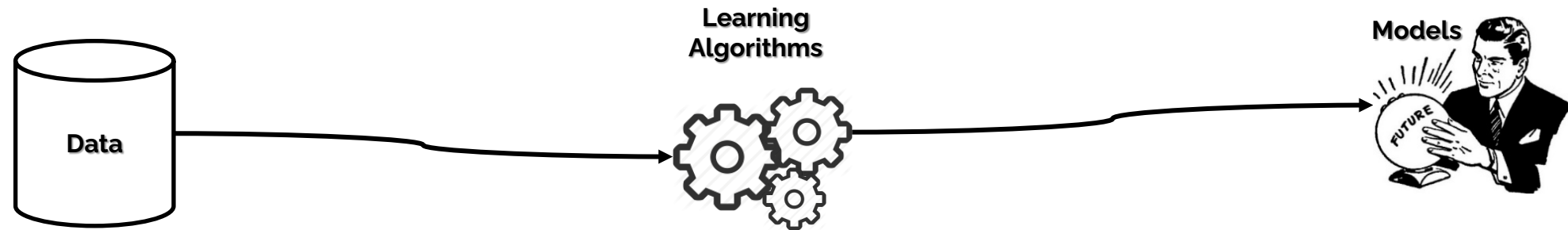
Understanding the structural roots of bias in AI → back to basics on how machines learn

- ML “gives computers the ability to learn without being explicitly programmed” (Arthur Samuel, 1959)
 - We don't codify the solution. We may not even know it!
- **Data** & the **learning algorithms** jointly shape model behavior



Where does bias come from in AI?

Is rarely “coded-in”, it emerges from data, objectives and models



Data is not neutral

AI-systems rely on **data generated by humans** or **collected via systems designed by humans**.

As a result, human biases:

- **enter** these systems through design, usage, and labeling.
- can be **amplified** by complex sociotechnical systems, such as the Web.
- can be **reinforced** through feedback loops and pipelines.

Learning algorithms ignore fairness

Optimize performance objectives such as:

- **Accuracy** in predictive tasks
- **Reconstruction error** in generative tasks

Fairness is not part of the learning objectives

- It is not encoded in standard loss functions.
- Group-level disparities are **neither measured nor reported**

Models exploit shortcuts & proxies

Models often rely on “**shortcuts**”: **quick-to-learn patterns** (e.g., spurious correlations or easy to learn features not related to the task semantics) that optimize objectives

Shortcuts can be wrong:

- A wolf detector learns snow instead of wolf
- A hiring model prefers male candidates via proxies

Proxy attributes: Attributes that correlate with protected characteristics

- Zip code for race, name/hobbies for gender

These shortcuts are not explicitly programmed; they **emerge from data**.

What do we mean by fairness

Operational definitions of fairness

- **Group fairness**

- similar outcomes/performance across demographic groups (e.g., women vs. men)

- **Example measures**

- Demographic (or statistical) parity (DP)
- Equal opportunity (EP)
- Equalized odds
- Conditional statistical parity (CDP)
- Treatment equality
- Test fairness

DP: no difference in **positive prediction rates (PPRs)** across groups

CDP: No difference in **PPRs** across groups within the same contextual group (e.g., occupation)

EP: No difference in **true positive rates (TPRs)** across groups.

- **Individual fairness**

- similar people should be treated similarly

- **Example measures**

- Fairness through awareness
- Fairness through unawareness
- Counterfactual fairness

- **Other definitions**

- beyond supervised learning tasks, e.g., based on diversity and coverage

Protected vs non-protected attributes

- **Protected:** sensitive attributes (e.g., gender, race, age) → **direct discrimination**
- **Non-protected:** other features
 - may still act as proxies for protected attributes (e.g., zipcode → race) → **indirect discrimination**

Impossibility of fairness

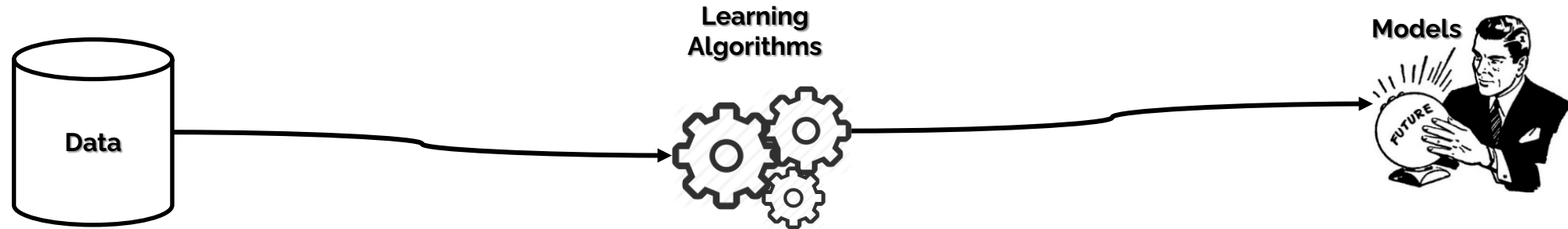
Fairness in ML involves both mathematical and sociotechnical trade-offs

