Can We Trust Fair-AI?

Salvatore Ruggieri[1], **Jose M. Alvarez**[1, 2], **Andrea Pugnana**[2],
Laura State[1, 2], **Franco Turini**[1]

[1] University of Pisa, Pisa, Italy
[2] Scuola Normale Superiore, Pisa, Italy
salvatore.ruggieri@unipi.it, jose.alvarez@di.unipi.it, andrea.pugnana@sns.it
laura.state@di.unipi.it, franco.turini@unipi.it

Abstract

There is a fast-growing literature in addressing the fairness of AI models (fair-AI), with a continuous stream of new conceptual frameworks, methods, and tools. How much can we trust them? How much do they actually impact society? We take a critical focus on fair-AI and survey issues, simplifications, and mistakes that researchers and practitioners often underestimate, which in turn can undermine the trust on fair-AI and limit its contribution to society. In particular, we discuss the hyper-focus on fairness metrics and on optimizing their average performances. We instantiate this observation by discussing the Yule's effect of fair-AI tools: being fair on average does not imply being fair in contexts that matter. We conclude that the use of fair-AI methods should be complemented with the design, development, and verification practices that are commonly summarized under the umbrella of trustworthy AI.

The Landscape of Fairness in AI

Fair-AI methods are designed with the purpose of controlling biased decisions in algorithmic decision making[1] (Schwartz et al. 2022; Ntoutsi et al. 2020). A highly relevant case of bias is discrimination against protected-by-law social groups (Altman 2020). Fairness, however, can imply different meanings depending on the context as well as the discipline (Mitchell et al. 2021; Mulligan et al. 2019). For instance, equity requires that people are treated according to their needs, which does not mean all people are treated equally (Minow 2021). To formalize and measure the degree of (un)fairness, quantitative metrics have been introduced in philosophy, economics, and machine learning in the last 50 years (Lee, Floridi, and Singh 2021; Hutchinson and Mitchell 2019; Binns 2018; Romei and Ruggieri 2014), amounting to more than 20 different definitions thus far (Castelnovo et al. 2022; Mehrabi et al. 2021; Berk et al. 2021; Verma and Rubin 2018; Zliobaite 2017). *Group fairness* metrics aim at measuring the statistical difference in distributions of decisions across social groups. *Individual fairness* metrics bind the distance in the decision space to the distance in the feature space describing people's characteristics. *Causal fairness* metrics exploit knowledge beyond observational data to infer causal relations between features and decisions, and to estimate interventional consequences. As with other quality objectives, the choice of a fairness metric is crucial for optimizing AI models. See the previous surveys and (Räz 2021; Wachter, Mittelstadt, and Russell 2021; Hertweck, Heitz, and Loi 2021; Binns 2020) for a discussion of the moral/legal bases and relative merits of the various fairness notions and metrics. Based on these fairness metrics, methods and tools have been proposed for bias detection (*discrimination discovery* or *fairness testing*) (Chen et al. 2022), for data de-biasing and data processing (*pre-processing approaches*) (Biswas and Rajan 2021), for training models and representations (*in-processing approaches*) (Wan et al. 2022), for correcting existing models (*post-processing approaches*), and for monitoring models' decisions (*monitoring*) (Kenthapadi et al. 2022). We also refer to (Pessach and Shmueli 2022; Hort et al. 2022; Mehrabi et al. 2021) and to (Fabris et al. 2022; Quy et al. 2022), respectively, for surveys of the techniques and of the datasets commonly used in the field.

Research in fair-AI originated from the supervised learning area, but it is rapidly expanding, e.g., to unsupervised and reinforcement learning, natural language processing (NLP), computer vision, and speech processing, among others. Major AI conferences regularly include papers and workshops on the topic. A few global events (ACM FAccT[2], AAAI/ACM AIES[3], ACM EAAMO[4], and FoRC[5]) are targeted at multidisciplinary aspects of fairness and other ethical issues in AI and algorithmic decision making. Similarly, several off-the-shelf software libraries are available to practitioners, expanding at a fast pace, yet with some critical gaps to be addressed (Richardson and Gilbert 2021; Lee and Singh 2021). A number of initiatives have started to standardize, audit, and certify algorithmic fairness, such as the IEEE P7003™ Standard on Algorithmic Bias Considerations[6], the IEEE Ethics Certification Program for Autonomous and Intelligent Systems[7], the ISO/IEC TR

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] Due to the large body of literature, we prioritize survey papers, where applicable, and recent works.

[2] https://facctconference.org/
[3] https://www.aies-conference.com/
[4] https://eaamo.org/
[5] https://responsiblecomputing.org/
[6] https://standards.ieee.org/project/7003.html
[7] https://standards.ieee.org/industry-connections/ecpais.html

24027:2021 - Bias in AI systems and AI aided decision making[8], and the NIST AI Risk Management Framework [9]. However, very few works attempt at investigating the practical applicability of fairness in AI (Madaio et al. 2022; Makhlouf, Zhioua, and Palamidessi 2021b; Beutel et al. 2019), whilst several external audits of AI-based systems have been conducted (Koshiyama et al. 2021), sometimes with extensive media coverage.

The issue of engineering fairness is challenging (Scantamburlo 2021), and likely to require domain-specific approaches (Lee and Floridi 2021). For instance, individual fairness metrics require to define (or to learn (Ilvento 2020)) a distance function to quantify how different two persons are with respect to the characteristics that matter for the decision. The distance function, if any, must be necessarily specific of the application domain (e.g., by appropriately weighting skills for job candidates, capacity to repay for credit applicants, etc.). Finally, on the educational side, bias and fairness have become common topics of university courses on technology ethics (Fiesler, Garrett, and Beard 2020), albeit they are not sufficiently included in core technical courses (Saltz et al. 2019) nor sufficiently transversal and interdisciplinary (Raji, Scheuerman, and Amironesei 2021).

