Achieving Socio-Economic Parity through the Lens of EU AI Act

Arjun Roy*
arjun.roy@unibw.de
Dept. Mathematics & CSc., Freie Universität Berlin
RI CODE, Universität der Bundeswehr München
Germany

Stavroula Rizou*
rizstavroula@iti.gr
Information Technologies Institute, CERTH
Greece

Symeon Papadopoulos
papadop@iti.gr
Information Technologies Institute, CERTH
Greece

Eirini Ntoutsi
eirini.ntoutsi@unibw.de
RI CODE, Universität der Bundeswehr München
Germany

Abstract

Unfair treatment and discrimination are critical ethical concerns in AI systems, particularly as their adoption expands across diverse domains. Addressing these challenges, the recent introduction of the EU AI Act establishes a unified legal framework to ensure legal certainty for AI innovation and investment while safeguarding public interests, such as health, safety, fundamental rights, democracy, and the rule of law (Recital 8). The Act encourages stakeholders to initiate dialogue based on existing AI fairness notions to address discriminatory outcomes of AI systems. However, these fairness notions often overlook the critical role of Socio-Economic Status (SES) and fail to capture the compounded effects of economic privilege, inadvertently perpetuating biases that favour the economically advantaged. This oversight is particularly concerning given that the principles of equalization advocate for equalizing resources or opportunities to mitigate disadvantages beyond an individual's control. While provisions for discrimination are laid down in the AI Act, specialized directions should be broadened, particularly in addressing economic disparities perpetuated by AI systems. In this work, we explore the limitations of popular AI fairness notions using a real-world dataset (Adult), highlighting their limitations, particularly their inability to address SES-driven disparities. To address this gap, we propose a novel fairness notion, Socio-Economic Parity (SEP), which incorporates SES and promotes positive actions for underprivileged groups while accounting for factors within individual's control, such as working hours, which can serve as a proxy for effort. We define a corresponding fairness measure based on this notion, and optimize a model constrained by SEP to demonstrate its practical utility. Our empirical results demonstrate the effectiveness of our approach in mitigating SES-driven biases. By analyzing the AI Act with our method, we lay a foundation for aligning AI fairness with SES factors while ensuring legal compliance in AI decisions.

CCS Concepts

• **Applied computing → Law, social and behavioral sciences**.

Keywords

AI fairness, AI Act, Socio-Economic Status

1 INTRODUCTION

> "*We seem to be back in the Dickensian world of Hard Times, with the haves facing off against the increasingly alienated have-nots, with no resolution in sight*" [7]

Unfair treatment and discrimination [17, 39] in AI systems — often described as AI bias or unfairness — are widely recognized as significant ethical concerns [47, 54, 55]. The pervasive integration of AI across various societal and economic domains has amplified the urgency of addressing these issues. This is further underscored by the increasing global adoption of AI-related legal frameworks [51], including the landmark EU AI Act proposed in 2021, which catalyzed a 1,200% increase[1] in regulatory activities in the following year. While the Act provides a comprehensive framework for addressing discrimination, its scope should be broadened to include specialized directions targeting economic disparities perpetuated by AI systems [33]. These disparities are deeply rooted in Socio-Economic Status (SES), which disproportionately affects historically marginalized groups [33]. For instance, individuals subjected to discrimination based on traditional grounds, such as sex and race, are often overrepresented among economically disadvantaged populations [32]. This intersectionality of SES with existing vulnerabilities further exacerbates systemic inequities [6], violating principles of equalization [24], which advocate for equalizing resources or opportunities to mitigate disadvantages beyond an individual's control. Notably, at the EU level, the pay gap between men and women was

*Both authors contributed equally to this research.

[1]Based on public data for 2021 and 2022 from the AI Index 2024 Annual Report by Stanford University [2].

13% in 2020 and has decreased by only 2.8 percentage points over a decade [16].

Despite advancements in AI fairness metrics [8, 14], most fail to capture the compounded effects of socio-economic privilege. This limitation creates critical gaps in addressing SES-driven disparities that produce unfair advantages across protected attributes such as sex [42] or race [44, 70]. To bridge this gap, we propose a novel fairness notion, Socio-Economic Parity (SEP), along with its stricter variant, Conditional Socio-Economic Parity (CSEP), explicitly incorporating SES into fairness considerations. Our contributions include introducing SEP and CSEP, demonstrating their effectiveness in mitigating SES-driven bias through evaluations on the Adult dataset [10], and analyzing their alignment with the AI Act to bridge fairness-aware ML research with regulatory frameworks. Using this approach, we emphasize positive actions for underprivileged groups while accounting for individual-level factors [13] (e.g., working hours). We demonstrate that existing fairness notions—such as Equal Opportunity [38], Demographic Parity [23], and Conditional Demographic Parity (CDP) [20]—fail to address SES-driven disparities and inadvertently reinforce biases favouring economically privileged groups. Our empirical results highlight the effectiveness of SEP and CSEP in mitigating SES-driven bias. Furthermore, through an analytical study of the AI Act in parallel with our proposed methods, we establish a foundation for fostering dialogues that align AI fairness in socio-economic applications, emphasizing the importance of addressing SES-driven disparities.

The remainder of the paper is organized as follows: We start by presenting the investigation over the protected grounds in the context of SES under EU law (Sec. 2). Sec. 3 contextualizes socio-economic disparities through the existing fairness definitions, underlying challenges, and the driving motivations and objectives behind addressing them. Sec. 4 outlines the proposed solution, focusing on key variables, and thresholds for fairness tests. Sec. 5 examines the proposed notion through justifications, drawing on the AI Act and outlining also the legal framework of anti-discrimination law, which the regulation assigns with addressing specific issues related to AI bias. Sec. 6 discusses the implementation of the assessments of the AI Act through fairness measures. Sec. 7 concludes the paper with challenges and limitations.

2 PROTECTED, NON-PROTECTED GROUNDS AND PROXIES IN THE CONTEXT OF SES UNDER EU ANTI-DISCRIMINATION LAW

The investigation of protected grounds[2], which can affect the identification of various vulnerable groups under SES, according to EU non-discrimination law, is based on Article 2 of the Treaty on European Union (TEU) and Articles 10 and 19[3] of the Treaty on the Functioning of the European Union (TFEU), while EU secondary law prohibits discrimination in relevant adopted Directives[4] on

defined grounds [71]. Additionally, Article 21 of the EU Charter of Fundamental Rights provides a broader, non-exhaustive list[5]. However, it should be mentioned that this extended list in the Charter is limited solely to discrimination by the institutions and bodies of the EU, when exercising powers conferred under the Treaties and by Member States only when they are implementing Union law [43], indicating its limited scope when also being examined in the context of algorithmic fairness. In parallel, the European Convention on Human Rights (ECHR) law of the Council of Europe also plays a key role on the protected grounds under Article 14[6] in the EU environment, as the Treaty of Lisbon permitted the EU to accede to the ECHR [22, 56]. The non-exhaustive list of Article 14 does not include SES as a protected ground either, although, it allows potential expansion by the European Court of Human Rights [30, 67]. Nevertheless, it should be noted that the European Court of Human Rights receives claims from victims of rights violations by States Parties to the Convention, with only a potential indirect effect on private actors [34].

Therefore, it becomes clear that SES is not considered a protected ground under EU law. It is nevertheless protected in the national law of certain EU Member States [71]. Namely, 10[7] of the 27 EU Member States, according to the 2019 comparative analysis of non-discrimination law in Europe [19], incorporate grounds related to SES into their anti-discrimination legislation with relative terminology as follows: social status, property, economic vulnerability, social origin, social standing, economic situation, profession and part-time employment. Apart from the investigation of protected grounds in the EU legal environment, and particularly in the case of SES, it should be noted that there are also non-protected categories that can serve as proxies for a protected ground [72]. Algorithms effectively identify correlations between factors that may not be identical to the protected ground; nevertheless, when considered in conjunction [34], these correlations can highlight the significance and the interference of proxies with the protected grounds and thus justify the existence of discrimination.

3 PROBLEM BACKGROUND AND MOTIVATION

We consider a supervised fairness-aware machine learning setup where a dataset $\mathcal{D}$ is drawn from an independent and identically distributed (i.i.d.) sample space $\mathcal{P}(X, S, Y)$. Here, X denotes unprotected attributes (e.g., education, job experience), S represents protected attributes (e.g., race, sex), and Y is the target class (e.g., loan approval). Protected attributes in S identify historically discriminated groups [61], which must be safeguarded in automated decision-making $E(h(x))$. $\mathbb{E}(h(x))$ is the expected favourable outcome from a predictor $h()$. Without loss of generality, we assume

---

[2]Anti-discrimination law provides protection on the grounds of protected attributes [3].