- **Mathematical impossibility of fairness** ([Kleinberg et al, 2017](#); [Chouldechova, 2017](#))
 - (Some) fairness metrics are **mutually incompatible** and cannot be satisfied simultaneously (except in trivial cases)
 - Trade-offs are inevitable (improving one may harm another)
 - ➔ We must choose which fairness definition to prioritize based on context and goals.
- **Conceptual impossibility** ([Selbst et al., 2019](#))
 - Formal fairness definitions require **abstraction** and **simplification**.
 - But fairness is socially situated, it depends on context, history, power, and values.
 - ➔ No definition is value neutral or universally correct



Mitigating unfairness

A broad spectrum of interventions across the AI pipeline



Pre-processing approaches

Intuition: Making the data “fairer” will result in a “less unfair” model

Core idea: “Balance” the representation of protected and non-protected groups in the dataset

Design principle: Use **minimal data interventions** to preserve data utility for the learning task

Example techniques:

- Sampling
- Class rebalancing (massaging)
- Synthetic data generation

In-processing approaches

Intuition: Working directly with the algorithm offers greater control over fairness behavior

Core idea: Explicitly incorporate fairness objectives into the learning process

Design principle: “**Balance**” predictive- and fairness-performance

Example techniques:

- Fairness constraints
- Regularization
- In-training altering of data distribution
- Adversarial debiasing

Post-processing approaches

Intuition: Start with predictive performance

Core idea: Apply fairness adjustments after training the model, no changes to the data or learning algorithm.

Design principle: **Minimal interventions** to preserve predictive performance

Example techniques:

- Shift decision boundary
- Calibration
- Relabel tree leaf nodes
- Wrap a fair classifier on top

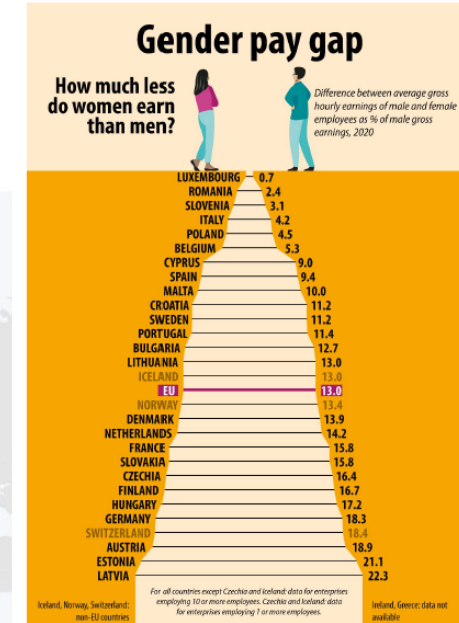
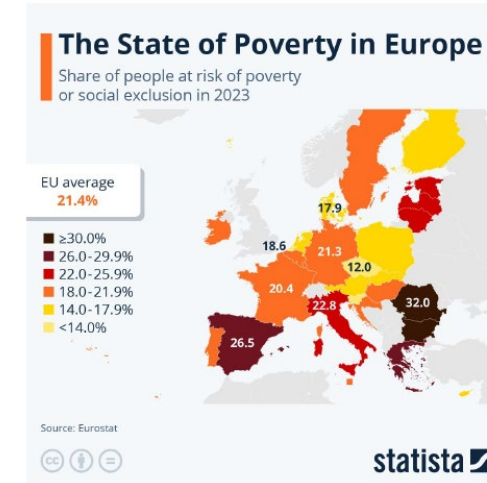
Agenda

- **Part 1:** Context and foundations of fairness-aware learning
- **Part 2:** Socio-economic fairness in AI systems

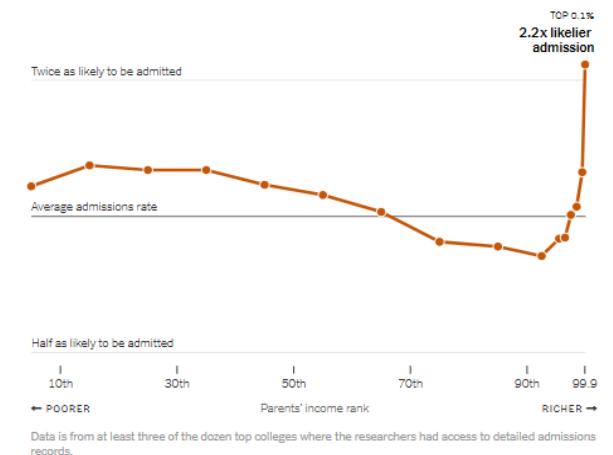
Socio-economic status (SES) and inequality

