Fairness and Ethics Under Model Multiplicity in Machine Learning

KACPER SOKOL, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia

MEELIS KULL, Institute of Computer Science, University of Tartu, Estonia

JEFFREY CHAN, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia

FLORA DILYS SALIM, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computer Science and Engineering, University of New South Wales, Australia

While data-driven predictive models are a strictly technological construct, they may operate within a social context in which benign engineering choices entail implicit, indirect and unexpected real-life consequences. Fairness of such systems – pertaining both to individuals and groups – is one relevant consideration in this space; it surfaces when data capture *protected* characteristics upon which people may be discriminated. To date, this notion has predominantly been studied for a *fixed* predictive model, often under different classification thresholds, striving to identify and eradicate undesirable, and possibly unlawful, aspects of its operation. Here, we backtrack on this assumption to propose and explore a novel definition of fairness where individuals can be harmed when one predictor is chosen ad hoc from a group of equally-well performing models, i.e., in view of utility-based *model multiplicity*. Since a person may be classified differently across models that are otherwise considered equivalent, this individual could argue for a predictor with the most favourable outcome, employing which may have adverse effects on others. We introduce this scenario with a two-dimensional example based on linear classification; then, we investigate its analytical properties in a broader context; and, finally, we present experimental results on data sets that are popular in fairness studies. Our findings suggest that such unfairness can be found in real-life situations and may be difficult to mitigate by technical means alone, as doing so degrades certain metrics of predictive performance.

Additional Key Words and Phrases: individual fairness, ethics, model view, model multiplicity, Rashōmon effect, machine learning, artificial intelligence, automated decision-making.

Highlights

- Utility-based individual fairness guarantees that a person receives the same prediction across a collection of models with identical or comparable predictive performance.
- When at least one prediction output by multiple equivalent models for a single individual is perceived as favourable, the person might argue for the precedence of this outcome.
- This notion of fairness is consistent with the Blackstone's ratio and "presumption of innocence" – lack of convincing evidence ought to warrant the most favourable treatment.
- Granting each person the most favourable decision afforded by a collection of equivalent models may degrade the overall predictive performance on the underlying task, especially when the employed family of models is highly expressive.
- A possible solution is to limit the number of admissible predictors by imposing appropriate modelling restrictions, which are consistent with the social context and the natural process governing the generation of the underlying data.

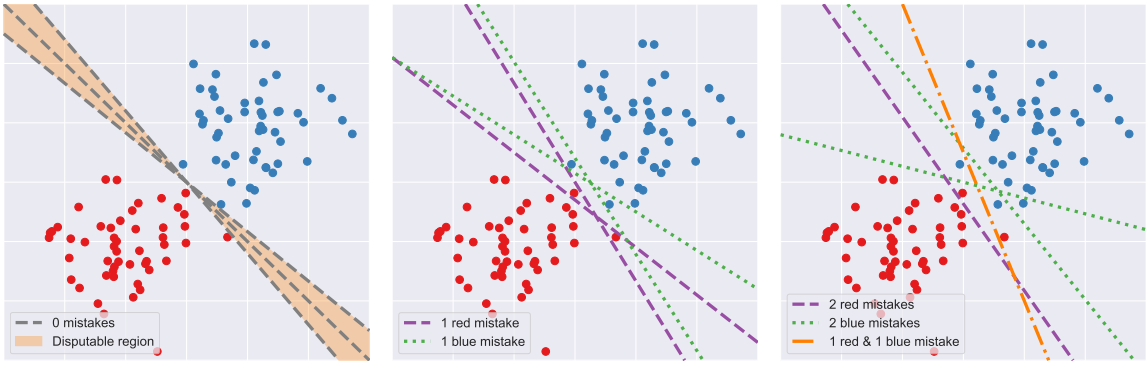
Source Code https://github.com/So-Cool/IndividualFairness

Authors' addresses: Kacper Sokol, Kacper.Sokol@rmit.edu.au, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Melbourne, Australia; Meelis Kull, Meelis.Kull@ut.ee, Institute of Computer Science, University of Tartu, Tartu, Estonia; Jeffrey Chan, Jeffrey.Chan@rmit.edu.au, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Melbourne, Australia; Flora Dilys Salim, Flora.Salim@unsw.edu.au, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia.

1 A NEW NOTION OF ALGORITHMIC FAIRNESS IN MACHINE LEARNING

Data-driven predictive models are making impressive strides across numerous domains, leading to their proliferation throughout businesses and society, where they either support decision-making or outright automate relevant tasks. This speedy adoption of Artificial Intelligence (AI) and Machine Learning (ML) algorithms, however, outpaces research investigating the potential harm of these techniques across different aspects of everyday life. While excluding human operators from the decisive process bears the promise of faster, more precise as well as consistent and replicable outputs that lack implicit human biases, pre-existing (historical) patterns concealed in training data may easily overshadow these benefits. The ubiquity of oftentimes erratically and narrowly tested and validated models can therefore contribute to and amplify problems with fairness, accountability and robustness of the predictive tasks being addressed. This is of particular concern when AI and ML models affect people – with possibly long-term or legally binding decisions – which has been documented in various contexts including banking, school admission, job screening and judiciary ruling [1, 20]. It is thus critical, among other desiderata, to ensure fairness of the resulting predictions with respect to protected characteristics, e.g., ethnicity or gender, when dealing with both individuals and population subgroups.

While mitigating disparate treatment of groups and individuals in view of protected attributes is of paramount importance, other notions of algorithmic fairness should not be neglected. Here, we explore a novel conceptualisation where instead of focusing on bias exhibited by a single model we deal with their collection [18] characterised by equal (or comparable) predictive performance according to a given evaluation strategy and metric [17]. For example, consider the classification scenario depicted in Figure 1a where three distinct predictors from the family of linear classifiers achieve 100% *accuracy* with respect to the displayed *validation set*. While models perfect in this respect do not make any *observable* mistakes, they may still suffer from *disputable regions* within which they disagree, giving rise to possible claims of unfair treatment voiced by previously unseen individuals residing in these spaces. Whenever we cannot guarantee perfect classification on the dedicated validation data, these considerations become more immediate as certain individuals from within this set may be treated differently across apparently equivalent models, as seen in Figures 1b and 1c. Notably, as we move towards more *expressive* families of predictive algorithms, their complexity and



(a) No errors – 100% accuracy. (b) One mistake – 99% accuracy. (c) Two errors – 98% accuracy.

Fig. 1. Multiplicity of linear models with predictive performance (accuracy) measured on dedicated validation data (scatter plot). Perfect classifiers – Panel (a) – may still yield unfair decisions for (initially unobserved) instances residing in disputable spaces. Non-perfect models can make the same type of a mistake on different data points – Panel (b) – which may not be reflected in the chosen performance metric; different types of a mistake – Panel (c) – may also go unnoticed in such a scenario.

parameterisation space grow, possibly making the observed phenomenon more prominent, especially for workflows whose optimisation is *greedy* or *stochastic*.

From the viewpoint of predictive performance there may be no objective reason to favour one model over another. This observation extends to confusion matrices (also known as contingency tables) since in certain scenarios they offer identical classification summaries despite distinct errors being made (as shown later in Table 1). In this setting a predictor could possibly be selected at random, however such an arbitrary choice may be dismissed by (adversely affected) individuals who are classified differently across this collection of models. Specifically, they can argue to be treated with the predictor granting them the most favourable outcome (especially if the selection procedure was ad hoc in the first place). Nonetheless, as the models under consideration are intrinsically different, many choices will inevitably result in some people benefiting at the expense of others, which in any case is difficult to justify. Notably, even if a collection of models boasts perfect performance with respect to a designated validation set, we still have to account for any disputable regions that are not covered by these data, hence unobserved, yet allow individuals who are placed therein to claim unfair treatment. This line of reasoning becomes especially important in the context of criminal justice, where a person might be entitled to be processed by the most advantageous model as long as it achieves certain performance for the (validation) population as a whole. This concept of fairness can be linked to the Blackstone's ratio [4] or the "presumption of innocence" [2], where lack of convincing evidence (or a unanimous vote) warrants the most favourable treatment – "It is better that ten guilty persons escape than that one innocent suffers."

To the best of our knowledge such considerations found limited recognition in fairness literature – albeit they have been reported in machine learning [17] – therefore our findings aim to establish solid foundations and provide initial analysis of *individual fairness under model multiplicity* [6]. With an abundance of predictive pipelines composed of diverse data processing techniques and model families, deriving bounds on the least and most favourable treatment of each individual in a dedicated *fairness* validation data set may be infeasible. Understanding the stability of each prediction under certain modelling assumptions could nonetheless enhance, or even guarantee, fairness, trustworthiness and accountability of important data-driven decisions, or hand them over to a human supervisor when necessary. Such a purview encourages incorporating various selection heuristics and criteria that account for properties beyond predictive performance; e.g., overall complexity or coverage of a model may be considered to increase the likelihood of its uniqueness, and classification with a reject/abstain option can be adopted to avoid individually unfair predictions [7, 8]. Another possible research question concerns cases where only relatively few classifiers that exhibit a given level of performance present an individual with a favourable outcome. For example, consider such a curated ensemble of predictive models taking the role of jurors where an overwhelming majority offers the less desirable decision. While the unfavourable ratio can be ignored with just a single model creating a precedent for a positive prediction, accounting for the proportion of classifiers and individual reasons for their respective outputs may provide important insights.