Some Issues with Fair-AI

The AI community tends to self-correct recurring research mistakes in cycles (Lipton and Steinhardt 2019; McDermott 1976). Now it is the turn of reflecting on how fairness in AI has been developing in the last fifteen years since (Pedreschi, Ruggieri, and Turini 2008; Kamiran and Calders 2009). Rather than discussing the frontiers of fair-AI research (Chouldechova and Roth 2020), we focus on a (necessarily incomplete) collection of issues that prevent well-established fair-AI methods from being impactful. We start by discussing a few theoretical limitations and practical problems of fair-AI, and refer to the appropriate literature.

Hyper-focus on abstract metrics. Theoretical results state that it is impossible to achieve different fairness notions at the same time, such as for instance *independence* $\hat{Y} \perp\!\!\!\perp R$ and *separation* $\hat{Y} \perp\!\!\!\perp R \mid Y$ in the case of group fairness metrics (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2017). Here, $\hat{Y}$ is a (random) decision variable, such as admission to university; R a socially sensitive feature, such as gender; and Y a ground truth decision. For simplicity, we restrict our focus here to binary decisions and binary sensitive groups. Independence requires that the decision (whether by humans or machines) is statistically independent from the sensitive feature. Separation requires that they are independent conditionally on the ground truth, i.e., on the merit of people (Barocas, Hardt, and Narayanan 2019). Further, fairness notions are in tension not only among them, but also with other quality requirements, such as predictive accuracy (Menon and Williamson 2018), calibration (Pleiss et al. 2017), and privacy (Cummings et al. 2019), for which Pareto optimality should be considered (Wei and Niethammer 2022). Nevertheless, a corollary of the theoretical results is that the AI designers should opt for one of the many fairness metrics. The choice, however, requires to account for contrasting objectives: business utility, human value alignment (Friedler, Scheidegger, and Venkatasubramanian 2021), people's actual perception of fairness (Saha et al. 2020; Srivastava, Heidari, and Krause 2019), and legal and normative constraints (Xenidis 2020; Kroll et al. 2017). (Makhlouf, Zhioua, and Palamidessi 2021a) provide a decision diagram for guiding practitioners, which highlights the complexity of the choice. Based on how the above constraints are formalized, the fairness metrics and, a fortiori, their impact can be different (Passi and Barocas 2019) – an instance of the *framing effect* bias. For example, in the famous case analysed by *ProPublica*[10], the COMPAS algorithm for recidivism prediction fails to meet equal false positive rate among groups, but it achieves equal calibration (Corbett-Davies et al. 2017). Even when restricting to a specific fairness notion, there is a problem on how to quantify the degree of unfairness. In the case of independence, for instance, association measures are typically adopted, such as risk difference (a.k.a. statistical/demographic parity):

$$P(\hat{Y} = 1|R = 1) - P(\hat{Y} = 1|R = 0) \tag{1}$$

or selective risk ratio[11]:

$$P(\hat{Y} = 1|R = 1)/P(\hat{Y} = 1|R = 0) \tag{2}$$

possibly together with confidence intervals (Pedreschi, Ruggieri, and Turini 2009). We refer to (Maity et al. 2021) for individual fairness confidence intervals, and to (Shah and Peters 2020) for a characterization of the hardness of statistical tests of (conditional) independence. The apparently innocuous choice between the algebraic operators (1) or (2), however, has an enormous impact on how decisions are affected. (Pedreschi, Ruggieri, and Turini 2012) show that the top-k sub-populations with the highest risk difference and with the highest selective risk ratio do not coincide. Therefore, optimizing an AI model w.r.t. risk difference or selective risk ratio affects the relative impact of AI decisions for a same sub-population.

The impossibility of fairness. In most cases, collecting the ground truth Y is impossible, expensive, or even unethical, as it would require to obtain counterfactual outcomes, such as releasing potential criminals, not treating sick patients, etc. In the analysis of the COMPAS algorithm, ground truth was approximated by the actual recidivism outcome of defendants in the two years period after they were scored. However, we do not know whether or not a defendant who was not released would have recidivated in case she/he would have been released. Similarly, we do not know whether an applicant with denied credit would have repayed the credit if granted, a sample selection bias problem tackled by reject inference in credit scoring (Ehrhardt et al. 2021). An idea close to reject inference has been considered in (Ji,

[8]https://www.iso.org/standard/77607.html

[9]https://www.nist.gov/itl/ai-risk-management-framework

[10]https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[11]Sometimes wrongly referred to as "disparate impact". See (Watkins, McKenna, and Chen 2022).

Smyth, and Steyvers 2020) for group fairness. Sampling bias in collected ground truth Y has been called *negative legacy unfairness* (Kamishima et al. 2012). In other contexts, such as in NLP, the ground truth is obtained by human annotation, typically aggregating annotators's labels through majority voting. Here, the simplifying assumption of a *single* ground truth is used. A perspectivist approach is emerging in favor of granting significance to divergent opinions, by designing methods over non-aggregated data (Cabitza, Campagner, and Basile 2023). In absence of unbiased ground truth, however, practitioners set Y to the target feature used for training an AI model. Any bias in the target feature risks to be lifted to the AI model with a false claim of fairness.

Lack of source criticism. The sensitive feature R, also known as *grounds of discrimination* (Romei and Ruggieri 2014), is a key input in the design of fair-AI systems. Fairness metrics boil down to compare AI model's performances or decisions across (individuals from) different social groups. (Stewart 2022) shows that optimizing for a fairness metric w.r.t. one way of classifying individuals can make it impossible to optimize the same metric for another way of dividing people up[12]. There are, however, inherent problems in coding human identity in raw data, an issue known as *datafication* (Mejias and Couldry 2019), which can become even more complex once we allow for identity to fluctuate (Lu, Kay, and McKee 2022), and care for the representativeness of grounds of discrimination in data, i.e., *representation bias* (Shahbazi et al. 2022). For instance, if gender is coded with a binary feature (male/female), then any further discrimination analysis is limited to contrasting only such two groups, excluding, e.g., non-binary people. More elaborate representations of human identity could benefit from ontologies for concept reasoning (Kronk and Dexheimer 2020). The issue of *source criticism* (Koch and Kinder-Kurlanda 2020), which is central in historical and humanistic disciplines, is still in its infancy in the area of AI. Source criticism attains at the provenance, authenticity, and completeness of data collected, especially in social media platforms. The adoption of source criticism practices in fair-AI would allow us to give a better picture of the data being used and the individual instances it contains.