- **Socio-economic status (SES)** refers to an individual's or group's social and economic position within society.
 - Common indicators include income, wealth, education, and occupation.
- **SES and discrimination are structurally intertwined**
 - EU gender pay gap (2020): 13% (reduction of only 2.8% over a decade)
 - Poverty and social exclusion remain unevenly distributed across Europe.
 - Access to elite education remains strongly associated with wealth and family background "being very rich is its own qualification"
 - Lower SES is associated with higher rates of chronic disease, disability, and adverse health outcomes (Robert Koch Institute)

These structural inequalities may also be **learned and reproduced by AI** systems.



Admissions rate at elite colleges among students with the same test scores



[Source](#)

SES in the EU AI Act

- The EU AI Act adopts a **risk-based framework** for regulating AI systems
 - including high-risk domains such as credit scoring, employment, education.
 - Requirements include bias detection, representative datasets, and Fundamental Rights Impact Assessments (FRIAs).
- These domains are **strongly linked to socio-economic disadvantage**
 - e.g., credit scoring, social benefits, access to essential services
- However, **SES is not explicitly recognized as a protected ground in EU law**
 - although 10 EU Member States incorporate SES-related concepts (e.g., social origin, economic vulnerability) into their national anti-discrimination laws.
- The AI Act requires fairness and bias mitigation but **does not specify how fairness should be operationalized technically**.
- **Existing fairness notions** often compare demographic groups while overlooking SES disparities within these groups.

Core gap: AI Act **recognizes discriminatory risks in SES-sensitive domains**, but current fairness notions largely ignore SES.

Experimental setup

How do existing fairness notions behave under SES?

- **Adult dataset¹**

- US census data
- Prediction task: income > 50 K vs. ≤ 50K
- Protected attribute: sex (binary)
- SES-related proxy attributes
 - Capital gain
 - Occupation
 - Hours per week

- **Compared fairness-aware classifiers**

- DP-aware model
- EP-aware model
- CDP-aware model

- All methods rely an in-processing fairness-aware classifier.

TABLE 2 Adult: attributes characteristics

Attributes	Type	Values	#Missing values	Description
age	Numerical	[17-90]	0	The age of an individual
workclass	Categorical	7	2,799	The employment status (private, state-gov, etc.)
fnlwgt	Numerical	[13,492-1,490,400]	0	The final weight
education	Categorical	16	0	The highest level of education
educational-num	Numerical	1-16	0	The highest level of education achieved in numerical form
marital-status	Categorical	7	0	The marital status
occupation	Categorical	14	2,809	The general type of occupation
relationship	Categorical	6	0	Represents what this individual is relative to others
race	Categorical	5	0	Race
sex	Binary	{Male, Female}	9	The biological sex of the individual
capital-gain	Numerical	[0-99,999]	0	The capital gains for an individual
capital-loss	Numerical	[0-4,356]	0	The capital loss for an individual
hours-per-week	Numerical	[1-99]	0	The hours an individual has reported to work per week
native-country	Categorical	41	857	The country of origin for an individual
income	Binary	{≤50K, >50K}	0	Whether or not an individual makes more than \$50,000 annually

Source: Le Quy et al, 2022.

¹ <https://archive.ics.uci.edu/ml/datasets/adult>

Ground truth: Imbalance

PPR (ground truth): % of positive instances (>50K)

- Ground truth distribution of positive (>50K) and negative (≤50K) **instances** for females and males by occupation.
 - Occupations are sorted from low to high number of positive labels.