This paper investigates the scope and implications of individual unfairness under model multiplicity for crisp binary classifiers where one outcome is preferred over the other, e.g., grant or decline parole. This is especially important for data-driven decisions that influence human affairs, where choosing an arbitrary model from such a group may disparately affect certain people, posing ethical dilemmas when building and deploying predictive pipelines. While these observations are explicit for instances contained in a dedicated fairness evaluation set – as well as for performance validation data when dealing with non-perfect models – they are just as concerning, yet less pronounced, in the broader scope determined by disputable regions, which pertain even to seemingly perfect models as demonstrated by Figure 1. Specifically, this paper identifies, introduces, conceptualises and formally defines a novel notion of *individual fairness* stemming from *disputable regions* arising under two distinct types of *utility-based model multiplicity*, which is considered

in view of either *strict* or *relaxed* predictive performance criterion (Section 2). It further offers analytical treatment of the problem (Section 3), investigating the influence of the expressiveness (flexibility) of predictive models on their fairness, and consequences of granting each individual the best possible outcome using a fair-by-design ensemble model. Next, we propose a novel visualisation toolkit to help discover and analyse the degree of unfairness across the multiplicity spectrum – both as a high-level overview and an in-detail, instance-specific perspective – which we apply to real-life data (Section 4). We support our results by computing *discrepancy* and *ambiguity* of the underlying predictive pipelines [17]; we also analyse the utility of the fair-by-design ensemble model for each data set. Finally, before we conclude our work and outline future research directions (Section 6), we discuss relevant literature (Section 5).

These contributions outline concepts fundamental to our notion of individual fairness. To introduce them, we first investigate diverse binary classification scenarios for a two-dimensional synthetic data set. In the process we uncover caveats and assumptions relevant to both the modelling task and the approach used to evaluate predictive performance; these findings help us to identify the core principles of fairness under utility-based model multiplicity and devise strategies to address (unintended) consequences of this phenomenon. Our preliminary results show that allowing an individual to choose the (most favourable) model can have detrimental effects on the overall predictive performance of the classification system, particularly so when the employed AI or ML model is relatively expressive, therefore its decision boundary is flexible. This observation promotes utilising an approach to limit the number of admissible predictors by imposing upon them (modelling) restrictions consistent with the natural process governing the generation of the underlying data, thus also making these models more interpretable [22]. Our graphical investigative tools as well as the ambiguity and discrepancy metrics support these findings; we apply and explain them for three real-life data sets popular in fairness studies: Credit Approval, German Credit and Adult. These experiments are based on top-performing classification workflows published in OpenML [24], thus ensuring credibility of our results.

2 NAVIGATING MODEL MULTIPLICITY IN VIEW OF INDIVIDUAL FAIRNESS

The notion of fairness studied by this paper is built upon the *model multiplicity* phenomenon understood here as existence of a collection of data-driven models that are indistinguishable in terms of their predictive performance under a fixed evaluation strategy [6, 17]. Furthermore, in this work we are interested in *crisp binary classification* in which one outcome is universally preferred to the other by individuals whose case is being decided. Therefore, a model $f : \mathcal{X} \to \mathcal{Y}$ classifies an instance $x \in \mathcal{X}$ as $f(x) = \hat{y}$, where $\hat{y} \in \mathcal{Y} \equiv \{0, 1\}$ and 1 represents a favourable outcome. Additionally, predictive performance of such a model is measured on a predetermined validation data subset $\widetilde{X} \subseteq \mathcal{X}$ accompanied by annotations $\widetilde{Y} \subseteq \mathcal{Y}$, using a selected metric $m : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ calculated as $m\left(f(\widetilde{X}), \widetilde{Y}\right)$. In our initial analysis we rely on linear, polynomial, k-nearest neighbours and decision tree classifiers applied to a synthetic two-dimensional data set, which helps us to demonstrate the principles of individual fairness under utility-based model multiplicity through visual inspection and hand-crafted examples; nonetheless, all of our results can be easily generalised beyond this restrictive setting.

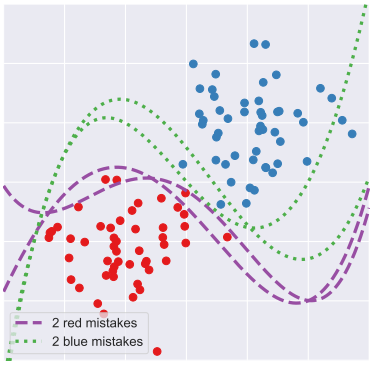


Fig. 2. Polynomial model multiplicity with two mistakes.

While in principle model multiplicity may span a diverse range of classifiers, we restrict our considerations to a *confined family of predictive functions* (inspired by such a formalisation of curves). This is desirable as distinct data-driven algorithms exhibit unique characteristics that translate into different shapes of their respective decision boundaries, allowing for diverse mistakes

to be made, as seen when comparing Figures 1c and 2. Therefore, a **family of data-driven models** $\mathcal{F}$ consists of classifiers $f \in \mathcal{F}$ based upon a single predictive algorithm and trained on a fixed data set, i.e., $\mathcal{F} \subseteq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$. It can further be constrained by imposing restrictions on the parameterisation space or optimisation approach, among other properties, used with the selected method, e.g., a collection of decision trees no deeper than 7. Notably, algorithms whose training procedure is stochastic or greedy may yield distinct models from a single specification when run multiple times. A family of models can also, by extension, include more complex predictive workflows built with pre- or post-processing steps such as input normalisation or output calibration.

To address unfairness under model multiplicity we further identify a subset of classifiers from a single family that all share a fixed level of predictive performance (i.e., *utility*). To this end, we need to select an evaluation metric; for simplicity, we employ accuracy throughout this paper, which is defined as $m_{\text{acc}}(\hat{Y}, Y) = \sum_{\hat{y},y \in \hat{Y},Y} \mathbb{1}_{\hat{y}=y} \;/\; |Y|$, where $\hat{Y} = f(X)$ are predictions and Y are annotations for a data set X. Moreover, we require a designated collection of labelled instances $(\widetilde{X}, \widetilde{Y})$ to serve as a dedicated *predictive performance validation set*. This adds to the existing training and validation data used to fit the model and tune its hyper-parameters, both of which constitute an integral part of any AI or ML workflow. Notably, the performance validation set $\widetilde{X}$ can also be employed to evaluate our notion of fairness, however in certain cases – reviewed in Section 3 – separating the two may be beneficial.

In this setting, *utility-based model multiplicity* $\mathcal{F}_\epsilon$ is determined by a fixed level of predictive performance $\epsilon \in \mathbb{R}$ shared by classifiers from a single family of data-driven models $\mathcal{F}$. In particular, there are two *meaningful* viewpoints on multiplicity:

population-based where the performance of each model is computed for the entire data space $\mathcal{X}$, which may be infeasible given a possibly infinite number of qualifying models (e.g., slight alterations of a linear classifier); and

validation-based where the performance is measured on a dedicated validation set $\widetilde{X}$, making the problem tractable.

In view of the former, the three models shown in Figure 1a are distinct, whereas based on the latter they are indistinguishable, i.e., they belong to a single multiplicity class $\mathcal{F}_\epsilon$. Notably, either of these two notions is distinct from a purely *theoretical non-uniqueness* of models within a single family, where the same decision boundary – thus unobservable changes – can be achieved with different realisations of a predictive pipeline; for example, see Figure 3 depicting two structurally different decision trees classifying the entire data space in the same fashion. Throughout this paper we are predominantly concerned with the validation-based multiplicity, which is outlined in Definition 1. Such a setting simplifies our considerations and allows us to avoid working with a possibly infinite number of models.

Definition 1. ***Utility-based model multiplicity*** *$\mathcal{F}_\epsilon$ is captured by a set of classifiers from a single family of models $\mathcal{F}$, all of which offer fixed predictive performance ϵ for a given metric m computed on a dedicated validation data set and*

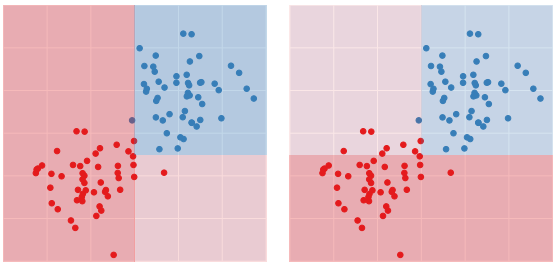


Fig. 3. Theoretical multiplicity of decision trees. While the models are distinct their predictions are identical for the entire data space.

its labels $(\widetilde{X}, \widetilde{Y})$*:*

$$\mathcal{F}_\epsilon \coloneqq \left\{ f \in \mathcal{F} : m\left(f(\widetilde{X}), \widetilde{Y}\right) = \epsilon \right\}.$$

It is important to observe that the choice of the performance metric determines which models are considered equivalent. For example, all of the classifiers shown in Figure 1c make two mistakes, hence are indistinguishable with respect to *accuracy*; however, a different metric, e.g., *precision*, *recall* or *specificity*, would make the violet, orange and green models distinct. Since all of the predictive performance metrics for classification tasks are derived from confusion tables, instead of relying on calculated numerical values we can refer directly to the underlying contingency matrices. Table 1 lists specific errors made for the red and blue classes by the models shown in Figure 1c, where the green predictors incorrectly classify two blue points, the violet model errs on two red examples and the orange classifier mistakes one blue and one red instance. However, even such a fine-grained approach is insufficient to capture the *same type of a mistake made on different data points*, i.e., individual predictions, motivating our notion of individual fairness under utility-based model multiplicity. We are therefore interested in classifiers that are indistinguishable in terms of performance measured on a dedicated population, but provide a distinct class assignment for certain individuals (who are not necessarily included in this data set).