Insufficient robust modeling. Purely observational approaches rely on correlation among features, and hence they are not able to account for spurious effects. A principled way of tackling unfairness is to rely on causal reasoning (Nogueira et al. 2022; Spirtes and Zhang 2016). Causal fairness metrics (Makhlouf, Zhioua, and Palamidessi 2020; Carey and Wu 2022b) overcome correlation shortcomings. On the other side, causal approaches require to know the structure of the causal relations among features, e.g., in terms of a causal graph. While approaches for causal discovery from data can be adopted, specifically in the context of fairness (Binkyte-Sadauskiene et al. 2022), they definitively need to be complemented with expert knowledge – but, with no guarantee of an unanimous agreement among experts (Rahmattalabi and Xiang 2022). Moreover, a number of assumptions are typically made which might not be met in practice, such as *sufficiency* (all causes are known), and *faithfulness* (the graph completely characterizes the conditional independences among features) (Spirtes, Glymour, and Scheines 2000). Overall, the causal fairness metrics may suffer from the *identifiability problem* (Makhlouf, Zhioua, and Palamidessi 2022), namely the impossibility to compute them from observational data only. Finally, arguments against the manipulability of the sensitive features, e.g., race, in counterfactual reasoning have been raised (Kohler-Hausmann 2019; Hu and Kohler-Hausmann 2020).

Lack of compositionality. Every (small or big) apparently-neutral technical decision in every step of the AI pipeline can impact the fairness of the final AI system. Fairness is affected by imputation of missing values (Caton, Malisetty, and Haas 2022), by encodings of categorical features (Mougan et al. 2022), by feature selection strategies (Galhotra et al. 2022), and even by hyper-parameter settings (Tizpaz-Niari et al. 2022), only to mention a few. Moreover, the composition of data transformations and models that are fair in isolation may not be fair in the end (Dwork and Ilvento 2019). Observe that this does not only apply to the compositions of AI systems, but also to the socio-technical systems resulting from the composition of AI, algorithms, people, and procedures. The lack of compositionality requires that the fairness analysis of a socio-technical system is conducted as a whole, not by pieces.

Can Fair-AI Be Unfair? The Yule's Effect

> [W]e cannot infer independence of a pair of attributes within a sub-universe from the fact of independence within the universe at large.
>
> *G. Udny Yule (Yule 1903, page 132)*

One additional issue, *the Yule's effect*, is introduced *by* the incorrect use of fair-AI methods. We discuss such an issue in detail in this section. First, we present a causal reasoning approach for correcting the unfairness of the decision procedure $\hat{Y}$. Next, we describe a common approach to the problem that adopts group fairness correction, possibly departing from the procedure grounded on the causal reasoning approach. Finally, we discuss the consequences of the common approach with an example, highlighting the Yule's effect of blindly correcting decision procedures.

Let us assume a scenario where we observe from historical data that $\hat{Y} \not\perp\!\!\!\perp R$, substantiated by a large risk difference.

What should be done? Risk difference, also called Total Variation in the causality literature, embeds direct, indirect, and spurious effects of R on $\hat{Y}$ (Plečko and Bareinboim 2022). Spurious effects are introduced by confounding variables, which cause both R and $\hat{Y}$. From a causal perspective, we are interested in measuring the direct and indirect effects

[12]Moreover, optimizing a fairness metric w.r.t. a collection of grounds of discrimination does not necessarily lead to being fair w.r.t. the intersection of such grounds (Stewart 2022) – the *intersectional fairness* problem. E.g., being fair w.r.t. black people and w.r.t. women does not necessarily imply being fair w.r.t. black women. See (Kong 2022) for a critique of current approaches to intersectional fairness, and (Wang, Ramaswamy, and Russakovsky 2022) for a few recommendations.

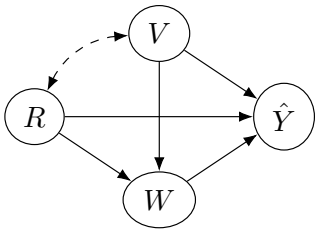

Figure 1: Standard Fairness Model (Plečko and Bareinboim 2022). Direct edges model possible causal dependencies. The dashed edge models spurious correlation induced by unobserved features. V is a confounder. W is a mediator.

only, whose sum is the Average Causal Effect (ACE):

$$P(\hat{Y}=1|do(R=1)) - P(\hat{Y}=1|do(R=0))$$

Let us formalize the causal relations among R, $\hat{Y}$ and other observed variables as in the semi-Markovian DAG in Figure 1, called the Standard Fairness Model (Plečko and Bareinboim 2022).

Consider now a third observed feature, called Z, which is the only input, together with R, to the decision procedure $\hat{Y}$. Let us develop a case-based reasoning on Z.

First, consider the case that Z is a mechanism through which the causal effect of R propagates to $\hat{Y}$ (it is a *mediator* between R and $\hat{Y}$), i.e., $W = Z$ and V is removed in Figure 1. Examples of mediators include legitimate business requirements such as level of education or prior experience in a job candidate selection. In this case, the ACE is equal to the risk difference metric, and since $\hat{Y} \not\perp\!\!\!\perp R$, it is non-zero. Therefore, the decision procedure leading to $\hat{Y}$ is unfair, and it *should be* corrected.