Observations

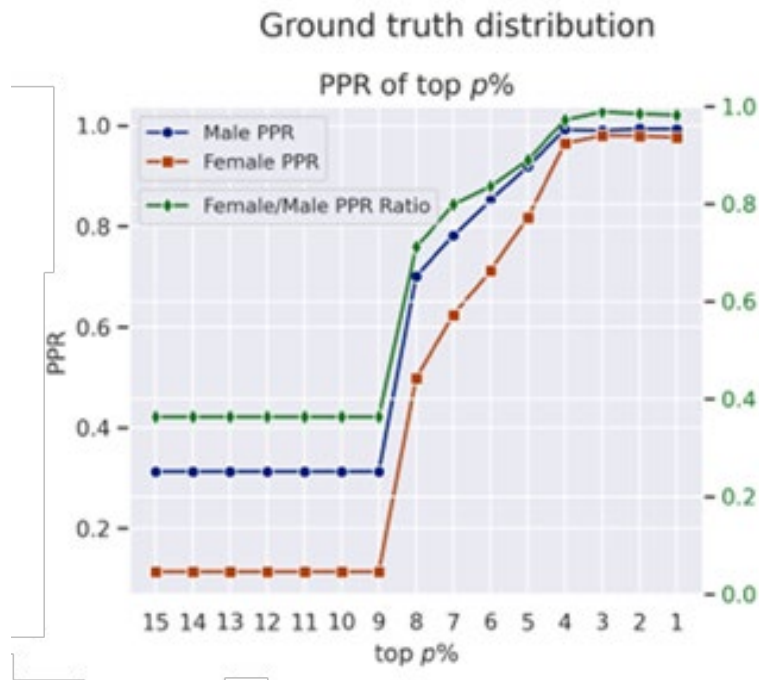
- **Strong class imbalance in the ground truth:** males receive nearly **3x more positive outcomes** (>50K) than females.
- **Occupation-dependent disparities:** some occupations exhibit much stronger class imbalance than others.

The ground truth is strongly imbalanced, and disparities are non-uniform across occupations.

Ground truth: Socio-economic privilege effect

PPR (ground truth): % of positive instances (>50K)

- Ground truth PPR for females and males across increasingly privileged subgroups
 - **Capital gain** is used as a proxy for SES privilege.
 - **Top p%** refers to individuals with the highest capital gain.



Observations

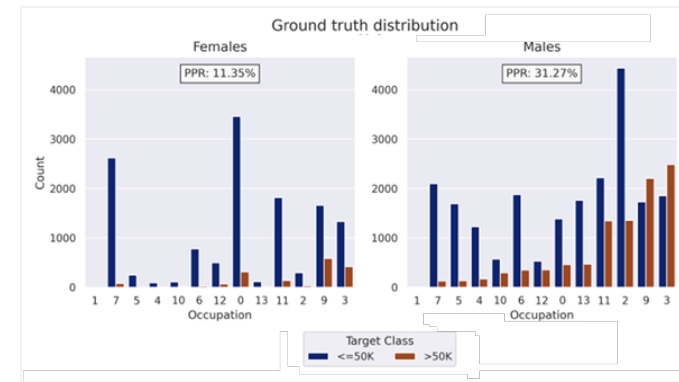
- **PPRs increase sharply with SES privilege:** both male and female PPRs rise as capital gain increases.
- **Gender disparities narrow among highly privileged groups:** female/male PPR ratio exceeds 80% within the top 5% capital-gain group.
- **Disparities remains strong outside the privileged groups:** for most of the population, the female/male ratio stays below 40%.

SES-privilege strongly shapes favorable outcomes in the ground truth.

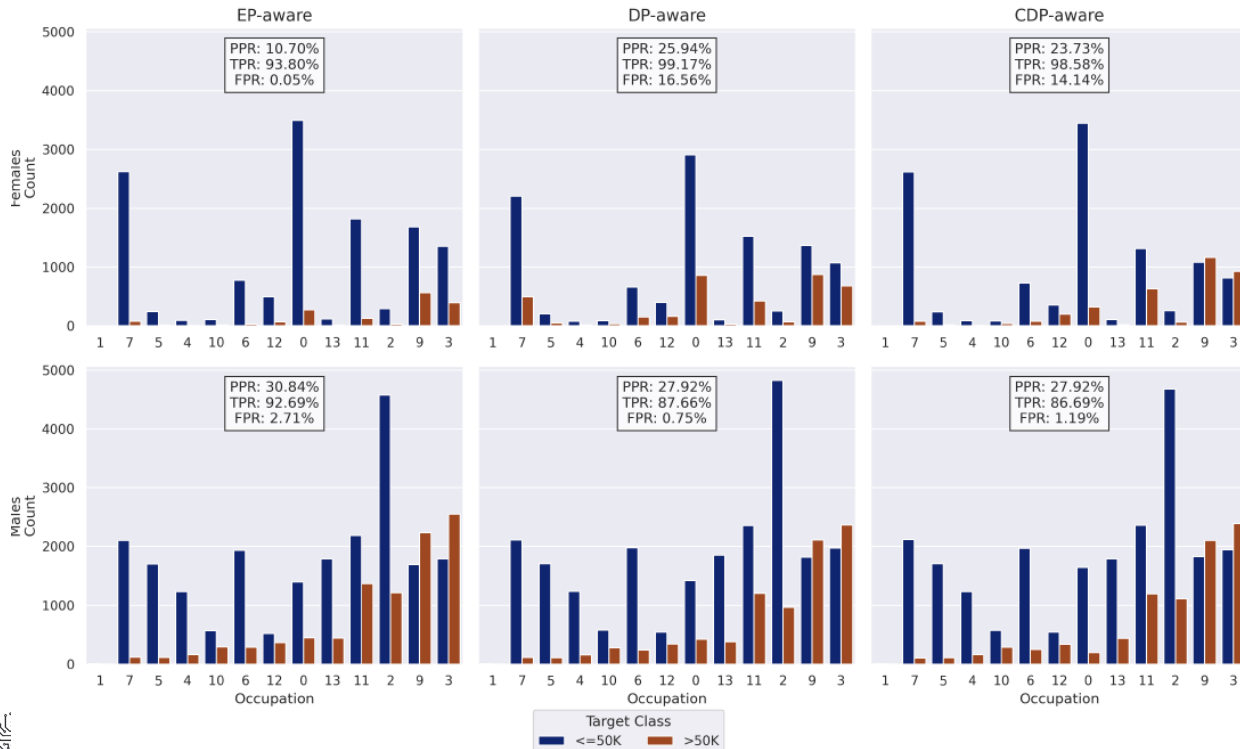
Existing fairness notions and fairness-aware classifiers