A generalisation of such individual mistakes – inspired by the *population-based* model multiplicity viewpoint – are *disputable spaces* $\mathring{\mathcal{X}}_{\mathcal{F}_\epsilon}$, an example of which is shown in Figure 1a. This extension – outlined by Definition 2 – is useful when the designated *fairness* validation data set $\mathring{X} \subseteq \mathcal{X}$ is sparse and cannot capture inconsistency of predictions at the desired level of detail, i.e., $\mathring{X} \notin \mathring{\mathcal{X}}_{\mathcal{F}_\epsilon}$, hence no instance from $\mathring{X}$ is placed within the disputable spaces $\mathring{\mathcal{X}}_{\mathcal{F}_\epsilon}$. Notably, the shape of these regions is determined by the model family $\mathcal{F}$; for example, consider the polynomial classifiers shown in Figure 2 for which such spaces are much more complex than in the case of linear classification (refer to Figure 1). Logical models, on the other hand, constrain disputable regions to (hyper-)rectangles since they impose axis-parallel splits on the data space (see Figure 3).

Definition 2. *A* ***disputable space*** *(or region)* $\mathring{\mathcal{X}}_{\mathcal{F}_\epsilon} \subseteq \mathcal{X}$ *for utility-based model multiplicity* $\mathcal{F}_\epsilon$ *is given by:*

$$\mathring{\mathcal{X}}_{\mathcal{F}_\epsilon} \coloneqq \left\{ x \in \mathcal{X} : \exists\, f_i, f_j \in \mathcal{F}_\epsilon \;\; s.t. \;\; f_i(x) \neq f_j(x) \right\},$$

for a chosen predictive metric m *and labelled (performance) validation set* $(\widetilde{X}, \widetilde{Y})$*, where* $\forall\, f \in \mathcal{F}_\epsilon \;\; m\left(f(\widetilde{X}), \widetilde{Y}\right) = \epsilon$.

If the main property considered while choosing a classifier $f \in \mathcal{F}_\epsilon$ is predictive performance, then from the perspective of utility-based model multiplicity $\mathcal{F}_\epsilon$ all such predictors may be viewed as equivalent, without any particular preference for a given classifier. Lacking some further, well-defined selection criteria, an arbitrary choice can nonetheless lead to unfair treatment of individuals given the existence of an equally suitable model that may provide these people with a more favourable outcome. For example, consider the green classifiers shown in Figure 1c; both of them have identical confusion matrices (Table 1c) yet only one of the two borderline blue individuals may be assigned the preferred outcome

	red	blue
red	48	0
blue	2	50

(a) Violet model.

	red	blue
red	49	1
blue	1	49

(b) Orange model.

	red	blue
red	50	2
blue	0	48

(c) Green models.

Table 1. Even though confusion matrices are different for the violet, orange and green classifiers shown in Figure 1c, they cannot convey the two distinct errors made by the green models on individual instances – see Panel (c) above.

– the red class – depending on the model choice. Framing such a scenario as an automated decision between granting (red) or denying (blue) parole in an (algorithmically-supported) judicial hearing, performance-based indistinguishability of models becomes an important factor. While ideally the task should be to minimise the scope of any disputable region, i.e., to deal with $\mathring{X}_{\mathcal{F}_\epsilon}$, in this paper we focus on a designated fairness validation set $\mathring{X} \subseteq \mathcal{X}$ as given by Definition 3, which outlines this type of *model-based disparate treatment*. Notably, such a notion of fairness can be expanded from individuals to *groups* by comparing the impact of multiplicity, e.g., ratios of affected instances, across them.

Definition 3. *A classifier $f \in \mathcal{F}_\epsilon$ is **fair** towards individuals $x \in \mathring{X} \subseteq \mathcal{X}$ in view of utility-based model multiplicity $\mathcal{F}_\epsilon$ iff $\forall\, f' \in \mathcal{F}_\epsilon\ \forall\, x \in \mathring{X}\ \ f(x) = f'(x)$.*

In such a setting, eliminating unfairness caused by disparate decisions found across admissible models entails treating each individual with the most beneficial predictor $f \in \mathcal{F}_\epsilon$. This observation leads us to define a new, universally fair model $f^\star$ built by aggregating all of the classifiers that are equivalent under utility-based model multiplicity $\mathcal{F}_\epsilon$. We do so by incorporating disputable regions into the decision space predicted with the more favourable outcome, therefore maximally growing its coverage. This approach may be considered as a simple model ensemble where we are interested in the best, rather than the average, result. Figure 4, for example, shows such a composition of classifiers for the models depicted in Figure 1b, assuming that the red class (1) is preferred to the blue class (0). The new decision boundary is constructed by joining linear models at their intersection – the two green segments in the visible frame – as to maximise the area predicted with the favourable outcome. Notably, it is possible that the fair model itself does not belong to the family, i.e., $f^\star \notin \mathcal{F}$, which is captured by the aforementioned example – the decision boundary of $f^\star$ is not linear. The individually fair model $f^\star$ is formalised in Definition 4.

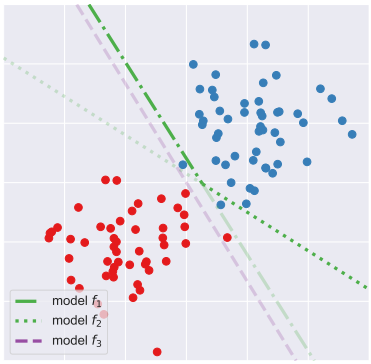


Fig. 4. Maximising the feature space predicted with the more favourable outcome (red) by joining (at intersections) the two green and one purple (top left and out of view) models $f \in \mathcal{F}_\epsilon$ shown in Figure 1b gives an individually fair classifier $f^\star$ under utility-based model multiplicity $\mathcal{F}_\epsilon$.

Definition 4. *An **individually fair classifier** $f^\star$ under utility-based model multiplicity $\mathcal{F}_\epsilon$ is defined as:*

$$f^\star(x) \coloneqq \max_{f \in \mathcal{F}_\epsilon} f(x),$$

for any instance $x \in \mathcal{X}$. (Recall that 1 is the preferred prediction in our binary classification setting.)

Thus far we operated under strict utility regime ϵ, nonetheless this notion can be relaxed by allowing a certain deviation from the desired level of predictive performance. We can argue in favour of such an approach given that the utility is measured on a data subset, an expansion or contraction of which is likely to yield some variation in predictive performance. We can specify the allowed *tolerance* through an additional parameter $\delta \in \mathbb{R}^+$ – where the performance band is denoted with $\epsilon \pm \delta$ – which extends the utility-based model multiplicity $\mathcal{F}_\epsilon$ provided in Definition 1 to $\mathcal{F}_{\epsilon\pm\delta}$ as outlined in Definition 5. An alternative operationalisation of this concept is to round the predictive performance ϵ to a specified decimal point δ, written as $\epsilon \simeq \delta$; for example, $\epsilon \simeq 10^{-2}$ indicates ϵ rounded to the second decimal place. We use the latter approach throughout this paper to streamline our exploration of real-life data sets.

Definition 5. ***Relaxed utility-based model multiplicity*** *$\mathcal{F}_{\epsilon\pm\delta}$ is a collection of classifiers from across a single family of models $\mathcal{F}$ that exhibit a level of predictive performance within a fixed range $[\epsilon - \delta,\ \epsilon + \delta]$ for a chosen metric m:*

$$\mathcal{F}_{\epsilon\pm\delta} \coloneqq \left\{ f \in \mathcal{F} : \epsilon - \delta \leq m\left(f(\widetilde{X}), \widetilde{Y}\right) \leq \epsilon + \delta \right\},$$

where $(\widetilde{X}, \widetilde{Y})$ is a dedicated validation data set with labels, and δ is the tolerance of the predictive performance level ϵ.

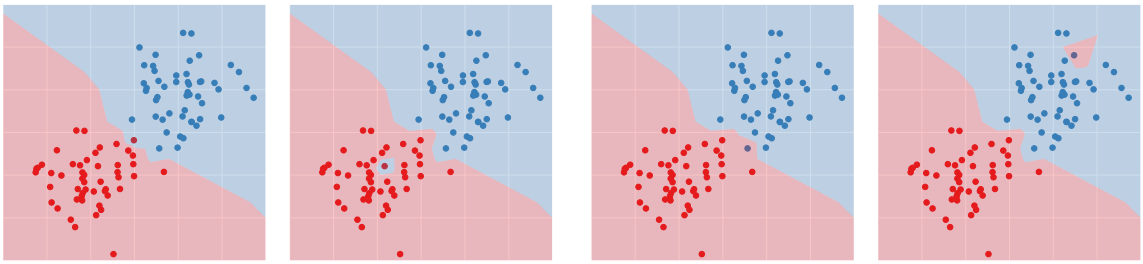
3 TOWARDS MODEL-BASED FAIRNESS OF INDIVIDUAL PREDICTIONS

Operationalising the proposed notion of fairness in view of utility-based model multiplicity brings about various challenges. Among others, the composition of the selected performance and fairness validation sets is a key to model equivalence and identification of unfair behaviour respectively. Their representativeness, density and relative distribution should therefore be carefully considered, and potentially expanded or augmented over time. In addition to the modelled data space, special attention ought to also be placed on inherent characteristics of the classifiers themselves. Overall stability of the model family, which may depend on the stochasticity or greediness of the corresponding training procedure, as well as proximity of instances to class boundaries – for example, conveyed by prediction uncertainty and model (over- or under-) confidence – and behaviour of a classifier in sparse data regions all play a role in the volatility of automated decision-making.

