AI and Ethics
https://doi.org/10.1007/s43681-022-00142-y

ORIGINAL RESEARCH

Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law

Daniel Vale[1] · Ali El-Sharif[2] · Muhammed Ali[3]

Received: 22 September 2021 / Accepted: 30 January 2022

Abstract

Organizations are increasingly employing complex black-box machine learning models in high-stakes decision-making. A popular approach to addressing the problem of opacity of black-box machine learning models is the use of post-hoc explainability methods. These methods approximate the logic of underlying machine learning models with the aim of explaining their internal workings, so that human examiners can understand them. In turn, it has been alluded that the insights from post-hoc explainability methods can be used to help regulate black-box machine learning. This article examines the validity of these claims. By examining whether the insights derived from post-hoc explainability methods in post-model deployment can *prima facie* meet legal definitions in European (read European Union) non-discrimination law, we argue that machine learning post-hoc explanation methods cannot guarantee the insights they generate.

Ultimately, we argue that the use of post-hoc explanatory methods is useful in many cases, but that these methods have limitations that prohibit reliance as the sole mechanism to guarantee fairness of model outcomes in high-stakes decision-making. By way of an ancillary function, the inadequacy of European Non-Discrimination Law for algorithmic decision-making is demonstrated too.

Keywords Artificial intelligence · Explainability · Discrimination · Law · Non-discrimination law · Machine learning

1 Introduction

The predictive accuracy of machine learning is increasingly publicly acknowledged and embraced in high-stakes decisions within both the public and private sectors, such as in the domains of criminal justice, medicine, and banking [4, 28, 49, 50, 62]. In an attempt to exploit this predictive accuracy, complex machine learning algorithms, often known as black-box models, are employed. The perception is that greater machine learning model complexity directly correlates with higher predictive accuracy and, in turn, better results [62]. For example, increasing the number of layers in a neural network or increasing the number of neurons in each layer can improve the performance of models [3]. Complex black-box machine learning models are difficult to maintain, debug, and test for robustness. However, the principal concern accompanying the implementation of black-box machine learning models is their lack of transparency [27, 53].

The opacity of black-box models' can arise from the distinct difficulty of interpreting prediction results in which accuracy is achieved through model complexity [13, 46, 55]. This, in turn, limits the ability for model oversight in post-deployment. Given the real-world application of black-box machine learning models in high-stakes decision-making, this is concerning [62]. It is well documented that, for high-stakes decision-making in the criminal justice, medicine, and banking domains, black-box machine learning models have generated unjustified social ills [4, 50, 75, 83]. Unfavourable parole denial and credit decisions based on race were identified by Larson et al. [39]. This is certainly not desirable, acceptable, nor legally permissible. As a result, there is growing recognition about the insufficiency of predictive accuracy offered by black-box machine learning models for high-stakes decision-making [62] and the need

✉ Daniel Vale
d.s.vale@law.leidenuniv.nl

[1] eLaw Centre, Leiden School of Law, Leiden University, Leiden, The Netherlands

[2] College of Computing and Engineering, Nova Southeastern University, Fort Lauderdale, USA

[3] UCL Knowledge Lab, University College London, London, UK

Published online: 15 March 2022

to incorporate adherence to legal requirements and social norms in model evaluation criteria [63]. To address concerns surrounding discrimination and generated social ills, post-hoc explainability methods are often alluded to as transparency tools that can be used to help guard against these [29, 54, 65, 82].

Explainability methods are advocated in the legal community as a pragmatic tool to help promote machine learning model transparency and, in turn, unearth model discrimination and other social ills [29, 54, 65, 83]. However, these suggestions seem insufficient as they do not include references to broader machine learning fairness metrics, such as outcome parity tests [29].

Post-hoc explainability methods approximate complex black-box machine learning models by generating simpler surrogate models [68]. In turn, these simpler surrogate models can be used by human examiners to comprehend and appreciate the inner workings of black-box machine learning models and, thereby, the real-world application of models' post-deployment [56, 57].