[3]The list of Article 19 includes the following protected grounds: sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation.

[4]Directive 2000/43/EC against discrimination on grounds of race and ethnic origin; Directive 2000/78/EC against discrimination at work on grounds of religion or belief, disability, age or sexual orientation; Directive 2006/54/EC equal treatment for men and women in the field of employment and occupation; Directive 2004/113/EC equal treatment for men and women in the access to and supply of goods and services.

[5]The list of Article 21 includes the following protected grounds:"…*on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation.*"

[6]The list of Article 14 includes the following protected grounds:"…*on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.*"

[7]Belgium (social origin and wealth/income/property), Bulgaria (social status), Croatia (property, social status, social origin), France (economic vulnerability), Greece (social status), Hungary (social origin, financial status, part-time nature of employment), Lithuania (social status), Romania (social status), Slovakia (social origin, property), Slovenia (social standing, economic situation).

$Y \in \{0, 1\}$, with $Y = 1$ indicating favourable outcomes. Using this setup, fairness notions can be generalized as:

$$\mathbb{E}(h(x) \mid C) = \mathbb{E}(h(x) \mid s, C)), \ \forall (x, s, y) \in \mathcal{D} \tag{1}$$

where s denotes protected identity (group or subgroup), and C defines the fairness condition [63]. This formula ensures that models favourable outcomes are comparable across groups under the specified fairness condition.

While numerous studies have explored various fairness notions [4], their evaluations [46, 63], and underlying concepts [40], to the best of our knowledge, none have specifically addressed the issue of socio-economic status (SES)-driven bias in the context of achieving demographic fairness. To understand the drawbacks of existing fairness definitions in recognizing SES disparities, we focus on three popular [4] notions: Equal Opportunity (EP) [38], Demographic Parity (DP) [23], and Conditional Demographic Parity (CDP) [41]. EP ensures fairness in true positive rates across groups and aligns with correctness in predictions. DP eliminates outcome dependency on protected attributes, CDP extends DP by conditioning equality on additional unprotected attributes like occupation, tackling nuanced disparities like those in "redlining" [62]. These notions were selected due to their legal [49, 62, 69], and practical [8, 20, 58, 65] significance in socio-economic contexts. Throughout the exploration, we use the popular Adult dataset [10] as $\mathcal{D}$, and for each fairness notion employ an in-processing fair-classifier [1] $h(X)$ constrained on respective fairness notion.

Equal Opportunity (EP)[38]: Eq.1 specifies EP under $C = (Y = 1)$, i.e., conditioned on the positive ground truth: $\mathbb{E}(h(x) \mid y = 1) = \mathbb{E}(h(x) \mid s, y = 1)$. While EP promotes prediction correctness across demographics, it often struggles with class imbalances in fairness-aware datasets [48]. For example, Fig.1 shows that an EP-aware classifier[1] can maintain high TPR ($P(h(X) = 1 \mid Y = 1)$) and low FPR ($P(h(X) = 1 \mid Y = 0)$) for both sex yet still exhibit substantial gaps in PPR ($P(h(X) = 1)$). Moreover, it exacerbates imbalances against females across occupation, as seen in Fig.1, thereby widening male-female disparities. The ground truth distribution (Fig.2) reveals that females experience fewer favorable outcomes (PPR Female/Male≈ 0.36) than males. Even with respect to capital-gain—a wealth proxy [52]—females consistently show lower PPR than males with gap being higher for the less wealthier (capital-gain below top 8%). Thus, EP by emulating ground truth, can amplify entrenched socio-economic inequalities.

Demographic Parity (DP) [23]: DP is achieved when $\mathbb{E}(h(x)) = \mathbb{E}(h(x) \mid s)$ in Eq 1, ensuring outcomes are independent of the protected attribute. Originally motivated by the need to combat racial discrimination in financial contexts (e.g., unequal credit)[65], DP aims to yield nearly equal PPR across all demographics, as shown by the DP-aware classifier[1] in Fig. 1 (PPR - females: 25.94%, males: 27.92%). This often relies on higher false positives for the disadvantaged group, resembling affirmative action'[8], yet its blind distribution of positive decisions raises concerns about exacerbating negative prejudices against protected groups [8]. As an example in Fig. 1, on high-income (#positive (>50K) > #negative (≤50K)) professions for males (c.f. Fig 2) such as Executive managerial' and

'Professional specialty', DP-aware approach fails to address class imbalances for females, missing opportunities for more targeted positive action that could help mitigate existing disparities.

Conditional Demographic Parity (CDP) [20]: CDP extends DP by conditioning equality in PPR on a categorical unprotected attribute (denoted by X_a). More formally based on Eq 1: $\mathbb{E}(h(x) \mid a) = \mathbb{E}(h(x) \mid s, a)$ for all $a \in X_a$. CDP draws inspiration from study of "redlining" [62] effect, where location proxies led to financial service disparities. For Adult, we trained a CDP-aware classifier with "occupation" as the conditional attribute. As shown in Fig. 1, it achieves nearly equal PPR for males (0.28) and females (0.24) with high TPRs (>0.86). Unlike EP- and DP-aware classifiers, the CDP-aware model achieves fairer outcomes across occupations (Fig. 1), predicting a higher likelihood of high income for females in high-income professions.

Should CDP be the gold standard of fairness in socio-economic contexts? Fig. 1 demonstrates that the CDP-aware classifier better promotes protected demographics. However, it remains unclear whether its decisions are influenced by factors such as SES-privilege. Socio-economic privilege, including inherited wealth [45, 53], often impacts opportunities [37]. Although the Adult dataset lacks some explicit wealth attribute, the capital-gain attribute could serve as a proxy [52]. Studies [60] suggest a strong correlation between capital gain and favourable outcomes.

Using capital gain, we define the "Privileged" group as the top 5% in capital gain and the "underprivileged" group as the remaining population. In Fig. 3, we observe stark disparities between these groups across males and females. In particular, privileged males and females achieve high PPRs (89.80% and 85.85%, respectively), whereas underprivileged groups see sharp declines (23.74% for males and 21.94% for females, respectively). Positive actions for underprivileged females (through the CDP-aware classifier) contribute to higher false positives (FPR 14.09%) compared to underprivileged males (1.20%), which some may argue is necessary to promote equity. However, as already mentioned, concerns arise about selecting undeserving candidates, potentially increasing biases against protected groups [8].

To address this, we evaluate individual effort using a proxy, the hours-per-week attribute[9] in Adult. In particular, we define individuals above the mean as "high-effort individuals" [11]. Fig. 3 shows improved PPR for high-effort subgroups (34.98% for females and 36.68% for males, respectively). Still, discrepancies in FPR (**6%↓**) and PPR (**59%↓**) between privileged and the underprivileged females show unfair treatment due to the privilege factor.

In conclusion, privileged groups benefit significantly from CDP, while underprivileged individuals, even with high effort, remain disadvantaged. Following the principle of equalization [24], affirmative actions should prioritize effort and need over privilege. Although, the CDP-aware classifier, addresses some disparities, it fails to explicitly account for wealth disparities and does not adequately reward high-effort, underprivileged individuals. Therefore, CDP may not suffice as the gold standard of fairness in socio-economic contexts. Building on these observations, we propose a new fairness notion that aims to capture both (demographic) need and

---

[8]"Affirmative action" or "positive action" (as referred to in the EU anti-discrimination Directives) refers to specific measures favouring members of a disadvantaged group [30, 49].