PPR (predictions): % of positive predictions (>50K)

- Distribution of positive (>50K) and negative ($\leq 50K$) **predictions** for females and males by occupation, as predicted by EP-aware, DP-aware, and CDP-aware classifiers.
 - Occupations are sorted from low to high number of positive (ground-truth) labels.



Predictions by EP-aware, DP-aware, and CDP-aware classifiers



Observations

- **EP**: preserves the original demographic imbalance
 - male PPR $\gg$ female PPR.
- **DP**: equalizes overall PPRs across males and females
 - $\rightarrow$ broad redistribution of positive predictions (FPR)
- **CDP**: produces more balanced occupation-level allocations.
 - $\rightarrow$ context-aware redistribution of positive predictions

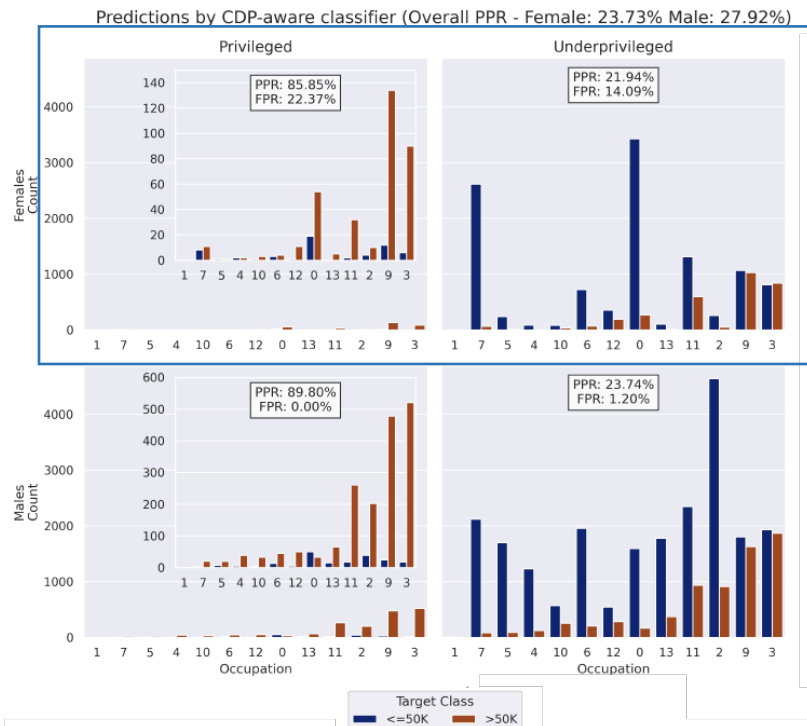
Fairness-induced increases in FPR can be interpreted as a proxy for positive actions, i.e., additional favorable allocations for the protected group

Existing methods mitigate demographic (male-female) disparities, but what about SES-privilege?

CDP under SES

Does the strongest baseline (CDP) address SES inequality?

- **CDP-aware predictions** across privileged and underprivileged groups for females and males by occupation
 - **Privileged:** individuals **within the top 5%** capital-gain group.



Observations

- **Privileged groups continue to receive substantially higher PPRs** than underprivileged groups.
- **Fairness-induced increases in favorable allocations (reflected in higher FPRs)** are substantially larger for privileged females, indicating that privilege still dominates allocation.