Moreover, the inherent expressiveness and flexibility of the chosen model, and more broadly any data modelling workflow that can be built upon it by incorporating new steps such as data pre-processing or feature engineering, can also be problematic. The operationalisation of the fair model given by Definition 4 should be scrutinised as well since it may lead to a drop in predictive performance, which can be seen in Figure 4 where $f^\star$ misclassifies two instances whereas the base models $f \in \mathcal{F}_\epsilon$ make just one mistake each. All of these observations create opportunities (predominantly on technical grounds) for individuals adversely affected by an automated decision to easily challenge fairness of this process under utility-based model multiplicity.

One conceivable attack on our notion of individual fairness can come from *under-specification* of the selected family of classifiers $\mathcal{F}$ or any predictive workflow built around it. Expressive classifiers may be flexible enough to single out individual instances and assign an arbitrary prediction to them, for example due to their inherent complexity or parameterisation scope. Such freedom can adversely influence model-based fairness by permitting any two instances to swap predictions – regardless of their placement in the data space – while maintaining the desired level of performance ϵ, i.e., operating within the designated utility-based model multiplicity class $\mathcal{F}_\epsilon$. This can easily be achieved for classifiers with considerable parameterisation scope such as deep neural networks, but it is also relevant to simpler models, e.g., k-nearest neighbours, when they are misconfigured as shown in Figure 5.

Flexibility of a model f from a given family $\mathcal{F}$ can be defined through a proxy such as complexity $\Omega(f)$ or Vapnik–Chervonenkis dimension [25], and imposed as an additional constraint. The precise specification of such a metric may be unique to each model family; for example, the number of non-zero parameters for linear models, the highest coefficient degree for polynomial classifiers and the depth, width or number of instances per leaf for decision trees. More broadly, expressiveness, hence flexibility, of models and workflows built from them may be controlled by ensuring diversity of training data, fixing a lower bound on confidence of each decision, abstaining from predictions, restricting model parameterisation, enforcing regularisation such as pruning for trees and LASSO for linear models, or limiting overfitting via alternative mechanisms. Otherwise, with a budget of errors given by the required level of predictive performance, an excessive number of people could abuse this notion and claim unfair treatment.



(a) One red instance is misclassified. (b) One blue instance is misclassified.

Fig. 5. Model multiplicity with one error (99% accuracy) for a k-nearest neighbours classifier ($k = 1$). Expressive models can be flexible enough to allow for an arbitrary classification of every instance, e.g., by manipulating training data. To mitigate such scenarios appropriate safeguards need to be established through constraints on model parameterisation (e.g., $k \geq 42$) or regularisation being required as part of a predictive workflow built on top of a selected model family $\mathcal{F}$.

The aforementioned strategies tasked with constraining the model space can be complemented by (use case-specific) operational requirements. For example, instead of a single performance metric employed to determine model equivalence, their hierarchy can be implemented, measuring accuracy first, and followed by precision and recall to address any ties. More broadly, such an approach could be applied directly to the confusion matrix of a classifier by sequentially imposing restrictions on its individual entries. Arguably, one could optimise exclusively for *recall* and *specificity* to maximise the number of favourable decisions; however, any such course of action is just a stopgap as it is likely to lack a solid foundation rooted in the domain-specific aspects of the modelled problem. A more meaningful heuristic should therefore rely on well-defined properties of the utilised model [6] – for example, monotonicity of a particular feature with respect to the prediction – which strategy is consistent with some definitions of explainability in AI and ML [19, 22, 23].

Lacking a (close to) *unique* classifier – obtained by imposing various desiderata to narrow down the scope of the chosen family of predictive models, thus making any alternative difficult or impossible to find – we may need to instead rely on the fair classifier $f^\star$ outlined by Definition 4. Treating each individual with the best available model in this fashion may however degrade the overall predictive performance for the task at hand, making the fair application of data-driven decisions on a par with less effective classifiers to begin with. For example, this phenomenon can be observed for the fair model $f^\star$ shown in Figure 4. Notably, any model family $\mathcal{F}_\epsilon$ subject to *population-* or *validation-based* multiplicity, i.e., with meaningful alternatives, is bound to suffer from disputable spaces. Depending on the density of data in these regions and their size, the fair model $f^\star$ will offer *no worse recall* but *no better specificity* than any individual model $f \in \mathcal{F}_\epsilon$ as stated by Proposition 1. This change in predictive performance is especially prominent (reaching its limit) when the chosen family of models $\mathcal{F}$ is expressive enough to single out individual data points, in which case the fair classifier $f^\star$ will assign the most favourable class to every instance – 0% specificity and 100% recall – as given by Proposition 2.

Proposition 1. ***Specificity*** *m_s of an individually fair classifier $f^\star$ under utility-based model multiplicity $\mathcal{F}_\epsilon$ is no better than that of any other classifier $f \in \mathcal{F}_\epsilon$:*

$$\forall\, f \in \mathcal{F}_\epsilon \;\; m_s\left(f^\star(X), \mathcal{Y}\right) \leq m_s\left(f(X), \mathcal{Y}\right),$$

where $(\mathcal{X}, \mathcal{Y})$ *is the labelled data space. Additionally,* ***recall*** m_r *of* $f^\star$ *is no worse than that of any* $f \in \mathcal{F}_\epsilon$*:*

$$\forall\ f \in \mathcal{F}_\epsilon \quad m_r\left(f^\star(\mathcal{X}), \mathcal{Y}\right) \geq m_r\left(f(\mathcal{X}), \mathcal{Y}\right).$$

The result above (Proposition 1) can be interpreted geometrically as expanding the space predicted as the more favourable outcome (red in our case) with the entirety of disputable regions – see Figure 4 for an example. It follows directly from observing individual predictions under $f^\star$.

- Each data point from the preferred class (red) predicted as such (true positive) will not be affected, but misclassified positive instances (false negatives) may be corrected by $f^\star$ – a possible improvement in **recall**.
- Each data point from the undesirable class (blue) predicted as such (true negative) may be (mis)classified as positive, but misclassified negative instances (false positives) will not be affected by $f^\star$ – a possible drop in **specificity**.

Furthermore, the fair classifier $f^\star$ for models $f \in \mathcal{F}_\epsilon$ from a family that is flexible enough to predict any individual with an arbitrary class will assign the preferred output to all data, from which Proposition 2 follows. These observations alone should encourage the developers of AI and ML systems to strive for high utility computed on a comprehensive and representative validation set to minimise the scope of any disputable regions.

Proposition 2. *When a family of classifiers* $\mathcal{F}_\epsilon$ *under utility-based model multiplicity is expressive enough to assign a selected class* $c \in \mathcal{Y}$ *to any data point* $x \in \mathcal{X}$*, i.e.,*

$$\forall\ x \in \mathcal{X}\ \exists\ f \in \mathcal{F}_\epsilon \quad s.t.\ \ f(x) = c,$$

the corresponding individually fair classifier $f^\star$ *achieves 0%* ***specificity*** *(*m_s*) and 100%* ***recall*** *(*m_r*):*

$$m_s\left(f^\star(\mathcal{X}), \mathcal{Y}\right) = 0\% \qquad m_r\left(f^\star(\mathcal{X}), \mathcal{Y}\right) = 100\%,$$

in which case every point is assigned the most favourable outcome.

Given the benefits of finding a classifier $f \in \mathcal{F}_\epsilon$ that is favourable for a specific individual, the concept of fairness under utility-based model multiplicity can be framed as an adversarial challenge. In such a setting people disparately affected by an automated decision can confront the owner of the underlying predictive model by building an equivalent classifier (performance-wise) that offers the desired outcome instead. These "adversaries" attempt to identify a predictor f' from within the employed family of models $\mathcal{F}_\epsilon$ – ensuring that it complies with all of the restrictions imposed by $\mathcal{F}_\epsilon$, including the predetermined level of predictive performance ϵ – that assigns a selected individual the most favourable class. The party responsible for building and deploying the challenged model, on the other hand, ought to minimise the possible number of such claims by reducing the size of any disputable regions. This tug-of-war is bound to have a positive effect on the predictive model in question by iteratively improving its accountability, robustness and overall quality.

Facilitating this back-and-forth process, however, requires releasing (a subset of) relevant training data as well as performance and fairness validation sets (in addition to the specification of the utilised model family $\mathcal{F}$), which may be problematic due to inherent trade secrets. While distributing training data cannot be easily sidestepped, predictive performance of a model may be assessed without access to evaluation data as the classifier itself can instead be submitted for testing – akin to how ML competitions are run. Notably, having a comprehensive performance validation set that faithfully represents all of the individual cases is advantageous for the model creators as it restricts the number of admissible classifiers. Lastly, publishing a fairness validation set may also be beneficial to the owner of the model

under investigation, encouraging a broader community to identify unfair predictions – and, more generally, disputable spaces – as well as engendering trust in the deployed model itself. This decoupling of the two – predictive performance validation data and fairness validation set – may therefore be desirable, especially that the latter *does not need to be labelled* since we are only interested in disparate treatment of these instances across equivalent classifiers $f \in \mathcal{F}_\epsilon$ to assess model-based individual fairness.