[9]Work hours in this example is just used as a reference and not necessarily our viewpoint. In reality measuring effort [13] can vary depending on policy. Our SEP and CSEP notions can comply with any such real-valued function measuring effort.
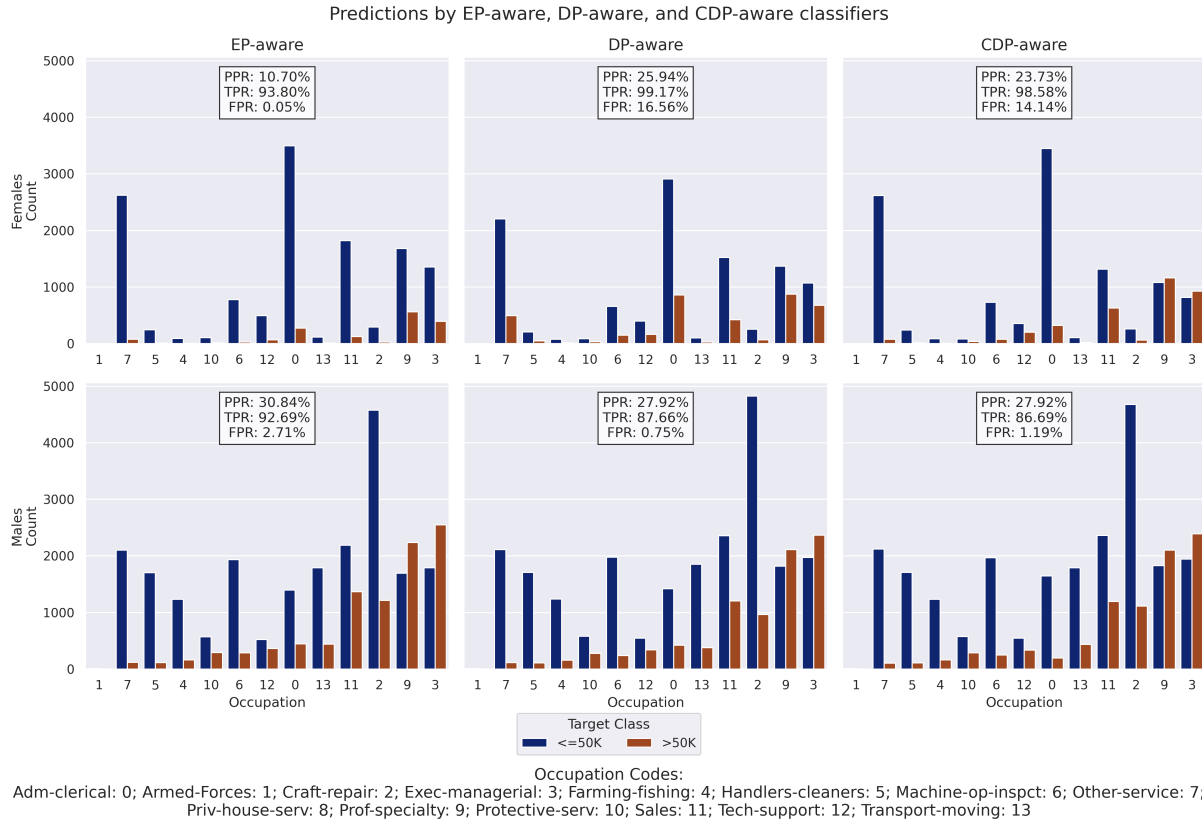
Figure 1: Distribution of positive (>50K) and negative (≤50K) classes for males and females predicted by EP-aware, DP-aware, and CDP-aware classifiers. Occupations are sorted in ascending order w.r.t. count of positive (ground truth) labels from left to right.
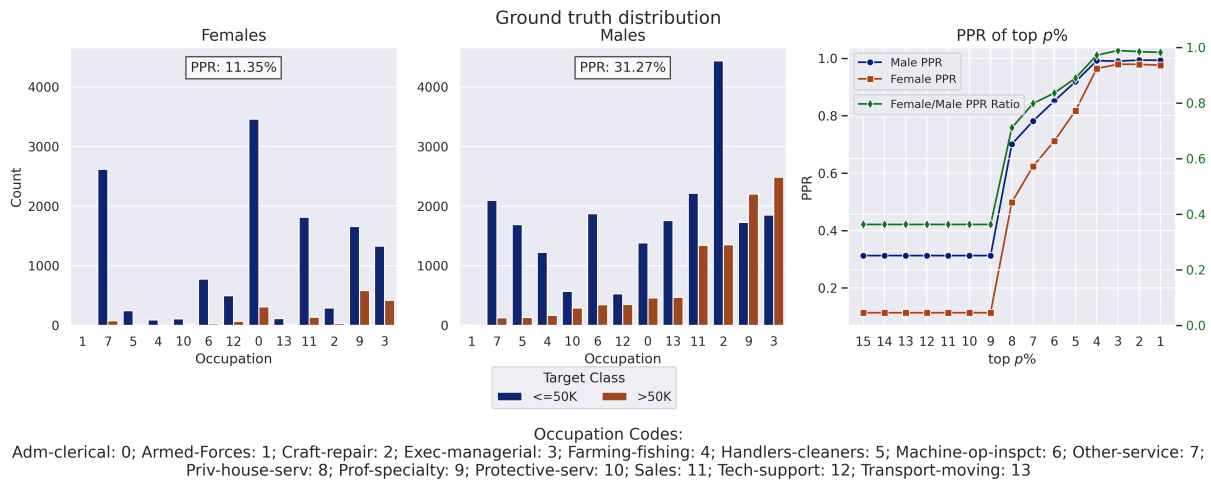

Figure 2: Ground truth distribution of positive (>50K) and negative (≤50K) classes for males and females by occupation. Occupations are sorted in ascending order w.r.t. count of positive labels from left to right. PPR scores refer to the % of positive labels in the ground truth. PPR of top p% represents the population with top p% capital gain.

(individual) effort by explicitly integrating SES and effort into its framework.

4 SOCIO-ECONOMIC STATUS-AWARE FAIRNESS NOTION

Our findings in Sec. 3 suggest that to be fair in the socio-economic context certain social factors, such as inherited privilege, and economic factors such as individual effort, need to be considered. More concretely:

a. The chances of favourable outcomes should not be hindered for socio-economically underprivileged individuals.
b. Individuals in underprivileged subgroups who put higher effort beyond a threshold should be given a higher chance of positive rewards, ideally proportional to their effort.
c. Positive action for privileged subgroups should be minimized in order to reduce the influence of privilege.

Building on these observations, we define two notions of SES-aware fairness (Socio-Ecomomic Parity - SEP and its stronger counterpart, Conditional SEP - CSEP) (Sec. 4.1) and their corresponding SES-aware fairness metrics (Sec. 4.2). We train fair models based on these metrics and evaluate their effectiveness in mitigating discrimination (Sec. 4.3).

4.1 Defining Socio-Economic Parity (SEP) and Conditional Socio-Economic Parity (CSEP)

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:*

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

such that:

$$\mathbb{E}(h(x) \cdot \zeta(x_e) \mid x_p < \tau_p, x_e \geq e) \geq \mathbb{E}(h(x) \mid x_p < \tau_p, x_e < e),$$
$$\mathbb{E}(h(x) \mid x_p \geq \tau_p, y = 0) \leq \mathbb{E}(h(x) \cdot \zeta(x_e) \mid x_p < \tau_p, x_e \geq e, y = 0) \tag{2}$$

where x_e is the value w.r.t. X_e-an identified non-protected attribute that can be expressed as a proxy for measuring effort put forth by an individual, ϵ is the disparity tolerance (ideally 0), and $\zeta(\cdot)$ is a real-valued function that relatively weights underprivileged individuals (to ≥ 1) based on their effort value x_e for individuals with $X_e \geq e$. The threshold τ_p defines the privileged group as the population with top $p\%$ value distribution of X_p.

In simpler words, SEP is achieved when the expected favourable outcome for any protected group s is independent of socio-economic privileges. Additionally, it ensures that the false positive rate for the privileged group is not higher than that of the underprivileged group exerting high effort. Furthermore, the cumulative rewards for underprivileged high effort individuals are proportionally weighted based on their evaluated effort. A stronger (and more desirable) version of the SEP fair notion, is what we call Conditional Socio-Economic Parity (CSEP):

DEFINITION 2. *A predictor h thats behold Socio-Economic Parity (SEP) across all the categories a defined by a categorical non-protected*

attribute X_a : $X_a \neq X_p, X_e$, is called Conditional Socio-Economic Parity (CSEP) -fair. Formally:

$$\mathbb{E}(h(x) \mid a) = \mathbb{E}(h(x)|s, a, x_p < \tau_p),$$
$$\forall_{a \in X_a, X_a \neq X_p, X_e}, \forall (x, s, y) \in D; \zeta : \mathbb{R} \to [1, \infty)$$

such that:

$$\mathbb{E}(h(x) \cdot \zeta(x_e, a) \mid a, x_p < \tau_p, x_e \geq e) \geq \mathbb{E}(h(x) \mid a, x_p < \tau_p, x_e < e),$$
$$\mathbb{E}(h(x) \mid a, x_p \geq \tau_p, y = 0) < \mathbb{E}(h(x) \cdot \zeta(x_e, s, a) \mid$$
$$a, x_p < \tau_p, x_e \geq e, y = 0) \tag{3}$$

Note that in Eq. 3, the function $\zeta(\cdot)$ takes the additional argument a. This is to ensure that the effort-based weighting also considers the localized condition. For example, in each job, effort weights will vary based on the threshold, such as the average working hours per week in that job. The notion can be further refined by adopting subgroup-specific thresholds for τ_p and e, in order to more accurately capture the socio-economic influences within each community.