Now consider the case that Z is correlated with R, but not causally (it is a *confounder*), i.e., $V = Z$ and W is removed in Figure 1. Examples of confounders include demographic and geographic features. In such a case, the ACE can be calculated by averaging the stratified risk difference on Z through the *adjustment formula*:

$$\textstyle\sum_z (P(\hat{Y}=1|R=1,Z=z) - P(\hat{Y}=1|R=0,Z=z))P(Z=z).$$

Let us now distinguish two sub-cases. The first one assumes $\hat{Y} \perp\!\!\!\perp R \mid Z$, and it is known as the Simpons' paradox[13];

$$\hat{Y} \not\perp\!\!\!\perp R \quad \wedge \quad \hat{Y} \perp\!\!\!\perp R \mid Z$$

It occurs when vanishing partial correlations in separate distributions do not produce a vanishing mixture. In such a case, each term in the sum above is zero, and, a fortiori, the ACE is zero. In summary, we *should not* correct the decision procedure leading to $\hat{Y}$. This reasoning extends to collapsible association measures, such as the selective risk ratio, for which the value in the mixture is a weighted average of the values in the separate distributions (Pearl 2009; Huitfeldt, Stensrud, and Suzuki 2019) For non-collapsible metrics, such as the odds-ratio, the value at the mixture can be outside of the range of the values in the separate distributions. Hence, for non-collapsible metrics, the decision procedure *should* or *should not* be corrected based on the value at the mixture (which can be computed from the adjusted formula). The second sub-case assumes $\hat{Y} \not\perp\!\!\!\perp R \mid Z$. At least one term of the sum above is non-zero. Also, terms can be of opposite sign, which means that the overall sum can be zero or non-zero. The decision procedure *should not* or *should* be corrected based on the result of the sum.

What is typically done? In causal approaches, it can be difficult to determine whether Z is a mediator or a confounder (Barocas, Hardt, and Narayanan 2019, Chapter 5). This may lead to the wrong action with regard to the correction of the decision procedure. Non-causal approaches dismiss the above case-based reasoning altogether. They test independence only (or separation only). As a consequence, a typical approach after observing $\hat{Y} \perp\!\!\!\perp R$ consists of blindly correcting the decision procedure leading to $\hat{Y}$. With a few exceptions that will be recalled next, research papers adopting non-causal approaches fall back to this.

What are the consequences? Assume now that the decision procedure is corrected and deployed. We would then observe (close to) zero risk difference, which would support the conclusion $\hat{Y} \perp\!\!\!\perp R$. Is everything all right? According to the case-based reasoning above, the correction of the decision procedure may have mitigated or may have worsened fairness of the procedure[14]. Let us consider an example based on the *ACSIncome* dataset – an excerpt of the U.S. Census data (Ding et al. 2021). With reference to Figure 1, we set R to be the race of individuals, $\hat{Y}$ the predicted income (above 50K USD or not), W the number of working hours per week, and V the state of residence. Moreover, let Y be the true income. We split the available data into 67% for training a classifier, and 33% for testing its predictive performances and fairness metrics. An initial classifier is built using LightGBM (Ke et al. 2017), a state-of-the-art gradient boosting approach. We adopt the separation metric of the Equality of Opportunity (EOP) (Hardt, Price, and Srebro 2016):

$$P(\hat{Y}=1|Y=1) - P(\hat{Y}=1|Y=1,R=i)$$

which is the difference between the recall of positives (income above 50K USD) at population-level and at the level of the i^{th} racial group. The larger the EOP, the worse is the ability of the classifier to recall positives of the group compared to the average recall. The EOPs observed over the test set are reported in Figure 2 (left) in blue, from which we clearly conclude $R \not\perp\!\!\!\perp \hat{Y} \mid Y$. Let us now correct the decision procedure by a post-processing method that specializes the decision threshold for each racial group of R (Hardt, Price, and Srebro 2016). The EOPs observed after this (global) correction are shown in in Figure 2 (left) in orange. They are clearly closer to the optimal value of zero.

[13]The term has been improperly extended to include the Yule's effect, see (Spirtes, Glymour, and Scheines 2000, Sect. 3.5.2).

[14]Interestingly, the graph in Figure 1 is not anymore faithful to the new data. Since faithfulness is required by many approaches for causal discovery (Binkyte-Sadauskiene et al. 2022), reconstructing the causal structure of the new data (e.g., in an external audit study) may become problematic.

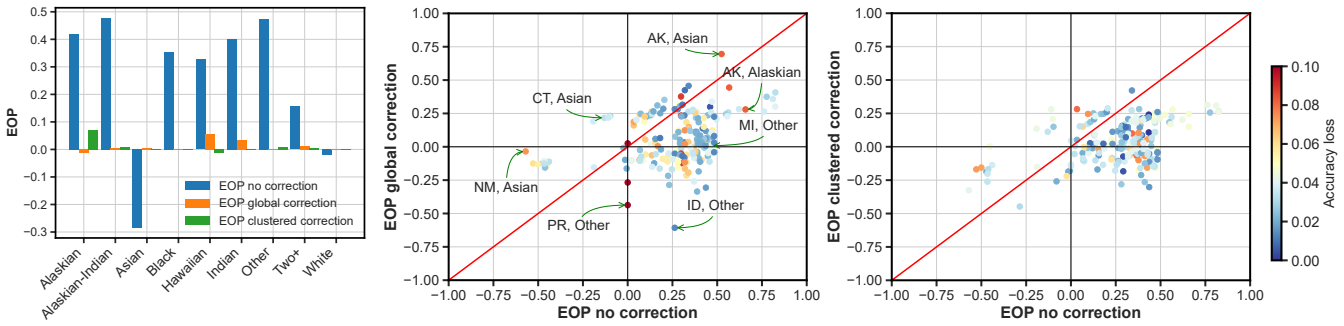

Figure 2: Left: EOPs for each racial group for classifiers with no correction, global correction, and clustered correction. Center: EOPs for each state and race, with color denoting the loss in accuracy after global correction. Right: EOPs for each state and race, with color denoting the loss in accuracy after clustered correction. Jupyter notebook available at https://github.com/ruggieris/DD/blob/main/notebooks/dd_ACSIncome_Yule.ipynb.

We would expect the corrected classifier to be fair not only at country level, but also at state level. However, the state is a confounder, and the correction of the classifier has not accounted for it. Figure 2 (center) shows that the EOPs of racial groups at each state have been affected by the correction in different ways. For instance, "Other races" in MI have a considerably lower EOP after correction. Asians in AK, instead, have a higher EOP metric. "Other races" in ID moved from being disfavored to being favored considerably, i.e., they moved from a recall much lower than average to a recall much higher than average. "Other races" in PR, which were not disadvantaged (close to zero EOP), result now to be advantaged (large negative EOP). Conversely, Asians in CT, which were favored, become disfavored after correction. Finally, notice that the loss in accuracy at state level after the correction, denoted by the color of dots, can be as high as 10% and it is not uniform across states, nor there is a clear pattern for how it is distributed.