4 REAL-LIFE IMPACT OF MODEL MULTIPLICITY ON INDIVIDUAL FAIRNESS

Having explored the theoretical aspects of individual fairness under model multiplicity, we shift our attention to the real-life occurrence and impact of this phenomenon, which we investigate by studying pre-existing collections of predictive pipelines. In particular, we focus on a common scenario when such considerations may arise, namely iterating over machine learning models during their development. We reconstruct this setting by employing a selection of top-performing binary classification workflows available in the OpenML repository [24], concentrating on data sets popular in fairness research. To study the extent of individuals being classified inconsistently for a chosen family of models $\mathcal{F}_\epsilon$, we propose a toolkit capable of:

- measuring the severity of this phenomenon,
- identifying the affected instances, and
- effectively communicating these findings.

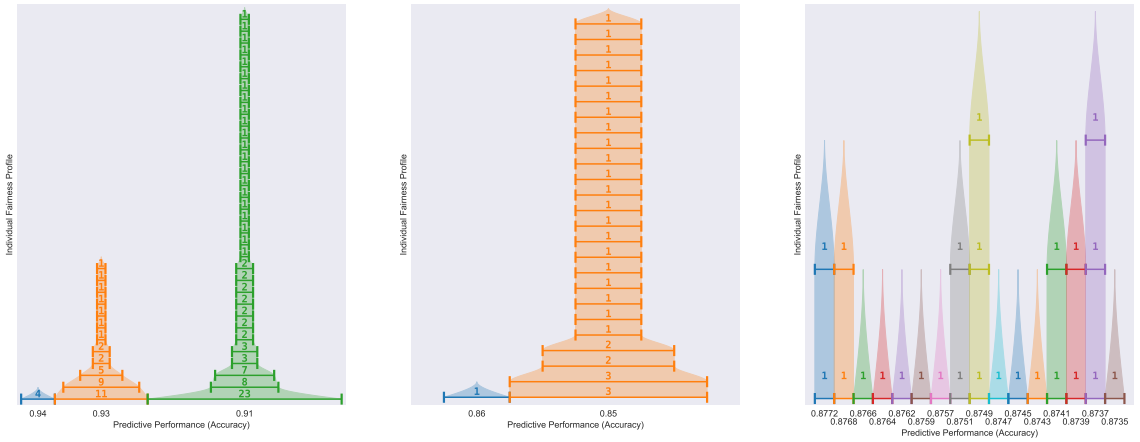
As we have observed earlier in Section 2, standard performance metrics derived from confusion matrices are insufficient to this end – see Table 1c for an example. To fill this gap we employ the numerical measures of multiplicity – *ambiguity* and *discrepancy* – introduced by Marx et al. [17], and build bespoke graphical analytic instruments around them. Additionally, we develop a comprehensive visualisation tool that summarises the prediction structure for each individual performance band ϵ of a model family $\mathcal{F}$ – a classifier perspective called *stability profile*. We complement this approach with a fine-grained inspection plot that captures precise classification results for all data points across chosen performance bands ϵ – a prediction view named *fairness profile*. We first introduce and explain these tools, and then apply them to real-life data sets that are common in fairness studies, using model families $\mathcal{F}$ available on OpenML. We expand our analysis by:

- investigating the same selection of properties when the performance bands are relaxed (Definition 5); and
- inspecting the change in utility (accuracy in our case) for fair models $f^\star$ (Definition 4) built for the aforementioned data.

The code needed to reproduce our experiments and the results reported in this section is published on GitHub[1].

Stability profiles are our first visual diagnostic tool; they provide an overview of predictive volatility with respect to the entire fairness validation set from the perspective of a group of classifiers sourced from the chosen model family $\mathcal{F}$ – see Figure 6. They capture the consistency of individual predictions across models belonging to distinct performance bands ϵ by grouping them in colour-coded pyramids. Each stack informs us of the number – quantity of horizontal segments – and frequency – width thereof – of unique prediction vectors (i.e., different class assignments) for the entire fairness validation data set. The most desirable, i.e., the most fair, shape of such a pyramid is one amassing all the models in a single segment at the bottom as this configuration indicates that every classifier offers the same set of predictions.

[1] https://github.com/So-Cool/IndividualFairness

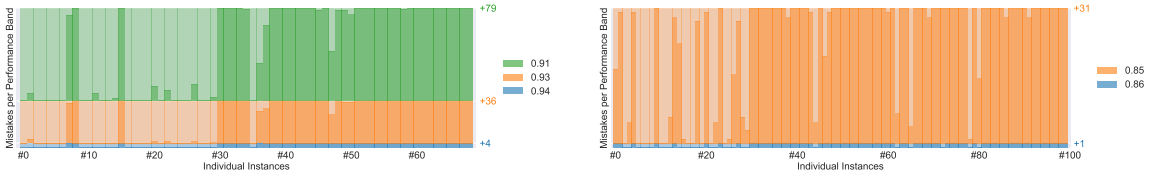


(a) Credit Approval – top 3 performance bands spanning 119 models.

(b) German Credit – top 2 performance bands spanning 32 models.

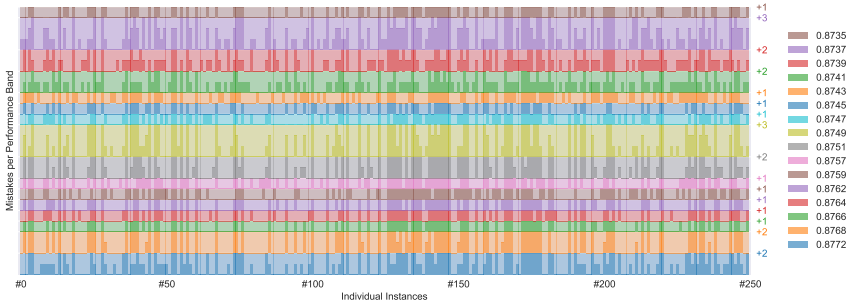
(c) Adult – top 16 performance bands spanning 25 models.

Fig. 6. Stability profiles for the top n performance bands of each data set. This visual inspection tool displays the counts of unique prediction vectors, i.e., class assignments across the entire fairness validation set, for a collection of models from a chosen family $\mathcal{F}$ grouped by predictive performance – depicted as stacks of different colours – measured on a dedicated validation set, in our case using accuracy. For example, the orange pyramid in Panel (a) reveals 36 models with 93% accuracy distributed across 12 distinct prediction vectors – the number of horizontal segments – as shown by the reported counts and reflected in the width of each bar.



(a) Credit Approval – top 3 performance bands with 69 instances.

(b) German Credit – top 2 performance bands with 100 instances.

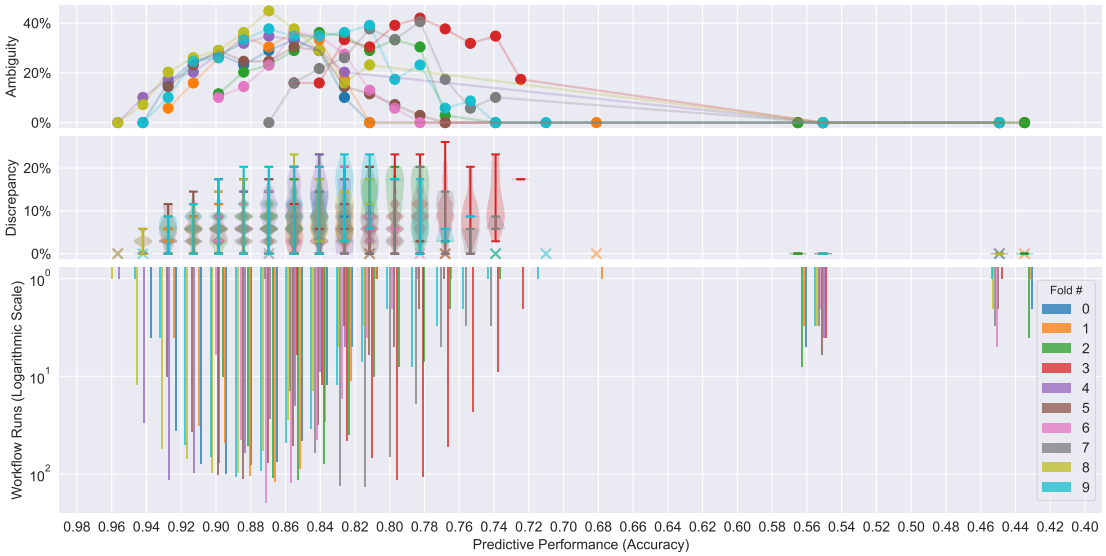


(c) Adult. For clarity, the profile shows a sample of 250 instances out of 586 individuals treated unfairly by any of the 25 models spanning the top 16 performance bands depicted in Figure 6c. There are a total of 4,885 data points in this validation set.