4.2 Measuring the violation of SEP and CSEP

Based on definitions of the SEP, and CSEP notions introduced in Sec. 4.1, here we introduce the AI fairness metric measuring the SES-aware bias/discrimination. Following Eq. 2, violation of SEP can be measured as:

$$\forall s \in S, \ |P(h(X) = 1) - P(h(X) = 1 \mid X_p < \tau_p, S = s)|$$
$$+ |\frac{1}{A} \sum_{x:x_e < e} P(h(X = x) = 0 \mid X_p < \tau_p, S = s, X_e < e)$$
$$- \frac{1}{B} \sum_{x:x_e \geq e} \zeta(e, s) P(h(X = x) = 0 \mid X_p < \tau_p, S = s, X_e \geq e)| \tag{4}$$
$$+ |\frac{1}{C} \sum_{x:x_p \geq \tau_p} P(h(X = x) = 0 \mid X_p \geq \tau_p, Y = 0)$$
$$- \frac{1}{B} \sum_{x:x_e \geq e} \zeta(e, s) P(h(X = x) = 0 \mid X_p < \tau_p, s, X_e \geq e, Y = 0)| \leq \epsilon$$

where $A = \sum_{x_p < \tau_p, x_e < e} 1$, $B = \sum_{x_p < \tau_p, x_e \geq e} \zeta(e, s)$, and $C = \sum_{x_p \geq \tau_p} 1$, ensures the respective parts remains a probability distribution. Note that the second and the third part of the Eq. 4 focuses on equating the probability negative predictions i.e., $h(X = x) = 0$, contrary to the favourable (positive) outcome expectation goal in Eq. 2. This is intentional and necessary, to avoid penalizing individuals exerting high effort. Similar to Eq. 4 we can trivially extend the measure following Eq. 3 to evaluate CSEP violation conditioned on non-protected condition X_a.

In absence of mechanism to evaluate effort, or to simplify computation and interpretation, violation of SEP (Eq. 4) can be relaxed to: $\forall s \in S, \ |P(h(X) = 1) - P(h(X) = 1 \mid X_p < \tau_p, s)| \leq \epsilon$ where the relaxed goal is to only capture the discrimination towards all the underprivileged demographic subgroups.

4.3 Empirical effectiveness of SES-aware fairness

To examine the effectiveness of our CSEP notion in reducing the effect of SES, similar to our previous exploration, we optimize a model [1] constrained on our notion (Eq. 3) and test it on the Adult data. We plot the results in Fig 4. The key things to notice here is
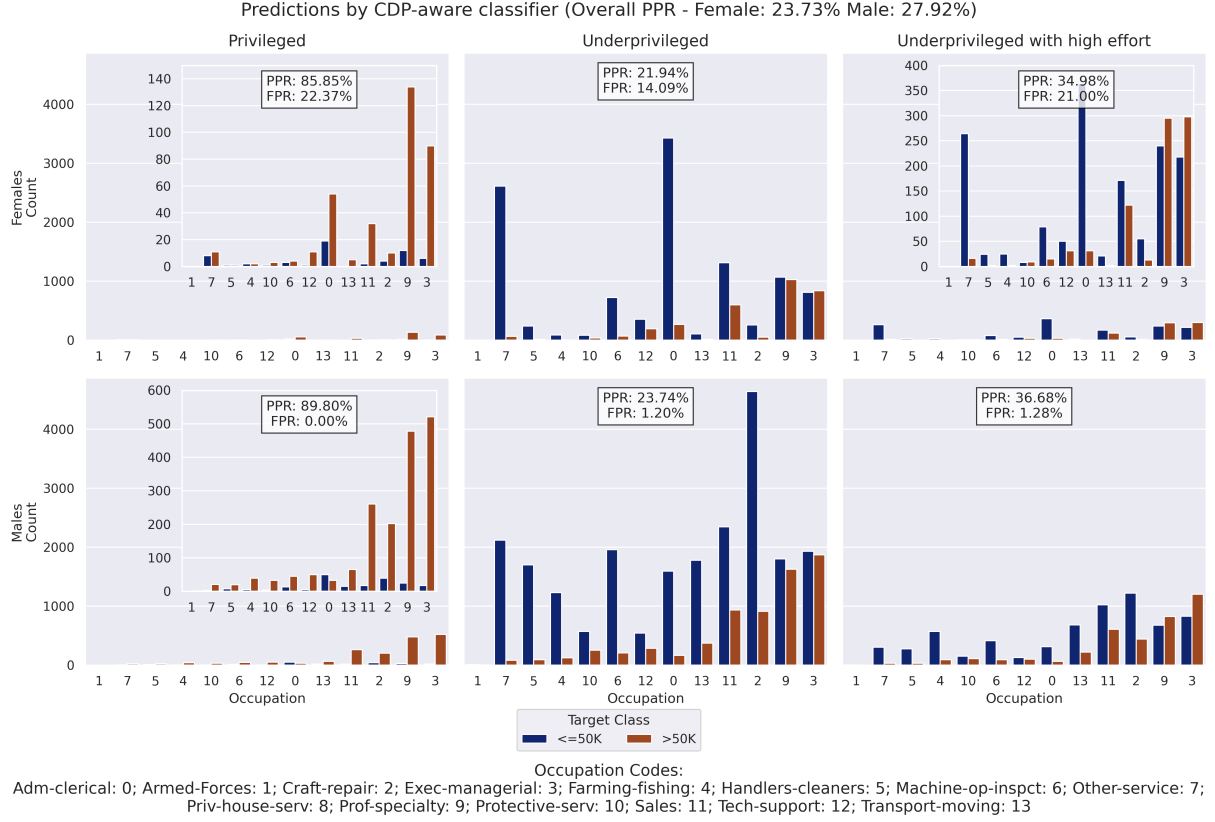
Figure 3: Predictions by CDP-aware classifier on female and male demographics, across different occupations, for socio-economic subgroups defined as privileged (individuals among top 5% capital gain), underprivileged (individuals below the top 5% capital gain margin), and underprivileged with high efforts (underprivileged individuals with working hours per week greater than mean of their demographics). Subplots that have fewer data (due to data scarcity of the subgroup) are presented with a zoomed version inside the respective subplots to provide a better view of the results.

that our CSEP-aware model starkly improves the PPR of high-effort underprivileged protected group to 56.44% (by 59%↑) by awarding more positive actions for this niche subgroup (FPR 47.09%) while keeping the false predictions for the overall underprivileged females at par with that of CDP-aware model (Fig. 3). Further, our CSEP-aware model also brings down the positive actions for the privileged females (FPR 58%↓) while keeping the overall PPR between male (27.84%) and female (24.19%) nearly equal. Fig. 5 further compares PPR rates by hours per week for underprivileged males and females under EP-, DP-, CDP-, and CSEP-aware models. Notably, CSEP allocates positive outcomes proportionally to effort among high-effort underprivileged females (≥ mean effort), nearly doubling the female-to-male PPR ratio (0.98 to 2.1) and mitigating both overall demographic disparities and those observed among protected underprivileged subgroups (Fig. 2). Specifically, the PPR for underprivileged females increases from below 0.3 for those working fewer than the mean hours to above 0.6 for those exceeding it, reaching nearly 0.7 for those working 55 hours[10] or more. In contrast,

---

[10]Work hours in this example is just used as a reference and not necessarily our viewpoint. In reality measuring effort [13] can vary depending on policy. Our SEP and CSEP notions can comply with any such real-valued function measuring effort.

EP- and DP-based methods under-reward these high-effort females, resulting in lower PPRs than their male counterparts. Although the CDP-aware model provides higher PPR for all underprivileged individuals, it does so uniformly across both demographics, failing to deliver the additional positive actions needed to support high-effort underprivileged females and address remaining disparities.