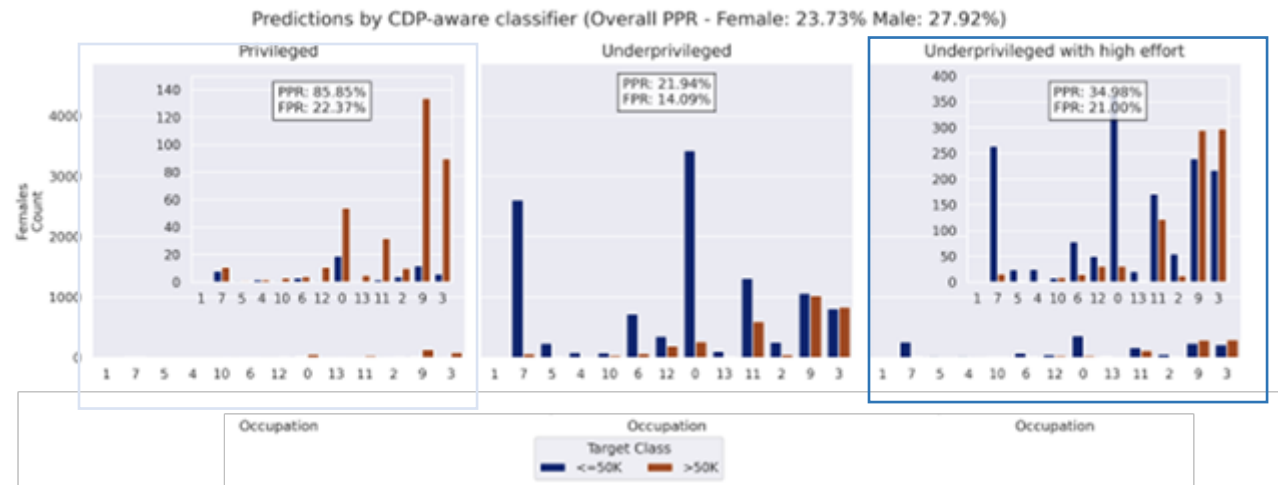
Who needs a positive action?

Fairness-interventions should prioritize **SES-underprivileged individuals within protected groups** rather than disproportionately benefiting already privileged individuals.

An effort-aware perspective under CDP

Should positive actions also account for effort?

- **CDP-aware predictions** for underprivileged individuals with high effort
 - **Underprivileged with high effort:** underprivileged individuals with working hours above the demographic mean
 - Working hours used as a **proxy for individual effort**



Observations

- **High-effort underprivileged individuals** still receive substantially fewer favorable outcomes than privileged groups.
- **Comparable FPRs:** Positive actions are not sufficiently aligned with both need and effort.
- Underprivileged females **remain disadvantaged even with high effort.**

Who needs a positive action?

Fairness-aware allocation should account for **both SES-disadvantage and individual effort.**

Towards SES-aware fairness

- Existing fairness notions primarily target **demographic disparities** (e.g., gender fairness).
- **SES-disparities can persist within protected groups, even under fairness-aware learning.**
 - Fairness-aware redistributions may disproportionately benefit already privileged subgroup members.
- Fairness notions may therefore need to account for:
 - SES-disadvantage
 - Individual effort
 - contextual structure
- **Socio-Economic Parity (SEP)** and **Conditional SEP (CSEP)** represent steps toward SES-aware fairness-aware learning.
 - extends SEP by enforcing SES-aware fairness within contextual groups (e.g., occupations).

Socio-Economic Parity (SEP)

- **Socio-economic parity (SEP)**

Favorable outcomes should not be hindered for socio-economically underprivileged individuals.

Underprivileged individuals who make greater effort (above a threshold) should receive rewards proportional to their effort.

Positive action for privileged groups should be minimized to reduce the influence of privilege.

DEFINITION 1. A predictor h is said to behold Socio-Economic Parity (SEP) fairness if for any given economically privileged group defined as $(X_p \geq \tau_p)$, and underprivileged sub-groups defined as $(X_p < \tau_p, S = s)$ by a socio-economic attribute X_p , a privilege threshold τ_p , a protected attribute S , and demographic identity s , it holds:

$$\mathbb{E}(h(x)) = \mathbb{E}(h(x) | s, x_p < \tau_p), \quad \forall (x, s, y) \in D; \zeta : \mathbb{R} \rightarrow [1, \infty)$$

such that:

$$\mathbb{E}(h(x) \cdot \zeta(x_e) | x_p < \tau_p, x_e \geq e) \geq \mathbb{E}(h(x) | x_p < \tau_p, x_e < e),$$