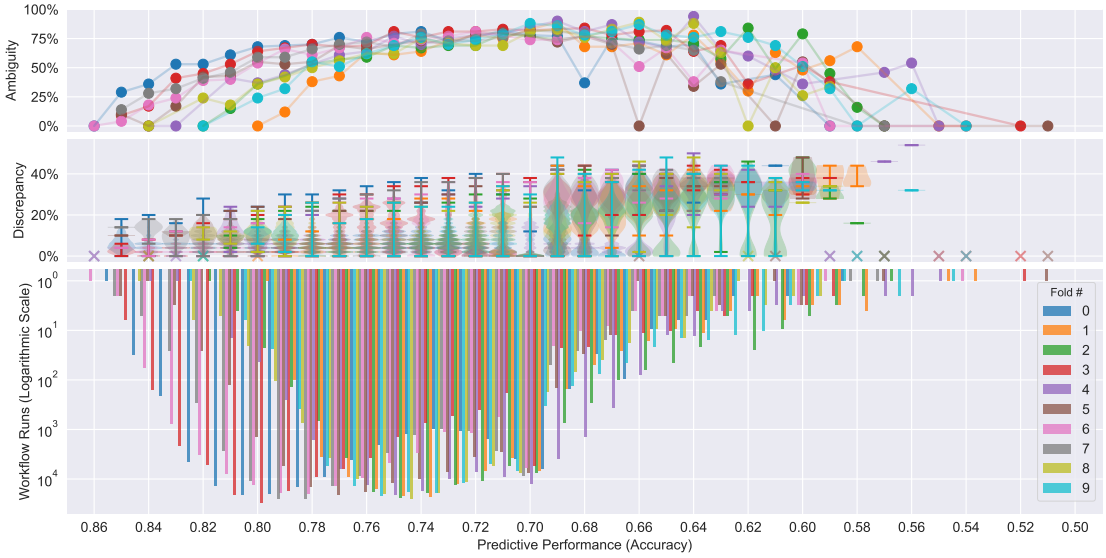
Fig. 7. Fairness profiles plotted for the chosen data sets. This visual inspection tool depicts the behaviour of individual predictive models – spread across rows – for each instance – captured by a single column – in the fairness validation set. The performance bands are colour-coded and the saturation of cells differentiates between the two possible predictions. The *summary variant* of a fairness profile – used here – sorts the predictions vertically in every column, separately for each performance band, to improve readability of large fairness validation data sets, especially since they cannot be easily ordered. For example, Panel (c) displays the behaviour of 25 models (rows) split across 16 (colour-coded) performance bands shown previously in Figure 6c for a selection of individual instances.

Fairness profiles complement stability profiles by focusing on the behaviour of each model with respect to individual instances as shown in Figure 7. They visualise how distinct classifiers – one per row – grouped together by predictive performance bands – captured by colour-coding – assign a (binary) decision – differentiated by the saturation of each cell – to every instance – one per column – in the designated validation data set. These plots shed light on the volatility of predictions across individuals, hence fairness of their classification. The stability and fairness profiles are linked through shared colouring of the predictive performance bands, for example, compare Figures 6a and 7a. A fairness profile can be plotted *faithfully*, thus accurately depicting the classification of every instance for all the models; alternatively, a *summary* variant that aggregates distinct predictions across models within a performance band (i.e., a coloured segments) by sorting them for each instance (i.e., a column) separately may instead be preferred due to its improved readability – we use this approach in Figure 7. Ideally, each column ought to be in a single – light or dark – shade, which indicates a consistent, thus fair, prediction output by models in a single level or throughout different levels of predictive performance.

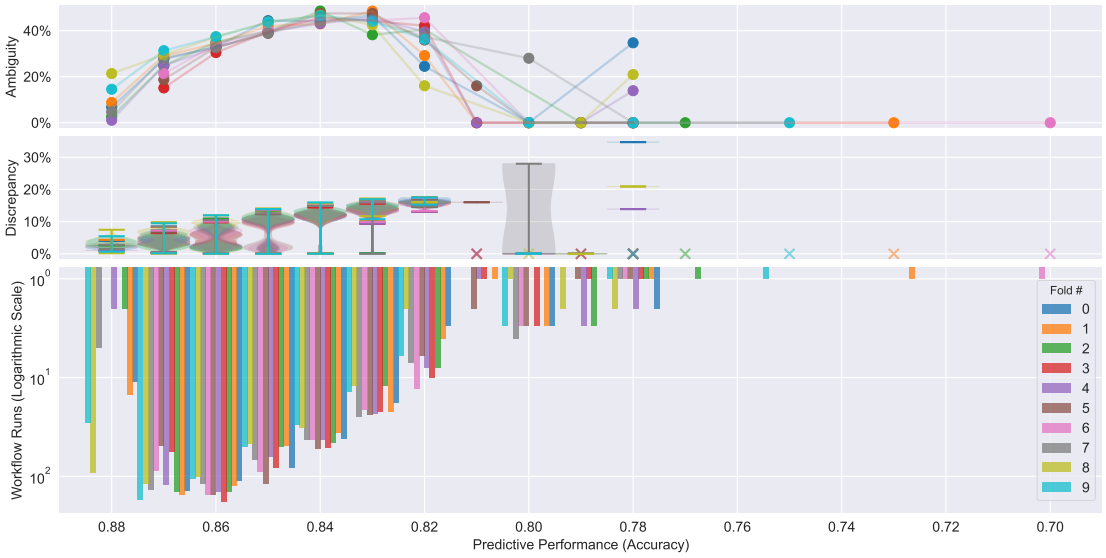
The remaining instrument in our toolkit is a visual depiction of the multiplicity metrics [17] – see Figure 8 – which we generate for our data sets using each of their 10 (colour-coded) cross-validation folds as the fairness test set (refer to the next paragraph for more details). *Ambiguity* quantifies the (percentage) proportion of instances treated unfairly in each performance band discovered for a collection of predictive workflow runs, i.e., models from a single family. *Discrepancy*, on the other hand, measures the (percentage) proportion of instances for which predictions change between any two models from within a given performance band. We present the former as scatter plot-based trajectories, with flat shapes hovering near 0% being optimal for individual fairness under model multiplicity – they indicate consistent classification of individuals. We depict the latter as violin plots, showing the overall distribution as well as the minimum and maximum number of instances treated unfairly when switching between two arbitrary models within a fixed performance level – short violins capture more stable, hence fair, treatment of individuals. Both figures are accompanied



(a) Credit Approval.



(b) German Credit.



(c) Adult with performance bands relaxed – $\epsilon \simeq 10^{-2}$ – to improve readability of the plot.

Fig. 8. Ambiguity (top), discrepancy (middle) and count of predictive models, i.e., workflow runs, (bottom, logarithmic scale) across performance bands for each data set over its 10 cross-validation folds used as fairness test sets. Ambiguity calculates the (percentage) proportion of instances subject to unfairness in view of utility-based model multiplicity; discrepancy measures the (percentage) proportion of instances for which predictions change between any two performance-equivalent models. The (a) Credit Approval and (b) German Credit data sets are displayed *without* relaxing the performance bands. The (c) Adult data set is shown with the performance bands relaxed by rounding them to the 2$^{\text{nd}}$ decimal place, i.e., $\epsilon \simeq 10^{-2}$, to improve readability of the plot (see Figure 9 for ambiguity without rounding and with $\epsilon \simeq 10^{-3}$). The discrepancy for each performance band is based on up to 500 (randomly selected) workflow runs given the overwhelming number of their pairs for the full collection of models. The $\times$ symbol in the discrepancy plots indicates a single run of a predictive workflow, and the – marker is a flat violin plot, i.e., all workflows offering identical predictions.

Data Set	Data ID	Task ID	Flow ID	Execution Count
Credit Approval	29	29	12736	442
German Credit	31	31	6794	102,306
Adult	1590	7592	6970	411

Table 2. OpenML workflow execution counts, i.e., the number of distinct models, as well as identification numbers (IDs) of data sets, tasks and flows used in our experiments.

by a bar plot illustrating (on a logarithmic scale) the number of distinct models, i.e., workflow runs, across all utility bands to help better assess the trustworthiness and reliability of these results.

Equipped with the right tools, we are in a position to assess individual fairness under utility-based model multiplicity for real-life data sets popular with the research community; in particular, we look into *Credit Approval*, *German Credit* and *Adult*. In lieu of designing and testing our own classifiers, we turn to OpenML – a reproducibility repository for machine learning experiments [24]. Such an approach ensures that the predictive workflows used for our study are realistic, especially that we opt for one of the top-performing models for each data set: histogram-based gradient boosting classification tree, random classification forest and decision tree-based AdaBoost respectively. In each case, the ML task is (supervised) crisp binary classification run on 10-fold cross validation; unless otherwise stated, we select the model trained on folds 2–10 whose performance is evaluated on fold 1, which also serves as the fairness validation set. Figures 6 and 7 show the *stability* and *fairness profiles* for the aforementioned data sets; Figure 8 depicts the *ambiguity* and *discrepancy* metrics of model multiplicity. As explained earlier, the former two display results for fold 1 only, whereas the latter spans all of the folds. A summary of this setup given as identification numbers (IDs) of OpenML data sets, tasks and flows is provided in Table 2.

Prior to reviewing individual fairness under model multiplicity, we inspect the influence of using non-zero tolerance δ (Definition 5), i.e., relaxing the performance bands, on the quantity of unique utility levels (computed with accuracy). We investigate this phenomenon by rounding predictive performance ϵ to the 3$^{\text{rd}}$ ($\epsilon \simeq 10^{-3}$) and 2$^{\text{nd}}$ ($\epsilon \simeq 10^{-2}$) decimal place, with the results reported in Table 3. While applying these tolerance values has no effect on Credit Approval and German Credit – given that their performance validation sets have only up to 100 instances – it affects Adult with its 4,885 data points. This observation suggests that for large validation sets we may need to relax the performance bands to get digestible and meaningful results, however such an approach makes it easier to find alternative models and challenge individual fairness. In general, this procedure is likely to degrade individual fairness as it combines predictions

Data Set	Test Points	Performance Bands		
		ϵ	$\epsilon \simeq 10^{-3}$	$\epsilon \simeq 10^{-2}$
Credit Approval	69	12	12	12
German Credit	100	26	26	26
Adult	4,885	185	57	9

Table 3. Count of test instances and number of performance bands with different relaxation criteria for each selected data set. Specifically, we study the quantity of unique utility levels when predictive performance ϵ is rounded to the 3$^{\text{rd}}$ ($\epsilon \simeq 10^{-3}$) and 2$^{\text{nd}}$ ($\epsilon \simeq 10^{-2}$) decimal place as well as without rounding (ϵ). The number of performance bands is likely to decrease when the utility measurements are relaxed, especially so for large validation sets.