The Yule's effect. The Yule's effect occurs when vanishing correlation in the mixture of a few distributions does not produce vanishing partial correlation in separate distributions:

$$\hat{Y} \perp\!\!\!\perp R \quad \wedge \quad \hat{Y} \not\perp\!\!\!\perp R \mid Z$$

The Yule's effect can occur when positive and negative associations between $\hat{Y}$ and R when conditioning on Z cancel out. This is precisely what has been pointed out in the example above. Whenever we aim at group fairness, such as independence $\hat{Y} \perp\!\!\!\perp R$, but we wrongly disregard to control for Z, fair machine learning algorithms may result into disparate effects on separate distributions, with some impacted positively (higher fairness) and other impacted negatively (lower fairness). The combined Simpson's paradox and Yule's effect can be summarized in a well-known general statement about conditional independence:

For $\mathbf{W} \subset \mathbf{Z}$*, then* $\hat{Y} \perp\!\!\!\perp R \mid \mathbf{W}$ *neither implies nor is implied by* $\hat{Y} \perp\!\!\!\perp R \mid \mathbf{Z}$.

As a consequence, independence fairness ($\hat{Y} \perp\!\!\!\perp R$) does not imply nor is implied by conditional independence fairness ($\hat{Y} \perp\!\!\!\perp R \mid Z$). Also, separation fairness ($\hat{Y} \perp\!\!\!\perp R \mid Y$) does not imply nor is implied by conditional separation fairness ($\hat{Y} \perp\!\!\!\perp R \mid Y, Z$). Moreover, when multiple confounders are present, conditional independence/separation w.r.t. all of them does not imply nor is implied by conditional independence/separation w.r.t. *a subset* of them. It means we should be aware of all the confounders in order to control for all of them, or, alternatively, we should know a more detailed structural causal model that allow for finer reasoning.

Conditional independence and conditional separation metrics have been addressed by (Kamiran, Zliobaite, and Calders 2013; Corbett-Davies et al. 2017). However, they are rarely adopted, mainly for two reasons. The first one is because it is difficult to determine whether a feature is a confounder or a mediator. Notice that by wrongly controlling for a mediator, we only measure the direct causal effect of R on $\hat{Y}$ (assuming that we are controlling on all confounders), hence we are ignoring the indirect effect. The indirect effect can be positive, if the mediator positively affects the advantageous decision and social groups have disproportionate distributions over the mediator. This is the case, for instance, of education level for job candidate selection, since social groups have disproportionate acckess to education. The indirect effect can also be negative, if the mediator results from the implementation of positive actions, e.g., quotas in favor of disabled people. The second reason why conditional independence/separation is not used in practice, is because the number of strata to control for can be very high. In the above example, there are 51 states. In general, the number of strata is equal to the product of the cardinalities of the domains of the features to control for. As a partial solution, (Kamiran, Zliobaite, and Calders 2013) propose to cluster the strata into a few groups to control for. Figure 2 (right) reports the result of separately correcting the classifier for each of five groups of states. The groups are obtained by clustering states based on the probability distribution of races within them using the k-means algorithm. Compared to the global correction, the clustered one is beneficial with respect to both EOP and accuracy loss. The mean absolute EOP is 0.258 for the uncorrected classifier, 0.119 for the globally corrected one, and 0.105 for the clustered corrected classifier.

Trusting Fair-AI

The previous two sections highlighted a few issues with fair-AI that may undermine the impact on society and ultimately the trust in fair-AI. The Yule's effect is only one example of worsening-off protected groups as an unintended consequence of fair-AI. Other worsening effects are described in the literature, e.g., regarding the long-run effects of imposing fairness constraints (Liu et al. 2018), or the impact of parameters in fair-AI methods (Ben-Porat, Sandomirskiy, and Tennenholtz 2021). In this section, we discuss some remedies and research pathways.

Accounting for multiple stakeholders. (Carey and Wu 2022a; Weinberg 2022) survey the existing critiques on the hegemonic theory of fairness that draw from non-computing disciplines, including philosophy, law, critical race and ethnic studies, and feminist studies. The hegemonic theory reduces the fairness problem to statistical parity or other metrics to be used in end-to-end optimization. Some of the issues with fair-AI discussed earlier in this paper clearly show this is insufficient. The usage of fair-AI methods does not necessarily guarantee the fairness of AI-based complex socio-technical systems (Kulynych et al. 2020). This is because the fairness objectives of the designers of AI, of the users of AI, and of the population subject to the AI decisions are unlikely to be the same. Fair-AI methods are currently not sufficiently robust and they can be incomplete in modelling the complexity and dynamic of the deployment scenario. Multi-stakeholders participatory design and policy actions that take into account qualitative contextual information and feedback from reality may be a valid alternative to technological solutionism. For instance, (Scott et al. 2022) adopt a participatory approach in the design of algorithmic systems in support of Public Employment Services.

The need for trustworthy AI. We think that the use of fair-AI methods should be complemented with design, development, and verification practices that are commonly summarized under the umbrella of *trustworthy AI* (Kaur et al. 2023). They include: human agency and oversight, accountability, explainability, robustness and safety, privacy, diversity, reproducibility, and societal and environmental wellbeing. The research on the interplay between fairness and those other non-functional requirements has been developing at different speed. We refer to surveys on human-centered algorithmic fairness (Wu and Liu 2022), differential privacy and fairness (Fioretto et al. 2022), fairness and diversity constraints in ranking (Zehlike, Yang, and Stoyanovich 2023), trust and fairness (Knowles, Richards, and Kroeger 2022), and fairness and robustness (Lee et al. 2021).