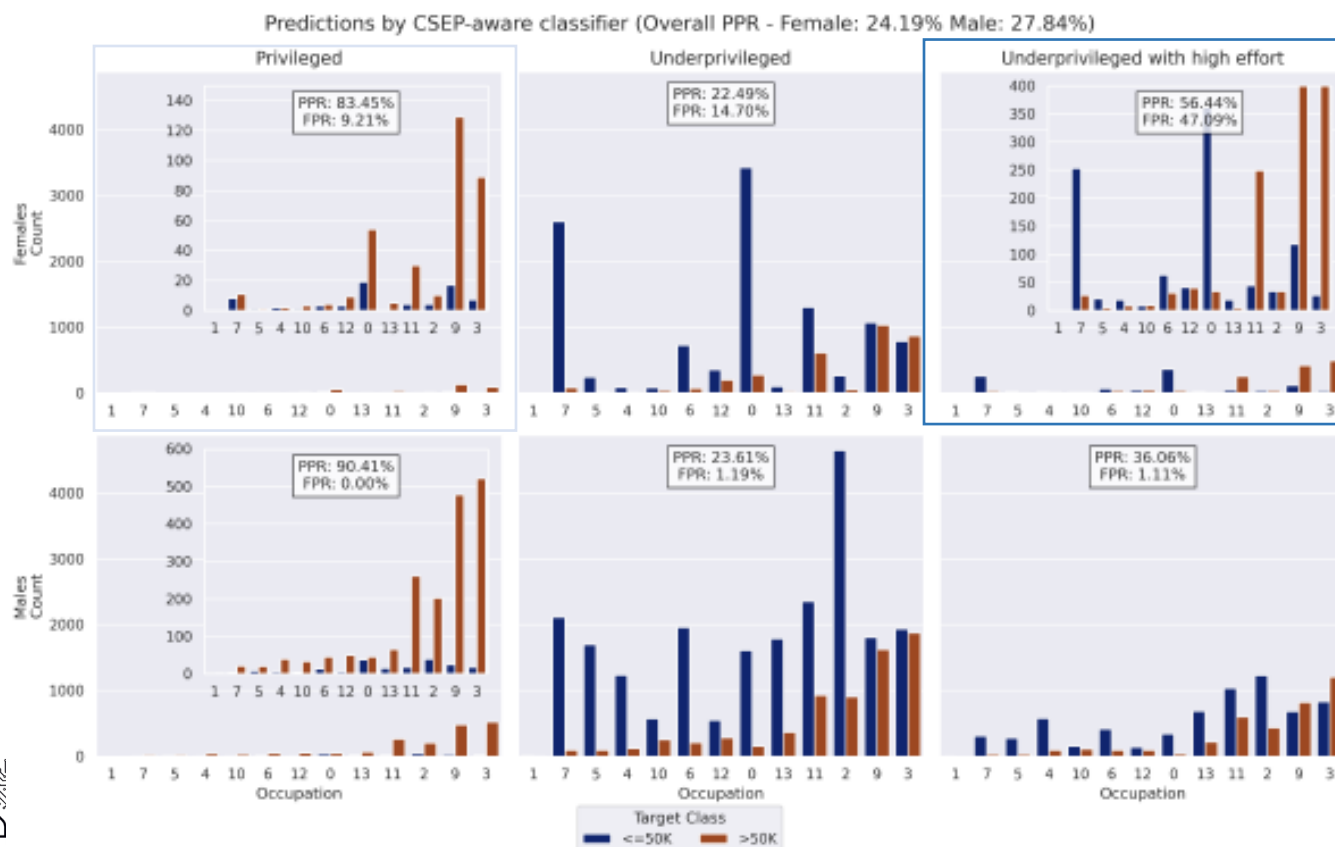
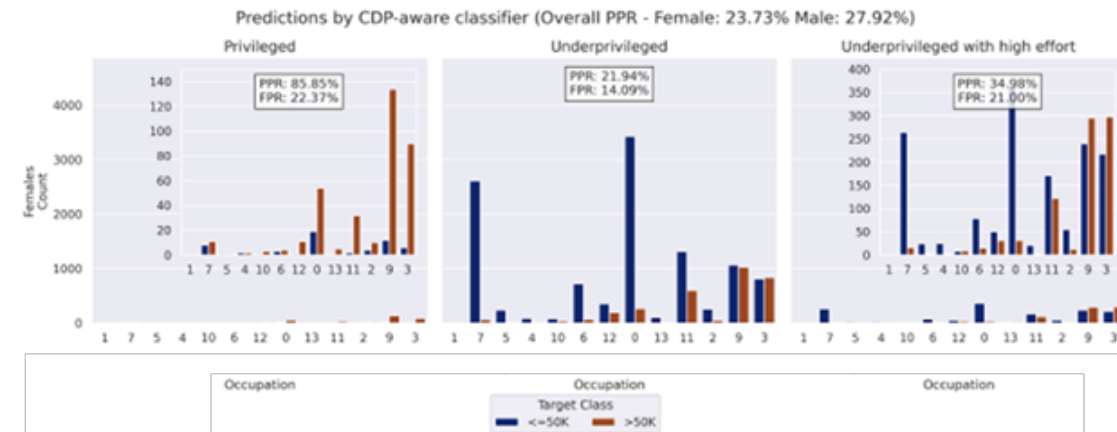
$$\mathbb{E}(h(x) | x_p \geq \tau_p, y = 0) \leq \mathbb{E}(h(x) \cdot \zeta(x_e) | x_p < \tau_p, x_e \geq e, y = 0)$$

- **Conditional Socio-economic parity (CSEP)**

- extends SEP by enforcing SES-aware fairness within contextual groups (e.g., occupations)

Empirical evaluation of CSEP

- Does CSEP improve allocation for underprivileged high-effort protected groups?