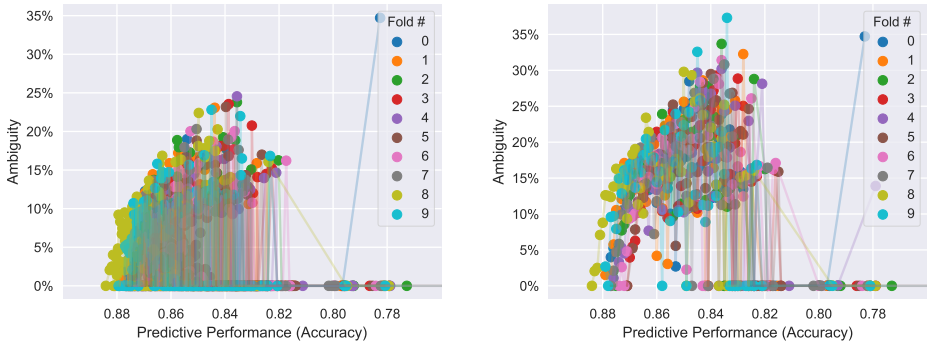
from across strict, i.e., non-relaxed, performance bands, creating a more diverse sample that possibly disagrees on a larger subset of data points.

A stability profile for each selected data set is shown in Figure 6. Given that our analysis produces a large number of (accuracy-based) performance bands, the plots show only a collection of top-performing models. While for the Credit Approval and German Credit data sets (Figures 6a and 6b) the high-utility classifiers appear fair – the stack corresponding to the most accurate models (left-most, depicted in blue) has only one segment – this ceases to hold for classifiers with slightly lower accuracy. In contrast, the predictive workflow used for the Adult data set (Figure 6c) generates a large collection of models whose accuracy differs beyond the second decimal point – multiple stacks composed of at most 3 segments of width 1 – highlighting potential fairness issues. One contributing factor may be significantly larger predictive performance and fairness validation sets: 4,885 instances as compared to 69 and 100 for the other two; however, as we proceed with our investigation we see that this size discrepancy is not entirely to blame.

To paint a more accurate picture we generate the summary variant of fairness profiles – shown in Figure 7 – for the selected data sets. Recall that consistent shading within each column, i.e., for individual data points, both within a single and across the colour-coded performance bands, conveys a stable, thus fair, prediction. A vertical bar spanning a single performance band that changes its saturation in the middle signifies the maximum disagreement between the models of this utility. While the profiles for the Credit Approval and German Credit data sets (Figures 7a and 7b) appear relatively consistent, the one for Adult (Figures 7c) is more jittery (note that for legibility reasons the plot only shows a subset of 250 instances out of 586 individuals who are treated unfairly, with the entire data set composed of 4,885 points).

To further investigate individual fairness under model multiplicity we analyse ambiguity and discrepancy of our predictive workflows using every fold of the data as the fairness validation set; the results are shown in Figure 8 and corroborate the insights gathered thus far. Note that for the Credit Approval and German Credit data sets (Figures 8a and 8b) we work with strict model multiplicity $\mathcal{F}_\epsilon$ (Definition 1), whereas Adult (Figure 8c) is investigated with relaxed performance bands $\mathcal{F}_{\epsilon\simeq\delta}$ (Definition 5), which in this case is achieved by rounding utility to the second decimal point ($\epsilon \simeq 10^{-2}$). We apply this procedure since the stability (Figure 6c) and fairness (Figure 7c) profiles for this particular data set suggest that ambiguity and discrepancy may otherwise be uninformative. Based on the former we can see an overwhelming number of unique prediction vectors – visible as stacked segments of width 1 – in each colour-coded performance band, signifying unfair treatment of many instances; the latter plot paints a similar picture. We validate these observations by calculating the number of performance bands as well as computing ambiguity without and with rounding to the third ($\epsilon \simeq 10^{-3}$) and second ($\epsilon \simeq 10^{-2}$) decimal points – refer to Table 3 and compare Figures 8c and 9.

For all the data sets, ambiguity raises quite sharply as we move away from an optimal model, i.e., the top performance band, reaching between 40 and 90% at its peak, which draws attention to a surprisingly high degree of individual unfairness for these classifiers. Discrepancy, on the other hand, shows that while in many cases it is possible to switch between two (performance-wise) equivalent models without affecting individual predictions, a non-negligible collection of instances – captured by tall violin plots – is likely to be treated differently. More generally, the predictive workflows used with Credit Approval and German Credit data seem to behave consistently without the need of relaxing performance bands, whereas the one used with Adult only provides discernible patterns when utility is rounded to the third ($\epsilon \simeq 10^{-3}$) or second ($\epsilon \simeq 10^{-2}$) decimal point. We suspect that in the latter case the culprits are volatility and high expressiveness of the underlying predictive model – decision tree-based AdaBoost – however this intuition cannot be confirmed without further analysis. Importantly, relaxing performance bands merges distinct levels of utility, thereby introducing more disagreements on the level of individual predictions. This phenomenon can be observed in



(a) Adult without rounding performance bands.

(b) Adult with performance bands rounded to the 3^{rd} decimal point, i.e., $\epsilon \simeq 10^{-3}$.

Fig. 9. Ambiguity range for the Adult data set increases from (a) 0–25% to (b) 0–35% when predictive performance bands are rounded to the 3^{rd} decimal place. The plots are truncated on the right side; see Figure 8c for a reference of the full scale and comparison to rounding performance bands to the 2^{nd} decimal place, i.e., $\epsilon \simeq 10^{-2}$, which yields ambiguity range of 0–50%.

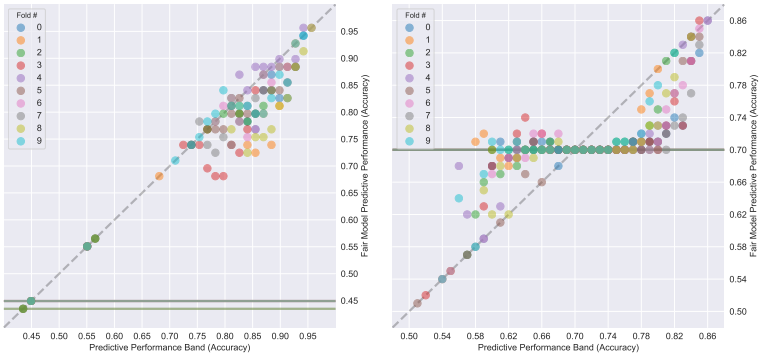
the ambiguity range being stretched from 0–25% without rounding predictive performance to 0–35% and 0–50% when utility is rounded to the third and second decimal places respectively (see Figures 8c and 9 for reference).

Finally, we analyse the utility – measured with accuracy – of individually fair models $f^\star$ (Definition 4) constructed for every performance band of each selected data set across all the folds. In addition to strict model multiplicity $\mathcal{F}_\epsilon$ (Definition 1), we investigate relaxed performance bands $\mathcal{F}_{\epsilon\simeq\delta}$ (Definition 5) exclusively for Adult, rounding accuracy to the third ($\epsilon \simeq 10^{-3}$) and second ($\epsilon \simeq 10^{-2}$) decimal point. The results – presented in Figure 10 – show a nearly universal drop in predictive performance when dealing with $f^\star$. An interesting exception is German Credit – Figure 10b – for which the proportion of the favourable class is 70% in each fold (based on the ground truth labels). This class imbalance improves the accuracy of individually fair models for performance bands (x-axis) below 70%, whereas after this mark the utility drops as expected. All in all, our comprehensive inspection toolkit appears to serve its purpose well.

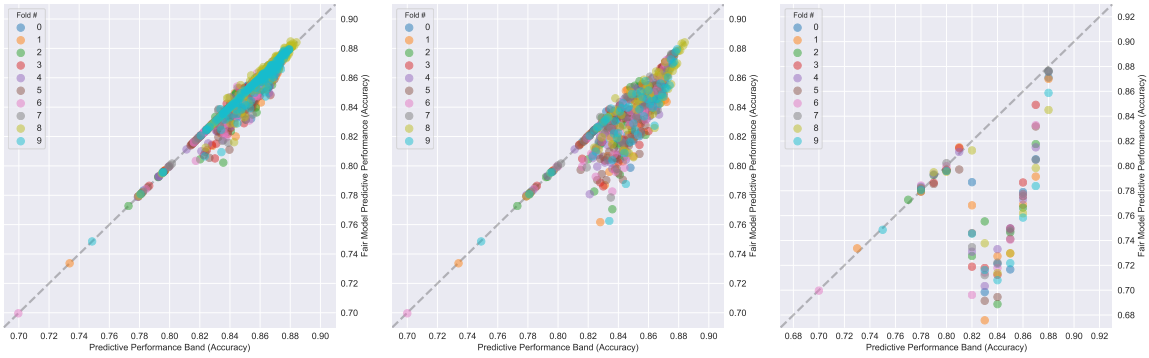
5 RELATED WORK

Fairness of artificial intelligence and machine learning algorithms has attracted considerable attention in recent years following proliferation of data-driven automated decision-making systems across real-life applications [20]. Two main themes dominate this research field: individual and group-based fairness, both focusing on strategies to identify and mitigate disparate impact that manifests itself through discriminatory behaviour and diverse biases [3]. Nonetheless, complementary viewpoints also emerge, investigating this topic under substantially different assumptions such as social and population changes over time, which bring to light considerations of the delayed impact of enforcing fair decisions [16]. Regardless, a nearly universal assumption found in the literature is to presuppose a *fixed predictive model*, thus overlooking its provenance, evolution as well as any implicit choices related to it. This is at odds with the model *multiplicity* phenomenon [6] – i.e., possible existence of a *collection* of equally capable classifiers – which has been largely neglected by the fairness community [11, 17].