A large potential stems from the convergence of fairness and eXplainable AI (XAI) (Balkir et al. 2022; Zhou, Chen, and Holzinger 2020). XAI methods for model inspection, such as variable importance, can be used to test the influence/independence of R on $\hat{Y}$ (Grabowicz, Perello, and Mishra 2022). Adding explanations to an AI system's output can increase users' trust and fairness perception (Tal, Kuflik, and Kliger 2022) and ultimately control for the exercise of power (Lazar 2022). In particular, local explanation methods that describe why a specific output was produced (factual explanation) and what could have changed the output (counterfactual explanation) can help to identify reasons of discriminatory decisions (Manerba and Guidotti 2021).

Incorporating the option to reject. An underdeveloped research line consists of rejecting the output of an AI system in favor of escalating the decision to a human agent who could possibly take into account additional (qualitative) information. This is considered in the area of classification with a *reject option* (or *selective classification*) (Hendrickx et al. 2021). There is a trade-off here between the performance of an AI system on the accepted region, which should be maximized, and the probability of rejecting, which should be minimized (as human agents' effort is limited). Standard techniques for selective classification may worsen the fairness metrics over the accepted region (Jones et al. 2021). It would be interesting to explore methods specifically designed for rejecting[15] unfair predictions. For example, under which conditions on states and races (and possibly other features), a classifier in the examples in Figure 2 should abstain in order to prevent the most unfair predictions? A promising work adds risk difference constraints in the problem of determining the accepted region (Schreuder and Chzhen 2021).

Conclusions

The critiques to the hegemonic theory of fairness (Weinberg 2022), which reduces the fairness problem to a numeric optimization of some fairness metric, are not new to the AI community. For instance, (Wagstaff 2012) questioned the hyper-focus of Machine Learning on abstract metrics "in that they explicitly ignore or remove problem-specific details, usually so that numbers can be compared across domains" but the true significance and impact of the metrics is neglected. We have covered this and other issues, simplifications, and mistakes in fair-AI research and practice such as the impossibility of fairness due to observational ground truth, the lack of source criticism in data collection and representation, insufficient robust modeling, and the lack of compositionality in fairness analysis. The Yule's effect is an example of additional unfairness introduced by an erroneous use of fair-AI methods. These and other issues put fair-AI at risk of being untrusted and, a fortiori, of being limited import to society. Pathways for research and practice include multi-stakeholders participatory design, integration with other trustworthy tools for AI, notably explanation methods, and the option to reject unfair AI outcomes.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Actions (GA number 860630) for the project "NoBIAS - Artificial Intelligence without Bias". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

[15]The early work of (Kamiran, Karim, and Zhang 2012) considers a reject region, but, instead of abstaining on it, the approach sets the prediction to a favorable (resp., unfavorable) decision for the protected (resp., unprotected) social groups.

References

Altman, A. 2020. Discrimination. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/entries/discrimination/.

Balkir, E.; Kiritchenko, S.; Nejadgholi, I.; and Fraser, K. C. 2022. Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. *CoRR*, abs/2206.03945.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Ben-Porat, O.; Sandomirskiy, F.; and Tennenholtz, M. 2021. Protecting the Protected Group: Circumventing Harmful Fairness. In *AAAI*, 5176–5184. AAAI Press.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1): 3–44.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Woodruff, A.; Luu, C.; Kreitmann, P.; Bischof, J.; and Chi, E. H. 2019. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In *AIES*, 453–459. ACM.

Binkyte-Sadauskiene, R.; Makhlouf, K.; Pinzón, C.; Zhioua, S.; and Palamidessi, C. 2022. Causal Discovery for Fairness. *CoRR*, abs/2206.06685.

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *FAT*, volume 81 of *Proc. of Machine Learning Research*, 149–159. PMLR.

Binns, R. 2020. On the apparent conflict between individual and group fairness. In *FAT**, 514–524. ACM.

Biswas, S.; and Rajan, H. 2021. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *ESEC/SIGSOFT FSE*, 981–993. ACM.

Cabitza, F.; Campagner, A.; and Basile, V. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. In *AAAI*. AAAI Press.

Carey, A.; and Wu, X. 2022a. The statistical fairness field guide: perspectives from social and formal sciences. *AI Ethics*.

Carey, A. N.; and Wu, X. 2022b. The Causal Fairness Field Guide: Perspectives From Social and Formal Sciences. *Frontiers Big Data*, 5: 892837.

Castelnovo, A.; Crupi, R.; Greco, G.; Regoli, D.; Penco, I. G.; and Penco, A. C. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1): 4209.

Caton, S.; Malisetty, S.; and Haas, C. 2022. Impact of Imputation Strategies on Fairness in Machine Learning. *J. Artif. Intell. Res.*, 74: 1011–1035.

Chen, Z.; Zhang, J. M.; Hort, M.; Sarro, F.; and Harman, M. 2022. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *CoRR*, abs/2207.10223.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.

Chouldechova, A.; and Roth, A. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5): 82–89.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic Decision Making and the Cost of Fairness. In *KDD*, 797–806. ACM.

Cummings, R.; Gupta, V.; Kimpara, D.; and Morgenstern, J. 2019. On the Compatibility of Privacy and Fairness. In *UMAP (Adjunct Publication)*, 309–315. ACM.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *NeurIPS*, 6478–6490.

Dwork, C.; and Ilvento, C. 2019. Fairness Under Composition. In *ITCS*, volume 124 of *LIPIcs*, 33:1–33:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Ehrhardt, A.; Biernacki, C.; Vandewalle, V.; Heinrich, P.; and Beben, S. 2021. Reject inference methods in credit scoring. *J Appl Stat.*, 48(13-15): 2734–2754.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic Fairness Datasets: the Story so Far. *Data Min. Knowl. Discov.*, 36: 2074–2152.

Fiesler, C.; Garrett, N.; and Beard, N. 2020. What Do We Teach When We Teach Tech Ethics?: A Syllabi Analysis. In *SIGCSE*, 289–295. ACM.

Fioretto, F.; Tran, C.; Hentenryck, P. V.; and Zhu, K. 2022. Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey. In *IJCAI*, 5470–5477. ijcai.org.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4): 136–143.

Galhotra, S.; Shanmugam, K.; Sattigeri, P.; and Varshney, K. R. 2022. Causal Feature Selection for Algorithmic Fairness. In *SIGMOD Conference*, 276–285. ACM.