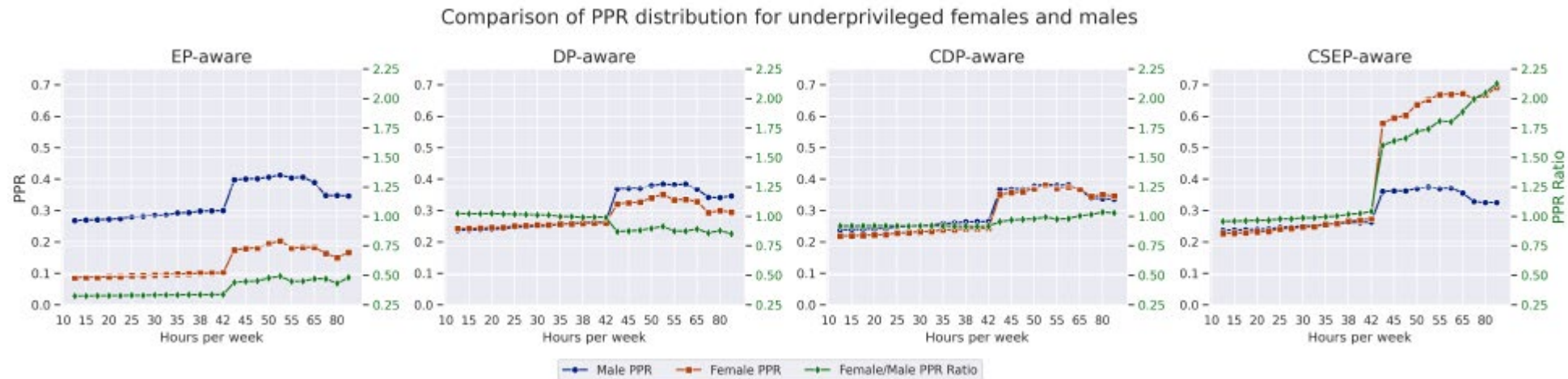


Observations

- CSEP reallocates positive outcomes **towards SES-underprivileged high-effort females**.
- Compared to CDP, CSEP **better aligns fairness-aware redistribution with SES-disadvantage and effort**.

Existing fairness-notions vs. CSEP under effort

PPR distributions across increasing effort levels for underprivileged males and females.



Observations

- **EP:** largely preserves the original imbalance, despite increasing effort
- **DP:** improves demographic fairness, but redistribution is broad rather than effort-aware
- **CDP:** improves contextual fairness across occupations, yet weakly accounts for effort
- **CSEP:** explicitly aligns positive allocations with both disadvantage and effort

CSEP is the only approach that explicitly aligns positive allocations with both SES disadvantage and effort.

Wrapping up

- **SES and discrimination are deeply interconnected.**
- The EU AI Act raises concerns about SES-related discriminatory risks, yet current fairness notions weakly operationalize SES.
- Existing fairness-aware learning methods operationalize fairness primarily **at the demographic-group level**, and insufficiently account for SES-privilege, and effort within protected groups.
 - Even CDP fail under SES-sensitive settings
- SEP/CSEP **explicitly incorporate socio-economic disadvantage and effort** into fairness-aware allocation.
 - Fairness-aware redistributions can therefore be redirected toward socio-economically underprivileged and high-effort protected groups.
- Our approach relies on proxy variables and threshold-based definitions
 - capital gain → SES privilege
 - hours-per-week → effort
 - top 5% capital-gain group → privileged subgroup
 - working hours above the demographic mean → high-effort individuals
- **Beyond SES:** We need targeted fairness interventions rather than broad redistribution strategies.



Thank you for your attention! I look forward to the exchange

Prof. Dr. Eirini Ntoutsis

Chair for Open Source Intelligence
Artificial Intelligence & Machine Learning (AIML) Lab
Institute for Data Security INF7
Research Institute CODE & UniBw Munich

Email: eirini.ntoutsis@unibw.de
[www](http://www.unibw.de/aiml); <https://www.unibw.de/aiml> || <https://aiml-research.github.io/>