Post-hoc explainability methods are broad in scope. They include methods supporting models based on different data types including tabular data [32], computer vision [45], and natural language text [52]. The output of post-hoc explanatory methods varies amongst text, visual explanations, and feature relevance/feature importance [6]. Given the broad scope of the field, an exhaustive review of post-hoc methods is beyond the scope of this paper. We make no claim that the limitations presented in this manuscript apply to all methods and all contexts. The goal is to identify potential risks and limitations that are present in some popular methods, and in doing so, help legal professionals be educated consumers of post-hoc explanatory methods. This is an important caveat to note when engaging with this paper.

We raise questions about whether the insights generated by some post-hoc explainability methods in post-model deployment can be used to pragmatically regulate black-box machine learning models. This paper specifically examines the questions within the context of non-discrimination law. It does so by theoretically interrogating whether the insights generated from post-hoc explainability methods in post-model deployment can *prima facie* meet the legal definitions of direct or indirect discrimination in European (read European Union) non-discrimination law. We proposed that if the insights generated from post-hoc explainability techniques can *prima facie* meet these legal definitions, they can, in turn, be used to regulate black-box machine learning models post-deployment without fear. However, if the insights generated from post-hoc explainability techniques cannot *prima facie* meet these legal definitions, their use to regulate or "govern" black-box machine learning models' post-deployment ought to be met with skepticism.

This paper ultimately argues that post-hoc explanation methods cannot guarantee that the insights they generate demonstrate the absence of discrimination (the null hypothesis). Therefore, their insights cannot *prima facie* meet, with sufficient consistency and certainty, the definitions of direct or indirect discrimination in European Law. By way of an ancillary function, the inadequacy of the European non-discrimination law for machine learning, especially in the context of high-stakes decision-making, is demonstrated. This paper's target primary audience is legal scholars, practitioners, regulators, lawyers, and jurists. Its secondary audience is machine learning researchers, ethicists, and engineers.

This paper is structured as follows: first, a brief background on non-discrimination law and the legal definitions of direct and indirect discrimination in European Law are presented. Proposed alternative "statistical fairness tests" in this regard are also introduced. These definitions are then mapped against acknowledged types of model outcome discrimination. Thereafter, an explanation of machine learning, model outcome bias (discrimination), post-hoc explanatory methods, and their limitations are submitted. Finally, a theoretical analysis of the technical limitations of post-hoc explanatory methods against mapped legal definitions and "statistical fairness tests" is tendered.

1.1 Non-discrimination law: legal definitions

The subject of fairness and non-discrimination within machine learning has generated much literature, including varying definitions [22, 26, 36, 48, 71, 80]. However, given the scope of this article, understandings of fairness and non-discrimination will be limited to a European legal perspective.

At the international level, non-discrimination law can be found in a plethora of human rights treaties, such as the International Covenant on Civil and Political Rights, the International Covenant on Economic Social and Cultural Rights, and the International Covenant on the Elimination of All Forms of Racial Discrimination [24].

Within the European Union ("EU"), the European Convention on Human Rights ("ECHR") is the legal bastion against discrimination [38, 83]. This results from a lack of a human rights framework within original EU treaties, later remedied by the Amsterdam and Lisbon Treaties when read in conjunction with the EU Charter of Fundamental Rights [19, 33, 38]). Article 14 of the ECHR reads: "*The enjoyment of the rights and freedoms set forth in the European Convention on Human Rights and the Human Rights Act shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.*" EU Directives 2000/43/

EC, 2000/78/EC, 2004/113/EC, and 2006/54/EC further confirm this prohibition in context-specific areas, such as labour law and the access and supply of goods and services [24].

It is evident within ECHR jurisprudence that both *direct* discrimination (read "disparate treatment" in North America labour law) and *indirect* discrimination (read "disparate impact" in North America labour law) are prohibited [83]. Although a heterogeneity of interpretations and applications of EU non-discrimination law exist [43, 74], conceptually direct discrimination can be distilled as the use of an overtly prohibited ground of discrimination to unlawfully differentiate between persons or classes thereof in consequence in specified protected sectors [20, 24, 42, 43, 75]. Indirect discrimination entails the use of seemingly neutral criterion to do the same [20, 24, 43, 75, 83]. Intent is largely superfluous to these considerations [69]. A further distinction between direct and indirect discrimination is that the former appeals to formal equality (read "equality" in North America), while the latter advocates substantive equality (read "equity" in North America) [24, 69, 75]. The above immediate difference and its varying manifestations and duties are beyond the bounds of this paper.