Individual fairness deals with disparate treatment of a single person based on selected comparison criteria such as relevant *protected characteristics*. This can be captured by a dedicated *similarity metric* – assuming an intuitive



(a) Credit Approval. (b) German Credit.



(c) Adult. (d) Adult with $\epsilon \simeq 10^{-3}$. (e) Adult with $\epsilon \simeq 10^{-2}$.

Fig. 10. Predictive performance – accuracy – of fair models $f^\star$ (y-axis) compared to (a, b & c) strict ϵ as well as (d & e) relaxed $\epsilon \simeq \delta$ performance bands (x-axis) discovered for individual models $f \in \mathcal{F}_\epsilon$ and $f \in \mathcal{F}_{\epsilon \simeq \delta}$ respectively. For consistency, only Adult is investigated under relaxed utility. The diagonal line represents unchanged performance, with points above indicating an improvement and points below a decrease in the accuracy of individually fair models built upon their respective collections of fixed-utility classifiers. The horizontal lines, plotted separately for each fold, indicate the performance achieved by alway predicting the preferred class; most of them overlap in Panels (a) and (b), and they are not visible (24%) in Panels (c), (d) and (e). Overall, we see up to 15% drop in accuracy when using a fair model, however this pattern is broken for data sets with highly imbalanced classes as shown by Panel (b).

notion that similar people should be treated comparably – however defining such a measurement strategy is non-trivial [10]. An alternative approach to individual fairness is formulated via *counterfactuals*, whereby changing a (protected) attribute should not yield a different prediction. This framing is also very intuitive and relatively easy to test, however the concept of a "counterfactual individual" may be ill-defined. For example, altering ethnicity while preserving the remaining features intact may not be representative given that all these other personal traits and attributes are (implicitly) influenced by the single personal characteristic being manipulated. Notably, causality may be employed to partially alleviate such shortcomings [15]. *Group-based fairness*, on the other hand, deals with disparate treatment of sub-populations determined, for example, by partitions drawn across protected attributes, in which case parity may be achieved by tweaking the underlying model separately for each group, e.g., via their respective classification thresholds [13]. Importantly, many such notions of fairness are inherently incompatible [14], and enforcing some of them may require trading off a degree of predictive performance [12].

Multiplicity of data-driven predictive models is a well-observed phenomenon – sometimes called the Rashōmon effect of statistics – where a group of classifiers exhibits comparable utility despite intrinsic differences [6]. This collection of models may rely on different subsets of attributes or patterns found in the training data, with some arguing that exogenous information may be necessary to narrow down its scope [11, 18]. In relation to fairness, model multiplicity has been used to understand the dependence of (inaccessible) classifiers on selected (protected) attributes, which can help to robustly identify discriminatory practices [11]. More recently, model multiplicity was suggested as an additional metric for evaluating accountability of classifiers. Specifically, Marx et al. [17] showed how this phenomenon may be problematic (from ethical and fairness perspectives) when individuals receive conflicting predictions, and offered to address this issue by granting them the most favourable decision. While this line of reasoning is close to ours, their study is limited to linear models based on Mixed-Integer Programming, whereas the work presented here is rooted directly in fairness, providing a more fundamental and general treatment of model multiplicity that is supported by a comprehensive analysis of real-life predictive pipelines. Similarly, Pawelczyk et al. [21] explored the implications of model multiplicity on veracity of counterfactual explanations, which may be easily lost when switching between different models whose performance is comparable.

Fundamentally, individual fairness under model multiplicity may be abated (to an extent) by employing a predictive paradigm more complex than crisp (binary) classification. Section 1 briefly discusses strategies that can be used to increase the chances of a model being unique, e.g., by enforcing various heuristics and criteria with respect to its properties other than predictive performance. Adopting *classification with reject or abstain option* is another approach to mitigate predictions that are unfair from a perspective of an individual [7, 8]. Dedicated techniques such as *prediction intervals* or *conformal predictions* can also be deployed. The former marks any instance falling within a specified area constructed around a decision boundary as unreliable, allowing it to be assigned the most favourable decision; the latter follows a similar strategy, instead outputting a set of admissible classes, which can be further processed as desired. A possible alternative is to use *ensemble learning*, e.g., a random forest [5], where instead of a majority vote the prediction is derived through maximising the output of the contributing models. Machine learning algorithms with prediction margins, such as support vector machines [9], as well as probabilistic classifiers that offer confidence scores or trustworthiness measures are another option. While in most of such cases the acute signs of individual unfairness under model multiplicity can be eliminated, the underlying problem may not necessarily be fully resolved, causing troubles down the line.

6 CONCLUSIONS AND FUTURE RESEARCH

In this paper we formalised the notion of individual fairness under utility-based model multiplicity: a scenario in which a number of classifiers that are considered equivalent based on their predictive power assign different labels to certain data points. Such a situation may arise for distinct types of models, diverse parameterisation of the same model, or over its multiple training runs when the underlying process is, for example, greedy or stochastic. Additionally, high-dimensional and sparse data may contribute to this phenomenon given the curse of dimensionality (everything is far away from each other). In particular, we built these concepts around a user-specified *family of predictive models*, taking into consideration its *expressiveness*, which in the extreme may lead to awarding the most favourable prediction to every individual. We then defined *two meaningful cases of model multiplicity*: one determined by the entire data space and another measured with respect to a designated *fairness validation data set*. We generalised the former into *disputable spaces* – regions where at least two models disagree on the prediction – which may not necessarily be identified even with a comprehensive fairness validation set.

Next, we showed how to combine (performance-wise) equivalent models to present each individual with the best possible decision; we also demonstrated that such an approach may adversely affect the overall predictive power of the classification task at hand, especially for highly expressive models. As an alternative, we discussed imposing restrictions on the model space to limit the number of viable alternatives, arguing for enforcing (operational) constraints that are meaningful to each individual modelling problem, e.g., prediction monotonicity with respect to a chosen attribute, which is also recognised as a strategy for introducing (ante-hoc) explainability whereby predictions are aligned with human values and can be easily justified. Otherwise, being able to determine disputable spaces – "grey areas" or "edge cases" from a classification standpoint – can allow to abstain from making a decision or engage a human expert in the process, thus partially mitigating the prevalence of unfair predictions. To help identify individual unfairness stemming from model multiplicity we introduced a bespoke visualisation toolkit, which we explained on and applied to real-life data sets popular in fairness research, using predictive pipelines available in the OpenML repository. Our findings highlight the importance of considering utility-based model multiplicity as a new dimension of algorithmic fairness.

In future work, we will investigate numerical metrics and analytical tools to assess multiplicity-based unfairness in a broader context, extending the notion beyond crisp binary classification. For example, we will study some of the techniques discussed briefly in Section 5. Moreover, we will look into developing methods to derive bounds on the most and least favourable treatment of each individual from within a given data set for a selection of predictive models. This should facilitate systematic identification of the disputable and stable regions, possibly leading to a new optimisation objective – minimise the former or maximise the latter – for training individually fair, robust and effective predictive models. Finally, we will explore the lower limit of utility for a fair classifier that always uses the most beneficial model from their collection determined by a given level of predictive performance.

ACKNOWLEDGMENTS

This research was supported by the ARC Centre of Excellence for Automated Decision-Making and Society, funded by the Australian Government through the Australian Research Council (project number CE200100005); the Estonian Research Council (projects number PUT1458 and PRG1604); and the Dora Plus programme, sponsored by the European Regional Development Fund and Estonian government.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica.
[2] UN General Assembly. 1948. Universal declaration of human rights. *UN General Assembly* 302, 2 (1948), 14–25.
[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. https://fairmlbook.org.
[4] William Blackstone. 1830. *Commentaries on the Laws of England*. Vol. 2. Collins & Hannay.
[5] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[6] Leo Breiman. 2001. Statistical modeling: The two cultures. *Statist. Sci.* 16, 3 (2001), 199–231.
[7] Chi Keung Chow. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers* 4 (1957), 247–254.
[8] Chi Keung Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory* 16, 1 (1970), 41–46.
[9] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[11] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* 20, 177 (2019), 1–81.
[12] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

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

[14] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[15] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[16] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *Proceedings of the 35th International Conference on Machine Learning* (2018).

[17] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*. PMLR, 6765–6774.

[18] James W McAllister. 2007. Model selection and the multiplicity of patterns in empirical data. *Philosophy of Science* 74, 5 (2007), 884–894.

[19] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[20] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[21] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 809–818.

[22] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[23] Kacper Sokol and Peter Flach. 2021. Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence. (2021). arXiv:2112.14466

[24] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations* 15, 2 (2013), 49–60. https://doi.org/10.1145/2641190.2641198

[25] V. N. Vapnik and A. Ya. Chervonenkis. 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications* 16, 2 (1971), 264–280. https://doi.org/10.1137/1116025