Grabowicz, P. A.; Perello, N.; and Mishra, A. 2022. Marrying Fairness and Explainability in Supervised Learning. In *FAccT*, 1905–1916. ACM.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*, 3315–3323.

Hendrickx, K.; Perini, L.; der Plas, D. V.; Meert, W.; and Davis, J. 2021. Machine Learning with a Reject Option: A survey. *CoRR*, abs/2107.11277.

Hertweck, C.; Heitz, C.; and Loi, M. 2021. On the Moral Justification of Statistical Parity. In *FAccT*, 747–757. ACM.

Hort, M.; Chen, Z.; Zhang, J. M.; Sarro, F.; and Harman, M. 2022. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *CoRR*, abs/2207.07068.

Hu, L.; and Kohler-Hausmann, I. 2020. What's sex got to do with machine learning? In *FAT**, 513. ACM.

Huitfeldt, A.; Stensrud, M. J.; and Suzuki, E. 2019. On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology*, 16.

Hutchinson, B.; and Mitchell, M. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *FAT*, 49–58. ACM.

Ilvento, C. 2020. Metric Learning for Individual Fairness. In *FORC*, volume 156 of *LIPIcs*, 2:1–2:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Ji, D.; Smyth, P.; and Steyvers, M. 2020. Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. In *NeurIPS*.

Jones, E.; Sagawa, S.; Koh, P. W.; Kumar, A.; and Liang, P. 2021. Selective Classification Can Magnify Disparities Across Groups. In *ICLR*. OpenReview.net.

Kamiran, F.; and Calders, T. 2009. Classifying without discriminating. In *Int. Conference on Computer, Control and Communication*, 1–6. IEEE.

Kamiran, F.; Karim, A.; and Zhang, X. 2012. Decision Theory for Discrimination-Aware Classification. In *ICDM*, 924–929. IEEE Computer Society.

Kamiran, F.; Zliobaite, I.; and Calders, T. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3): 613–644.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *ECML/PKDD (2)*, volume 7524 of *LNCS*, 35–50. Springer.

Kaur, D.; Uslu, S.; Rittichier, K. J.; and Durresi, A. 2023. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.*, 55(2): 39:1–39:38.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *NIPS*, 3146–3154.

Kenthapadi, K.; Lakkaraju, H.; Natarajan, P.; and Sameki, M. 2022. Model Monitoring in Practice: Lessons Learned and Open Challenges. In *KDD*, 4800–4801. ACM.

Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*, volume 67 of *LIPIcs*, 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Knowles, B.; Richards, J. T.; and Kroeger, F. 2022. The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency. *CoRR*, abs/2208.00681.

Koch, G.; and Kinder-Kurlanda, K. 2020. Source Criticism of Data Platform Logics on the Internet. *Historical Social Research*, 45(3): 270–287.

Kohler-Hausmann, I. 2019. Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination. *Northwestern University Law Review*, 113(5): 1163–1227.

Kong, Y. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *FAccT*, 485–494. ACM.

Koshiyama, A.; et al. 2021. Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. *SSRN Electronic Journal*.

Kroll, J. A.; Huey, J.; Barocas, S.; Felten, E. W.; Reidenberg, J. R.; Robinson, D. G.; and Yu, H. 2017. Accountable Algorithms. *U. of Penn. Law Review*, 165: 633–705.

Kronk, C. A.; and Dexheimer, J. W. 2020. Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow. *J. Am. Medical Informatics Assoc.*, 27(7): 1110–1115.

Kulynych, B.; Overdorf, R.; Troncoso, C.; and Gürses, S. F. 2020. POTs: protective optimization technologies. In *FAT**, 177–188. ACM.

Lazar, S. 2022. Legitimacy, Authority, and the Political Value of Explanations. *CoRR*, abs/2208.08628.

Lee, J.; Roh, Y.; Song, H.; and Whang, S. E. 2021. Machine Learning Robustness, Fairness, and their Convergence. In *KDD*, 4046–4047. ACM.

Lee, M. S. A.; and Floridi, L. 2021. Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. *Minds Mach.*, 31(1): 165–191.

Lee, M. S. A.; Floridi, L.; and Singh, J. 2021. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI Ethics*, 1(4): 529–544.

Lee, M. S. A.; and Singh, J. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *CHI*, 699:1–699:13. ACM.

Lipton, Z. C.; and Steinhardt, J. 2019. Research for practice: troubling trends in machine-learning scholarship. *Commun. ACM*, 62(6): 45–53.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed Impact of Fair Machine Learning. In *ICML*, volume 80 of *Proc. of Machine Learning Research*, 3156–3164. PMLR.

Lu, C.; Kay, J.; and McKee, K. 2022. Subverting machines, fluctuating identities: Re-learning human categorization. In *FAccT*, 1005–1015. ACM.

Madaio, M.; Egede, L.; Subramonyam, H.; Wortman Vaughan, J.; and Wallach, H. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Maity, S.; Xue, S.; Yurochkin, M.; and Sun, Y. 2021. Statistical inference for individual fairness. In *ICLR*. OpenReview.net.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2020. Survey on Causal-based Machine Learning Fairness Notions. *CoRR*, abs/2010.09553.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2021a. Machine learning fairness notions: Bridging the gap with real-world applications. *Inf. Process. Manag.*, 58(5): 102642.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2021b. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explor.*, 23(1): 14–23.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2022. Identifiability of Causal-based Fairness Notions: A State of the Art. *CoRR*, abs/2203.05900.

Manerba, M. M.; and Guidotti, R. 2021. FairShades: Fairness Auditing via Explainability in Abusive Language Detection Systems. In *CogMI*, 34–43. IEEE.

McDermott, D. 1976. Artificial intelligence meets natural stupidity. *SIGART Newsl.*, 57: 4–9.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6): 115:1–115:35.

Mejias, U. A.; and Couldry, N. 2019. Datafication. *Internet Policy Rev.*, 8(4).

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *FAT*, volume 81 of *Proc. of Machine Learning Research*, 107–118. PMLR.

Minow, M. 2021. Equality vs. Equity. *American Journal of Law and Equality*, 1: 167–193.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8: 141–163.