5 INVESTIGATING SEP NOTION UNDER THE AI ACT AND EU ANTI-DISCRIMINATION LAW

By placing the provision of justifications at the forefront of the study, we aim to ensure accountability, as the definition and realisation of explicit selections of the important attributes, context and any trade-offs of a fairness-aware metric undertaken to protect the rights, freedoms, and interests of individuals [39]. Thus, drawing upon the elucidation of the field of SES under the AI Act, the regulation states that "...*AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems...*" (Recital 58). In parallel, from the combination of Article 6(2) and Annex II, par. 5(b) of the AI Act, it is explicitly indicated
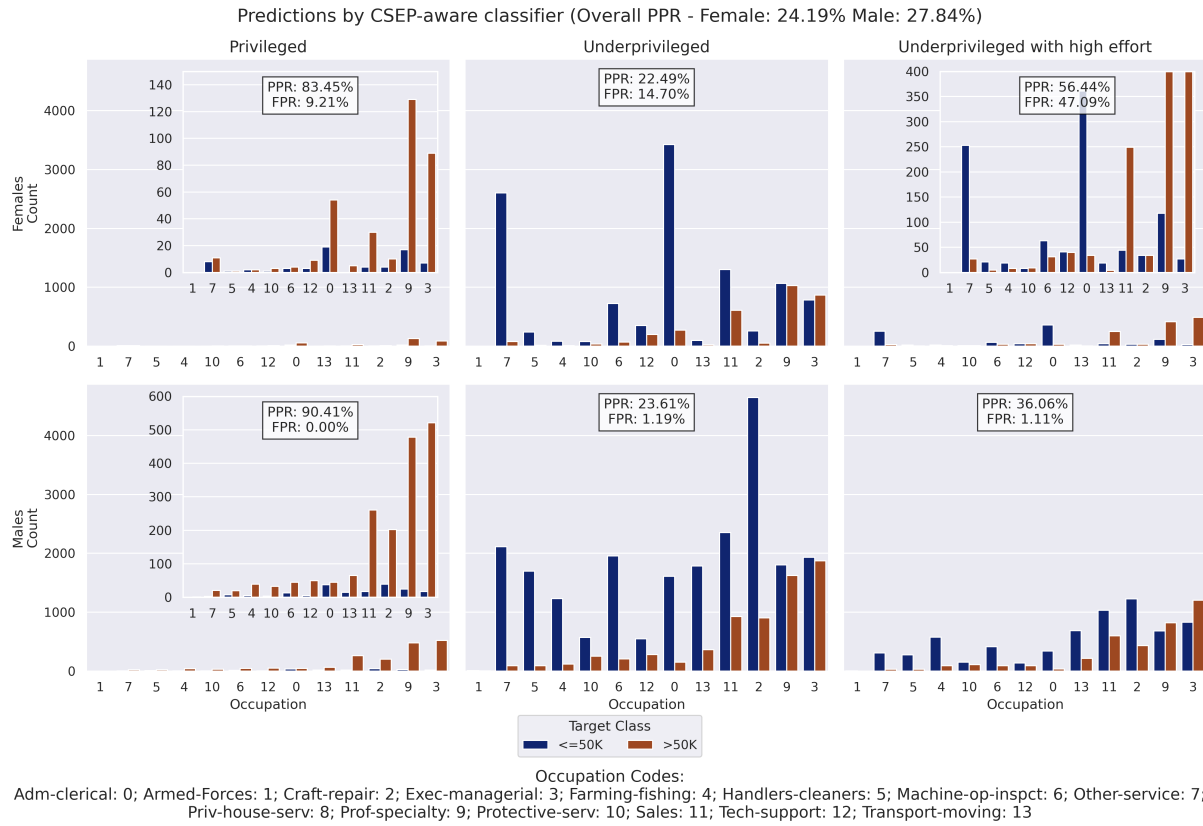
Predictions by CSEP-aware classifier (Overall PPR - Female: 24.19% Male: 27.84%)

Occupation Codes:
Adm-clerical: 0; Armed-Forces: 1; Craft-repair: 2; Exec-managerial: 3; Farming-fishing: 4; Handlers-cleaners: 5; Machine-op-inspct: 6; Other-service: 7; Priv-house-serv: 8; Prof-specialty: 9; Protective-serv: 10; Sales: 11; Tech-support: 12; Transport-moving: 13

Figure 4: Distribution of positive (>50K) and negative (≤50K) classes for males and females predicted by CSEP-aware classifier. Occupations are sorted in ascending order w.r.t. count of positive (ground truth) labels from left to right.
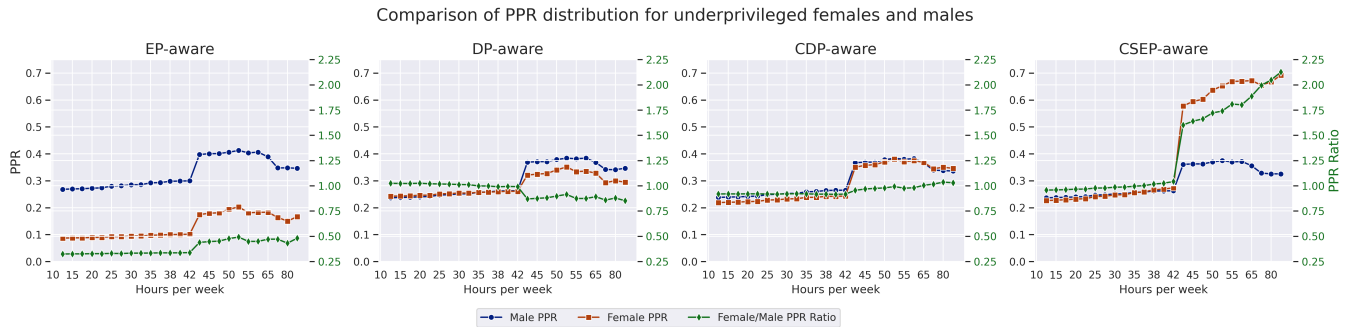

Figure 5: Comparison of our CSEP-aware model against EP, DP, and CDP -aware models in distributing positive action towards underprivileged females and males. Higher PPR ratio indicates higher preference of positive action for females.

that AI systems "*intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud*" are classified as high-risk AI systems, with the requirement to satisfy the provisions laid down in Articles 8-15 (Chapter III, Section 2) that should be assessed under the conduction of a conformity assessment [Article 3(20)]. In addition to those obligations, deployers of the systems that are described in Article 6(2) should perform

a Fundamental Rights Impact Assessment (FRIA) and notify the national authority of the results (under the circumstances of Article 27); thus, regarding AI bias, the FRIA aims to detect and mitigate the discriminatory risks of an AI system.

Regarding the topic of AI bias in AI systems, the AI Act formalises the requirement of identification and mitigation of unfair discrimination in high-risk AI systems [Article 10(2)(f), (g)]. Especially for the aspect of detection, AI system providers should first

identify the risk level of each AI system, recognising also the importance of preventing the (re)production of biases, that can lead to discrimination and taking into consideration the potential protected ground(s) of discrimination under EU law; the requirement of representative and ad hoc datasets [Article 10(3),(4)] also contributes to that direction. Moreover, the requirement for human oversight in Article 14(4)(b) of the AI Act shifts the focus to the identification of automation bias by the responsible person "in the loop" in the context of high-risk AI systems. Regarding accuracy, robustness and cybersecurity, Article 15(4) introduces the requirement to apply bias mitigation measures to the development of high-risk AI systems in order to avoid reproducing bias via "feedback loops". Furthermore, accountability is enhanced through the requirements of the prior design of the technical documentation (Article 11) and record-keeping of logs (Article 12).

Therefore, AI fairness-aware metrics should be designed through the notions of proving compatibility with the fundamental rights in order to be able to be utilized through the required assessments of the AI Act. The right not to be discriminated, which is a fundamental principle enshrined in EU law, should be explored and examined by every attempt to address AI bias. From the text of the AI Act, it is obvious that the regulation delegates the determination of discrimination to existing EU legislation [21].

The two main forms of discrimination, under EU law, include direct and indirect discrimination [68, 74]. Direct discrimination[11] occurs when a person is being undertaken less favourable treatment than another in a comparable situation, based on a protected ground, while indirect discrimination[12] exists when an ostensibly neutral provision, criterion or practice sets people with a particular protected characteristic in a disadvantaged position compared with others unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary[13] [29, 30]. Therefore, establishing indirect discrimination requires demonstrating the presence of two groups (comparators), one advantaged and one disadvantaged by the contested measure [30]. An expression of direct discrimination is discrimination by association [35], which can be identified when a person is treated less favourably due to another person's protected characteristic[14]. In cases where several grounds of discrimination exist, multiple discrimination takes place when the grounds operate separately, and intersectional discrimination occurs when the grounds interact, are inseparable and produce specific types of discrimination; nevertheless, the Court of Justice of the European Union (CJEU) has not recognized that a new category of discrimination can arise from the combination of more than one of those grounds [30, 63].

It has been widely claimed in the literature that AI bias should be viewed and addressed through the concept of indirect discrimination [9, 73, 75]. With the reasoning that direct discrimination will arise only if the explicit or implicit bias of the decision-maker influences the model, cases of unintentional discrimination, such as those caused by sampling errors or historical bias, are excluded from consideration [36] and due to the nature of AI systems, as they depend on inferences and proxies for target variables and protected attributes [68].