In an attempt to elaborate on these definitions, direct discrimination requires, more specifically, that (a) a protected class, (b) when compared to a non-protected class, (c) receives less favourable treatment (d) based on the application of a criterion that (e) directly appealed to a prohibited ground of discrimination [15]. Somewhat distinctly, *indirect* discrimination entails that (1) a protected class, (2) when compared to a non-protected class, (3) receives less favourable treatment (4) based on the application of a seemingly (5) neutral criterion that does not directly appeal to a prohibited ground of discrimination, but indirectly so [15]. In essence, what needs to be demonstrated is that (i) a claimant suffered harm (comparative unfavourable treatment), and (ii) that such harm stemmed from the application of a criterion that either directly or indirectly appeals to a prohibited ground of discrimination, i.e., that a sufficient legal *nexus* linking the two exists [58].

Notably, under EU Non-Discrimination Law, discriminatory practices—whether indirect or direct—will be lawful if a legitimate aim objectively justifies them and the means of achieving that aim are appropriate and necessary [15, 82, 83]. These conditions lie beyond the immediate purview of this paper.

Due to the normativity and politicized nature of the interpretation and application of EU non-discrimination law, it is criticized for being "*too contextual, reliant on intuition, and open to judicial interpretatio*n" [74]. The results are heterogeneous, fluid legal standards for non-discrimination that make the Law purportedly ill-suited for judging algorithmic and machine learning discrimination [74]. Consequently, Wachter et al. [74] propose two "statistical fairness tests" that can be used to unearth *prima facie* discrimination in algorithmic decision-making. The first test, Demographic Disparity, asserts that if protected and non-protected classes are not equally represented—in accordance with population demographics—in favourable or unfavourable model outcomes, discrimination will be *prima facie* present [74]. The second test, Negative Dominance, is two-part and necessitates (a) the majority of participants in unfavourable model outcomes are from protected classes, and (b) a minority of participants in favourable model outcomes are from protected classes [74]. Although the merit worthiness of these standards are not fully established, they serve as useful alternative formulations of discrimination.

Due to their consequential and representativeness focus, it is proposed here that both the Demographic Disparity and Negative Dominance tests can apply to instances of direct and indirect discrimination. The implemented group classifier for distinguishing between protected and/or non-protected classes is not discussed by Wachter et al. [74] when articulating both the Demographic Disparity and Negative Dominance tests. It is, therefore, assumed to hold no immediate value in such analyses, allowing both tests to accommodate the application of a classifier appealing to (a) an overtly prohibited ground of discrimination (as envisaged in direct discrimination), or (b) a seemingly neutral one, which serves as a proxy for a prohibited ground of discrimination (as practiced in indirect discrimination).

1.2 Mapping model outcome bias and non-discrimination law

A machine learning model is said to be discriminatory (hereafter "biased") in outcome if (1) group membership is not independent of the likelihood of a favorable model outcome ("Type (1) Model Bias"), or (2) under certain circumstances, membership in a subset of a group is not independent of the likelihood of a favourable model outcome ("Type (2) Model Bias") [34]. For example, Type (1) Model Bias will be actualised where a model utilizes gender as a marker to differentiate between group membership, such as male and female. Should membership to a specified gender, such as male, not lead to an independent favourable outcome, Type (1) Model Bias will be present. Type (2) Model Bias is in effect wherein, within the same example, transgender men, a subset within the group membership of male, do not have independent favourable outcomes. Note the tacit presumption of independence of the likelihood of favourable model outcomes for comparative groups and/or group subsets in both bias types. Should a machine learning model lack independence of the likelihood of favourable model outcomes for comparative groups and/or group subsets, it would generally be classified as inaccurate for all groups, as opposed to biased. Given Type (1) and Type (2) Bias above, the demands of non-discrimination law ought to be mapped accordingly.