Mougan, C.; Álvarez, J. M.; Patro, G. K.; Ruggieri, S.; and Staab, S. 2022. Fairness implications of encoding protected categorical attributes. *CoRR*, abs/2201.11358.

Mulligan, D. K.; Kroll, J. A.; Kohli, N.; and Wong, R. Y. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW): 119:1–119:36.

Nogueira, A. R.; Pugnana, A.; Ruggieri, S.; Pedreschi, D.; and Gama, J. 2022. Methods and tools for causal discovery and causal inference. *WIREs Data Mining Knowl. Discov.*, 12(2).

Ntoutsi, E.; et al. 2020. Bias in data-driven artificial intelligence systems - An introductory survey. *WIREs Data Mining Knowl. Discov.*, 10(3).

Passi, S.; and Barocas, S. 2019. Problem Formulation and Fairness. In *FAT*, 39–48. ACM.

Pearl, J. 2009. *Causality: models, reasoning and inference, Second Edition*. Cambridge University Press.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *KDD*, 560–568. ACM.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2009. Measuring Discrimination in Socially-Sensitive Decision Records. In *SDM*, 581–592. SIAM.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2012. A study of top-k measures for discrimination discovery. In *SAC*, 126–131. ACM.

Pessach, D.; and Shmueli, E. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.*, 55(3).

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J. M.; and Weinberger, K. Q. 2017. On Fairness and Calibration. In *NIPS*, 5680–5689.

Plečko, D.; and Bareinboim, E. 2022. Causal Fairness Analysis. Technical Report R-70, Causal Artificial Intelligence Lab, Columbia University.

Quy, T. L.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining Knowl. Discov.*, 12(3).

Rahmattalabi, A.; and Xiang, A. 2022. Promises and Challenges of Causality for Ethical Machine Learning. *CoRR*, abs/2201.10683.

Raji, I. D.; Scheuerman, M. K.; and Amironesei, R. 2021. You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *FAccT*, 515–525. ACM.

Räz, T. 2021. Group Fairness: Independence Revisited. In *FAccT*, 129–137. ACM.

Richardson, B.; and Gilbert, J. E. 2021. A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. *CoRR*, abs/2112.05700.

Romei, A.; and Ruggieri, S. 2014. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, 29(5): 582–638.

Saha, D.; Schumann, C.; McElfresh, D. C.; Dickerson, J. P.; Mazurek, M. L.; and Tschantz, M. C. 2020. Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics. In *ICML*, volume 119 of *Proc. of Machine Learning Research*, 8377–8387. PMLR.

Saltz, J. S.; Skirpan, M.; Fiesler, C.; Gorelick, M.; Yeh, T.; Heckman, R.; Dewar, N. I.; and Beard, N. 2019. Integrating Ethics within Machine Learning Courses. *ACM Trans. Comput. Educ.*, 19(4): 32:1–32:26.

Scantamburlo, T. 2021. Non-empirical problems in fair machine learning. *Ethics Inf. Technol.*, 23(4): 703–712.

Schreuder, N.; and Chzhen, E. 2021. Classification with abstention but without disparities. In *UAI*, volume 161 of *Proc. of Machine Learning Research*, 1227–1236. AUAI Press.

Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; and Hall, P. 2022. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Technical Report 1270, NIST Special Publication.

Scott, K. M.; Wang, S. M.; Miceli, M.; Delobelle, P.; Sztandar-Sztanderska, K.; and Berendt, B. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *FAccT*, 2138–2148. ACM.

Shah, R. D.; and Peters, J. 2020. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3): 1514 – 1538.

Shahbazi, N.; Lin, Y.; Asudeh, A.; and Jagadish, H. V. 2022. A Survey on Techniques for Identifying and Resolving Representation Bias in Data. *CoRR*, abs/2203.11852.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press.

Spirtes, P.; and Zhang, K. 2016. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(3): 39:1–39:38.

Srivastava, M.; Heidari, H.; and Krause, A. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *KDD*, 2459–2468. ACM.

Stewart, R. T. 2022. Identity and the limits of fair assessment. *J. of Theoretical Politics*, 34(3): 415–442.

Tal, A. S.; Kuflik, T.; and Kliger, D. 2022. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics Inf. Technol.*, 24(1): 2.

Tizpaz-Niari, S.; Kumar, A.; Tan, G.; and Trivedi, A. 2022. Fairness-aware Configuration of Machine Learning Libraries. In *ICSE*, 909–920. ACM.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *FairWare@ICSE*, 1–7. ACM.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *W. Va. L. Rev.*, 123(3): 735–790.

Wagstaff, K. 2012. Machine Learning that Matters. In *ICML*. icml.cc / Omnipress.

Wan, M.; Zha, D.; Liu, N.; and Zou, N. 2022. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Trans. Knowl. Discov. Data*. Published online.

Wang, A.; Ramaswamy, V. V.; and Russakovsky, O. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *FAccT*, 336–349. ACM.

Watkins, E. A.; McKenna, M.; and Chen, J. 2022. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. *CoRR*, abs/2202.09519.

Wei, S.; and Niethammer, M. 2022. The fairness-accuracy Pareto front. *Stat. Anal. Data Min.*, 15(3): 287–302.

Weinberg, L. 2022. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *J. Artif. Intell. Res.*, 74: 75–109.

Wu, D.; and Liu, J. 2022. Involve Humans in Algorithmic Fairness Issue: A Systematic Review. In *iConference (1)*, volume 13192 of *LNCS*, 161–176. Springer.

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

Yule, G. U. 1903. Notes on the Theory of Association of Attributes in Statistics. *Biometrika*, 2(2): 121–134.

Zehlike, M.; Yang, K.; and Stoyanovich, J. 2023. Fairness in Ranking, Part I: Score-based Ranking. *ACM Comput. Surv.*, 55.

Zhou, J.; Chen, F.; and Holzinger, A. 2020. Towards Explainability for AI Fairness. In *xxAI@ICML*, volume 13200 of *LNCS*, 375–386. Springer.

Zliobaite, I. 2017. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.*, 31(4): 1060–1089.