5.1 Automated contextual extraction of source of privilege (X_p)

Starting from the principle of proportionality under indirect discrimination, once it has been established, the burden of proof shifts to the defendant, effectively initiating a proportionality test [35]. It has been argued that AI fairness metrics should be based on the "proportionality test" (where the legitimate interest is pursued in a manner that is both necessary and proportionate) [68]; this approach can enhance the alignment between legal principles and AI-fairness metrics and, per analogiam, in the determination of the trade-offs and essential balancing of interests through fairness-aware mechanisms. Therefore, our solution includes the implications and essential justifications for the following question:

Are the selected attributes of the AI fairness-aware approach/ metric necessary and appropriate to protect the legitimate interests of the protected groups?[15] From a top-down approach, we start by identifying the source of biases that can affect the protected groups. The attributes identified by our approach, as most critical for accurately representing the comparators' groups, depend on statistical data and the ability of the trained models to identify bias. By implementing computational fairness notions to compare the treatment of different groups, clear statistical evidence of the existence of discrimination can be provided; this evidence can be taken into consideration by the court in cases involving allegations of discrimination[16] [31, 74].

Also, it is revealed that the attributes (including protected grounds and proxies) that are pivotal to influence the decisions as most critical for accurately representing the comparators' groups and prone to affect the model's learning bias, as the contribution of determining proxy variables, which are non-protected grounds ('capital-gain', 'work-hour-per-week', 'occupation') can guide the AI fairness-aware measures, through the investigated correlations to acknowledge protected grounds ('sex') and, as a result, contribute to the aim of definition and transforming the harmful discriminatory effects in SES context. The fact that the ground of SES is not covered by Article 19 TFEU [33] has provoked discussions on multiple levels; primarily, the lack of SES recognition [33], and secondarily, in algorithmic discrimination through harmful patterns that are present and do not directly relate to protected grounds under EU law [25].

Identifying attributes that can be considered as a source of privilege requires background knowledge about what can constitute socio-economic privilege and what is the historical impact of the

---

[11]Relevant provisions to EU law are included in Article 2(2)(a) of the Directive 2000/43/EC, Article 2(2)(a) of the Directive 2000/78/EC, Article 2(1)(a) of the Directive 2006/54/EC and Article 2(a) of the Directive 2004/113/EC, to the extent of the protected grounds of the directives.

[12]Relevant provisions to EU law are included in Article 2(2)(b) of the Directive 2000/43/EC, Article 2(2)(b) of the Directive 2000/78/EC, Article 2(1)(b) of the Directive 2006/54/EC and Article 2(b) of the Directive 2004/113/EC, to the extent of the protected grounds in the directives.

[13]Article 2(2)(b) Directive 2000/43/EC; Article 2(2)(b) Directive 2000/78/EC; Article 2(b) Directive 2004/113/EC; Article 2(1)(b) Directive 2006/54/EC.

[14]See: CJEU, C-303/06, S. Coleman v. Attridge Law and Steve Law, 17 July 2008.

[15]"necessary" and "appropriate" means that they "...must be suitable to achieve the desired end" [57].

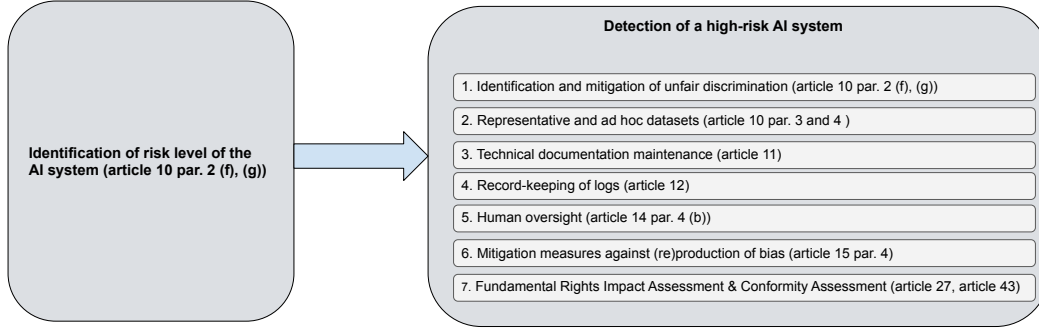[16]See Recital 15 of the Racial Equality Directive 2000/43/EC.

Figure 6: Specific provisions in the AI Act addressing AI bias

considered attribute in determining the current outcome. So, some may rightly argue that a fully automated determination of X_p may not be possible. However, since here we are interested in fairness-aware learning of an AI model, it can be intuitively useful if we focus on what the AI model considers as source of privilege. One simple strategy would be to identify an unprotected attribute which is important in producing favourable outcome for a non-protected demography (historically known to garner privilege), and inherently with its attribute values can segregate the population into different sections of hierarchy. We use a simple two-step process to approximate the above-mentioned idea, given as follows:

step1: Train a predictor $h_g(\cdot)$ with a subset $\mathcal{D}_g$ i.e., the subset of the data $\mathcal{D}$ drawn only from the distribution of the non-protected population $\mathcal{P}(X, g, Y)$.

step2: Among the subset of unprotected ordinal (X_o) e.g., "Education" [5], and numerical (X_n) e.g., "Capital gain" [70] attributes, using a feature importance method $\mathcal{FI}$ [28], identify the attribute $X_p \in (X_o \bigcup X_n)$ that has the highest impact on model's predictions for the privileged population $\mathcal{D}_g$, i.e., $X_p = \underset{X_i \in (X_o \bigcup X_n)}{\mathrm{argmax}} \mathcal{FI}(h(X), X_i)$. Note that the restriction of attribute X_p being either ordinal or nominal is a necessary condition to quantify the hierarchy of privilege.

5.2 Determining the value of p, which top $p\%$ privileged population to consider?

The identification of the comparators' groups plays a vital role both in the legal justification of the discrimination and in the implementation of the proposed fairness-aware notions. First, our approach initiates a comprehensive statistical investigation which is used to reveal the discrimination patterns and emphasizes the context-specific implementation of the fairness transforming methods, aiming to satisfy the goal of addressing AI bias with the ad hoc appropriate actions, which are considered proportionate. To align with the definition of indirect discrimination and address it in practice, we classify individuals into two comparator groups ("privileged" and "underprivileged"), based on the criterion of disadvantaged SES.

Therefore, the two groups, based on sex, have been determined according to the variables of `capital-gain`, `work-hour-per-week`, `occupation`. Second, our investigation shows in Fig. 5 that if the `underprivileged group`, when given certain advantages as the privileged, can perform equally based on the related variables. To enable this prediction, we set other two subgroups indicating the top $p\%$ in relation to the advantage of high capital gain. The positive prediction is based on counterfactual rationale [12], where these females will have a higher capital gain (if we increase the attribute value for the identified females that do not belong to the top $p\%$). It should be mentioned that this counterfactual reasoning lays down a "positive action"[17] which refers to the measure that promotes preferential treatment for disadvantaged groups [59] to eliminate, prevent or remedy past discrimination [26]. As "positive action" is, in general, adopted under EU law, the requirement should be for the AI fairness metrics to include the balancing to reach the goal of equal treatment, by not maintaining feedback loops that can deteriorate the SES of a population group.

Third, the determination of the comparators' groups based on a threshold (top 5% of males and females), in the case of fairness-aware methods, relies on a quantitative criterion. Initially, it should be mentioned that the evidential basis of the criterion aligns with this requirement of CJEU[18] with the aim of displaying that the contested rule is related to any form of discrimination [36]. Apart from clarifying the measurable nature of the predictive actions, the focus should also be shifted to its examination of satisfying the characteristics of "necessary" and "appropriate". In general, for the determination of comparators' groups through the definition of a threshold, it has been claimed that explicit quantitative limits derived from the EU legal framework are absent [64], and a hypothetical disadvantage is sufficient [36]. The necessity of determining this threshold lies in the objective of identifying the privileged category within the entire population, irrespective of protected identity. In particular, the positive class ratio between females and males surpasses the 80% rule only from the top $p\%$ onward. In parallel, the positive label distribution for the group of females exceeds the 80%

---

[17]Relevant provisions: Article 5 Directive 2000/43/EC; Article 7 Directive 2000/78/EC; Article 6 Directive 2004/113/EC; Article 3 Directive 2006/54/EC.
[18]CJEU, C-167/97, Seymour-Smith, 9 February 1999.

rule, thereby qualifying to be considered a privileged category, from the threshold of the top $p\%$ (as it is shown in Fig. 2), maintaining at the same time the appropriate size of the population.

6 STANDARDIZATION FOR THE AI ACT AND FAIRNESS MEASURES UTILIZATION

Initially, with regard to AI standards within the scope of the AI Act, the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CEN-ELEC) are considered permanent members of the advisory forum for the regulation, as specified in Article 67 of the AI Act. As the practical application of the AI Act will rely on the development and implementation of harmonized standards [15], in response to request [27] by the European Commission, CEN and CENELEC accepted to prepare a work programme of European standards and European standardization deliverables [66]. Notably, AI conformity assessment and risk management are included, as one of the targets (among others) of Task Group Inclusiveness, while the FRIA is not referred in the non-binding first newsletter [66]. As part of broader efforts to support standardization for legally and ethically fair AI systems, the critical actions necessitate a dialogue between the fairness-aware metrics and principles in light of the FRIA, and the conformity assessment (as described in Sec. 5). In parallel with the provisions of the AI Act, standardization efforts should also consider the 2019 Ethics guidelines for trustworthy AI [17] along with the Assessment List for Trustworthy AI [18], as the Guidelines are recalled in Recital 27 of the AI Act into the risk-based approach of the regulation.

Interference between fairness-aware machine learning measures and the assessments provided by the AI Act: During the conformity assessment process, several requirements must be examined, as illustrated in Fig. 6. In particular, there is a strong interference between the technical measures for ensuring AI fairness and the identification and mitigation of unfair discrimination in high-risk AI systems [Article 10(2)(f),(g)] in order to achieve the aims of the regulation. The detection of AI bias can be identified through technical measures that examine AI fairness, which also according to Article 9(5) of the AI Act can guide the design and development of the AI system to prevent or, where prevention is not possible, mitigate relevant risks [50]. The accountability of AI system providers should be placed at the forefront, supported by essential documentation, justifications regarding the process, and the technical solutions that must accompany this assessment to demonstrate practical compliance.

With regards to the FRIA, the risk assessment requirements [of Article 27(1)(a)-(f)] could play a vital role in technical metrics that address bias in the context of high-risk systems, as their implementation can help to evaluate, justify and measure under subpar. (d) "*the specific risks of harm likely to have an impact on the categories of natural persons or groups of persons...*" under statistical measurable criteria. As the FRIAs are expected to be implemented as context-based assessments [50], we would like to highlight the relevance of this condition, which can thus integrate specialized AI fairness metrics designed for individual contexts into the determination of explicit measures and attributes, as well as the setting of acceptable thresholds. More specifically, in terms of identifying/measuring

discriminatory risks under the assessment, the establishment of criteria based on the AI fairness metrics can contribute to evaluating the discriminatory risks that AI bias may pose to individuals. In particular, this can be achieved through the analysis of proxies, enabling the exploration of protected grounds and examination of important correlations that lead to discriminatory risks. Moreover, an additional element of the FRIA that could ultimately involve AI fairness metrics and solutions refers to Article 27(1)(f) and especially on "*the measures to be taken in the case of the materialization of those risks, including the arrangements for internal governance and complaint mechanisms.*" The implementation of technical solutions (in this case of the proposed method, concerning SES) could serve as a means to address the issue by reconfiguring AI bias in this area, and thus mitigate risks (under the FRIA) when an AI system exhibits risks related to the presence of discrimination. In parallel, the integration of specific AI fairness technical algorithmic solutions can contribute to enhancing a comprehensive management plan aimed at ensuring compliance, during the phase of addressing and responding to the potentially discriminatory outcomes derived from this assessment.

7 CONCLUSIONS, LIMITATIONS, AND FUTURE DIRECTIONS

In this work, we introduced Socio-Economic Parity (SEP), a novel AI fairness notion, alongside Conditional Socio-Economic Parity (CSEP), to address SES-driven disparities in AI systems, under the prism of the EU AI Act. This interdisciplinary study aims to bridge legal science and ML and highlight aspects that require further attention and exploration. Our empirical evaluation using Adult demonstrates that CSEP effectively rewards high-effort underprivileged protected groups while maintaining comparable error rates across the broader underprivileged female population, and accounting for equity across the conditional attribute categories. Despite these promising results, further investigation is required to address remaining limitations and ensure real-world applicability:

- *Scope of evaluation* The empirical evaluation is currently limited to Adult, a commonly used benchmark in fairness research. Future work should test the proposed framework on additional datasets representing diverse fairness scenarios to ensure its generalizability across different contexts.
- *Context-specificity* The proposed framework relies on context-specific definitions and proxies for socio-economic status (e.g., capital gain) and individual effort (e.g., working hours). These were chosen to illustrate the underlying concepts, however, such definitions and proxies should generally be tailored to the specific problem or domain.
- *Computational overhead* Additional constraints introduced in the notion increases the training computation, and might need relaxation of tolerance threshold ϵ for convergence.

Despite these limitations, the framework offers a strong *foundation for addressing financial exclusion through the exploration of (un)privileged groups*. The critical balance achieved by SEP relies on measurable evidence provided by the model, enhancing transparency and supported by comprehensive legal explanations for all decisions. In particular, the introduction of comparators' groups through the concept of indirect discrimination provides an

effective tool for identifying and addressing inequalities between privileged and underprivileged groups within the SES. Empirically, our experiments show that applying an SEP-driven fairness constraint can reward high-effort underprivileged subgroups while maintaining comparable error rates across the broader population, underscoring SEP's potential to narrow socioeconomic gaps in machine-learning applications without sacrificing performance.

Through our legal analysis, we also emphasized the importance of **_strengthening SES within the EU AI Act_**. The rapid development of the AI sector, highlights the need to refine SES-related evaluations, given its current status as a non-protected ground under EU law. Enhanced SES protection can be reinforced through AI Act assessments, especially considering discrimination risks.

A vital next step involves leveraging statistical data from sociological and economic studies focused on SES-specific contexts, which can further solidify SES-aware decisions in AI frameworks. We hope this work will inspire further studies on the role of SES in fairness-aware machine learning.

ACKNOWLEDGMENTS

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International conference on machine learning*. 60–69.

[2] Stanford University AI Index Steering Committee, Institute for Human-Centered AI. 2024. The AI Index 2024 Annual Report: Public Data.

[3] Jose M Alvarez, Alejandra Bringas Colmenarejo, Alaa Elobaid, Simone Fabbrizzi, Miriam Fahimi, Antonio Ferrara, Siamak Ghodsi, Carlos Mougan, Ioanna Papageorgiou, Paula Reyero, et al. 2024. Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology* 26, 2 (2024), 31.

[4] Guilherme Alves, Fabien Bernier, Miguel Couceiro, Karima Makhlouf, Catuscia Palamidessi, and Sami Zhioua. 2023. Survey on fairness notions and related tensions. *EURO journal on decision processes* 11 (2023), 100033.

[5] Yuval Arbel, Ronen Bar-El, and Yossi Tobol. 2017. *Equal opportunity through higher education: Theory and evidence on privilege and ability*. Technical Report. IZA Discussion Papers.

[6] Shreya Atrey. 2018. The intersectional case of poverty in discrimination law. *Human Rights Law Review* 18, 3 (2018), 411–440.

[7] Abhijit V Banerjee and Esther Duflo. 2019. *Good economics for hard times*. PublicAffairs.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

[9] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[10] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[11] Gary S Becker. 1985. Human capital, effort, and the sexual division of labor. *Journal of labor economics* 3, 1, Part 2 (1985), S33–S58.

[12] Ruth MJ Byrne. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning.. In *Twenty-Eighth International Joint Conference on Artificial Intelligence*. 6276–6282.

[13] Alexander W Cappelen and Bertil Tungodden. 2009. Rewarding effort. *Economic Theory* 39 (2009), 425–441.

[14] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 4209.

[15] CEN-CENELEC JTC 21. 2022. Business Plan for JTC 21. https://standards.cencenelec.eu/BPCEN/2916257.pdf

[16] European Commission. 2022. 2022 Factsheet on the gender pay gap. https://commission.europa.eu/document/download/61252380-e608-4157-818d-e81619715310_en?filename=equal_pay_day_factsheet_2022_en_1_0.pdf

[17] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI*. Publications Office.

[18] European Commission, Content Directorate-General for Communications Networks, and Technology. 2020. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office.

[19] European Commission, Directorate-General for Justice, Consumers, I Chopin, and C Germaine. 2020. *A comparative analysis of non-discrimination law in Europe 2019. The 28 EU Member States, Albania, North Macedonia, Iceland, Liechtenstein, Montenegro, Norway, Serbia and Turkey compared*. Publications Office. https://doi.org/doi/10.2838/797667

[20] Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky. 2012. *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Vol. 3. Springer Science & Business Media.

[21] Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, and Niklas Kühl. 2024. Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness. *arXiv preprint arXiv:2403.20089* (2024).

[22] Sionaidh Douglas-Scott. 2011. The European Union and human rights after the Treaty of Lisbon. *Human rights law review* 11, 4 (2011), 645–682.

[23] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[24] Ronald Dworkin. 2018. What is equality? Part 2: Equality of resources. In *The notion of equality*. Routledge, 143–205.

[25] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[26] Eurofound. 2011. *Positive action, European Industrial Relations Dictionary, Dublin*. https://www.eurofound.europa.eu/en/european-industrial-relations-dictionary/positive-action

[27] European Commission. 2023. Commission implementing decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence. https://ec.europa.eu/transparency/documents-register/api/files/C(2023)3215_0/de00000001048942?rendition=false

[28] Fiona Katharina Ewald, Ludwig Bothmann, Marvin N Wright, Bernd Bischl, Giuseppe Casalicchio, and Gunnar König. 2024. A guide to feature importance methods for scientific inference. In *World Conference on Explainable Artificial Intelligence*. Springer, 440–464.

[29] European Union Agency for Fundamental Rights. 2022. *Bias in algorithms - Artificial intelligence and discrimination*. Publications Office of the European Union. http://fra.europa.eu/en/publication/2022/bias-algorithm

[30] European Union Agency for Fundamental Rights and Council of Europe. 2018. *Handbook on European Non-discrimination Law–2018 edition*. Publications Office of the European Union. https://fra.europa.eu/en/publication/2018/handbook-european-non-discrimination-law-2018-edition

[31] Sandra Fredman. 2011. *Discrimination Law*. Oxford University Press, Oxford.

[32] Sarah Ganty. 2021. Poverty as misrecognition: What role for antidiscrimination law in Europe? *Human Rights Law Review* 21, 4 (2021), 962–1007.

[33] Sarah Ganty and Juan Carlos Benito Sanchez. 2021. Expanding the List of Protected Grounds within Anti-Discrimination Law in the EU. *S. Ganty and J.-C. Benito Sanchez, Expanding the List of Protected Grounds within Anti-Discrimination Law in the EU, Equinet* (2021).

[34] Janneke Gerards and Frederik Zuiderveen Borgesius. 2022. Protected grounds and the system of non-discrimination law in the context of algorithmic decision-making and artificial intelligence. *Colorado Technology Law Journal* 20 (2022), 1.

[35] Janneke Gerards and Raphaële Xenidis. 2021. *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law*. European Commission.

[36] Philipp Hacker. 2018. Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. *Common market law review* 55, 4 (2018).

[37] Victoria J Haneman. 2021. Wealth, Privilege, Power, and Opportunity. *Jotwell: J. Things We Like* (2021), 1.

[38] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[39] Information Commissioner's Office (ICO). 2023. *Guidance on AI and Data Protection*. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/

[40] Sofia Jaime and Christoph Kern. 2024. Ethnic classifications in algorithmic fairness: Concepts, measures and implications in practice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 237–253.

[41] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems* 35 (2013), 613–644.

[42] William Kilbourne and Susan Weeks. 1997. A socio-economic perspective on gender bias in technology. *The Journal of Socio-Economics* 26, 3 (1997), 243–260.

[43] Claire Kilpatrick and Hanna Eklund. 2022. Article 21–non-discrimination. In *The EU Charter of Fundamental Rights*. Nomos Verlagsgesellschaft mbH & Co. KG, 613–638.

[44] Rakesh Kochhar and Richard Fry. 2014. Wealth inequality has widened along racial, ethnic lines since end of Great Recession. *Pew Research Center* 12, 104 (2014), 121–145.

[45] Philipp Korom. 2016. *Inherited advantage: The importance of inheritance for private wealth accumulation in Europe.* Technical Report. MPIfG Discussion Paper.

[46] Emmanouil Krasanakis and Symeon Papadopoulos. 2024. Towards Standardizing AI Bias Exploration. *arXiv preprint arXiv:2405.19022* (2024).

[47] Mark Latonero. 2018. *Governing Artificial Intelligence: Upholding Human Rights and Dignity.* https://datasociety.net/library/governing-artificial-intelligence/

[48] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), e1452.

[49] William M Leiter and Samuel Leiter. 2011. *Affirmative action in antidiscrimination law and policy: An overview and synthesis.* State University of New York Press.

[50] Alessandro Mantelero. 2024. The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. *Computer Law & Security Review* 54 (2024), 106020.

[51] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. Artificial Intelligence Index Report 2024. *arXiv preprint arXiv:2405.19522* (2024).

[52] Tazra Mitchell. 2023. *Taxing Capital Gains More Robustly Can Help Reduce DC's Racial Wealth Gap.* Technical Report. DC Fiscal Policy Institute. https://www.dcfpi.org/all/taxing-capital-gains-more-robustly-can-help-reduce-dcs-racial-wealth-gap/ Revenue & Budget / Taxes, Tax Justice Is Racial Justice.

[53] Randall Morck, David Stangeland, and Bernard Yeung. 2000. Inherited wealth, corporate control, and economic growth. *Concentrated corporate ownership* 3 (2000).

[54] Catelijne Muller. 2020. The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law. https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da

[55] Access Now. 2018. *Human rights in the age of artificial intelligence.* Technical Report. Access Now. https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

[56] Publications Office of the European Union. 2017. *European Convention on Human Rights (ECHR).* https://eur-lex.europa.eu/EN/legal-content/glossary/european-convention-on-human-rights-echr.html

[57] Publications Office of the European Union. 2023. *Principle of proportionality.* https://eur-lex.europa.eu/EN/legal-content/glossary/principle-of-proportionality.html

[58] Executive Office of the President, Cecilia Munoz, Domestic Policy Council Director, Megan (US Chief Technology Officer Smith (Office of Science, Technology Policy)), DJ (Deputy Chief Technology Officer for Data Policy, Chief Data Scientist Patil (Office of Science, and Technology Policy)). 2016. *Big data: A report on algorithmic systems, opportunity, and civil rights.* Executive Office of the President.

[59] Colm O'Cinneide. 2011. *Positive Action and EU Law.* https://www.era-comm.eu/oldoku/SNLLaw/04_Positive_action/2011-111DV20-O'Cinneide_EN.pdf

[60] Emmanouil Panagiotou, Manuel Heurich, Tim Landgraf, and Eirini Ntoutsi. 2024. TABCF: Counterfactual Explanations for Tabular Data Using a Transformer-Based VAE. In *5th ACM International Conference on AI in Finance.* 274–282.

[61] European Parliament and Office for Official Publications of the European Communities. 2007. Charter of fundamental rights of the European Union.

[62] Willy E Rice. 1996. Race, gender, redlining, and the discriminatory access to loans, credit, and insurance: An historical and empirical analysis of consumers who sued lenders and insurers in federal and state courts, 1950-1995. *San Diego L. Rev.* 33 (1996), 583.

[63] Arjun Roy, Jan Horstmann, and Eirini Ntoutsi. 2023. Multi-dimensional discrimination in law and machine learning-A comparative overview. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 89–100.

[64] Fredman Sandra. 2011. *Discrimination law.* Oxford University Press, Oxford.

[65] Leslie Scism and Mark Maremont. 2010. Insurers test data profiles to identify risky clients. *The Wall Street Journal* 19 (2010).

[66] Task Group (TG), CEN-CENELEC Joint Technical Committee (JTC) 21. 2024. Artificial Intelligence standardisation. Inclusiveness Newsletter. https://www.cencenelec.eu/news-and-events/news/2024/newsletter/ots-49-etuc/

[67] Christa Tobler. 2014. Equality and Non-Discrimination under the ECHR and EU Law. A Comparison Focusing on Discrimination against LGBTI Persons'.

Zeitschrift für ausländisches öffentliches Recht und Völkerrecht 74, 3 (2014), 521–561.

[68] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.* 123 (2020), 735.

[69] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567.

[70] Robert B Williams. 2017. Wealth privilege and the racial wealth gap: A case study in economic stratification. *The Review of Black Political Economy* 44, 3-4 (2017), 303–325.

[71] Raphaële Xenidis. 2020. Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law* 27, 6 (2020), 736–758.

[72] Raphaële Xenidis and Linda Senden. 2020. *EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination.* Kluwer Law International.

[73] Tal Z Zarsky. 2017. An analytic challenge: discrimination theory in the age of predictive analytics. *I/S: A Journal of Law and Policy for the Information Society* 14 (2017), 11.

[74] Frederik Zuiderveen Borgesius, Philipp Hacker, Nina Baranowska, and Alessandro Fabris. 2024. Non-discrimination law in Europe, a primer. Introducing European non-discrimination law to non-lawyers. *arXiv preprint arXiv:2404.08519* (2024).

[75] Frederik J Zuiderveen Borgesius. 2020. Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights* 24, 10 (2020), 1572–1593.