In terms of Type (1) Model Bias, it seems that both instances of direct and indirect discrimination can arise. The important cursors for Type (1) Model Bias are (a) delineation of group membership and, consequently, (b) the non-independence of the likelihood of a favorable outcome. For direct discrimination, the group membership classifier would overtly utilize a prohibited ground of discrimination (hereafter "Prohibit Ground"). Subsequently, any lack of independence of the likelihood of a favorable model outcome that directly correlates along group membership would be demonstrative of comparative disadvantage. In turn, an instance of direct discrimination is apparent. Note again the presumption of independence of the likelihood of favorable outcomes for comparative groups. Instances of indirect discrimination are somewhat analogous, except that the classifier for group membership would not overtly appeal to a Prohibited Ground, but indirectly so. A seemingly neutral classifier would act as a proxy for determining group membership along with a Prohibited Ground and, subsequently, if non-independence would follow, indirect discrimination would be present. Again, note the aforementioned presumption, which is tacitly demonstrative of comparative disadvantage.

Type (2) Model Bias can meet instances of indirect discrimination seamlessly, but not direct discrimination. For Type (2) Model Bias, again, the delineation of group membership and consequent non-independence of the likelihood of favorable outcomes are important requirements. In terms of indirect discrimination, rather effortlessly, all that need to be demonstrated is that group subsets, which correlate along Prohibited Grounds either directly or indirectly do not have independent favorable outcomes. This is because, by definition, the classifier for group membership cannot appeal to a Prohibited Ground. If it did, it would amount to direct discrimination under Type (1) Model Bias. Clearly, a neutral classifier must be in used within this context. Absent of a classifier overtly appealing to a Prohibited Ground, direct discrimination categorically cannot be found. The result is that Type (1) Model Bias when found can amount to either direct or indirect discrimination. Type (2) Model Bias, however, can only be demonstrative of indirect discrimination. Again, note the presumption of independence of the likelihood of favorable model outcomes for comparative group subsets, which would be demonstrative of disparity.

The Demographic Disparity and Negative Dominance tests rightly—in terms of EU non-discrimination law—challenge the presumption of independence of the likelihood of favorable model outcomes for comparative groups or group subsets. Instead, they require *prima facie* proof of such—or, rather, greater disparity. This means that the non-independence of the likelihood of a favorable model outcome (Type (1) and/or Type (2) Bias) alone cannot demonstrate discrimination in terms of EU non-discrimination law. Rather, it must be accompanied by clear outcome disparity—either through (a) non-representativeness in favourable and unfavourable model outcomes for (intra) model groups (Demographic Disparity); and/or (b) the over- and under-representative of non-protected and protected classes in favourable and unfavourable model outcomes (Negative Dominance).

1.3 Post-hoc explainability methods and black-box models: key concepts

Black-box models refer to automated decision systems that map user features into a decision class without exposing how and why they arrive at a particular decision [46, 55]. The internals of black-box models are either unknown or not clearly understood by humans [13, 32]. The terms black-box, grey-box, and white-box refer to the level of exposure of the internal logic to the system user—i.e., human examiners [1].

Although interpretability can be intentionally obstructed to protect (trade) secrets, to protect against gaming the system, and maintain a competitive advantage [11], black-box models' opacity can arise from the distinct difficulty of interpreting their classification results, which leverage large datasets and achieve accuracy through model complexity. This means that deriving interpretability is extremely challenging. Despite this, post-hoc explainability methods are often touted as a solution.

The primary characteristics of successful predictions are accuracy and interpretability [32]. Predictive accuracy establishes "what" is the correct label on unseen data, while interpretability answers "how" and "why" a prediction was made and what features influenced the prediction [5, 39]. Interpretability is, therefore, essential in establishing human oversight for models and, more directly, warranting against discrimination. A further distinction between interpretability and explainability has been made in the literature. Explainability refers to understanding the outcome of the model (why was a decision made), while interpretability describing the internal logic of the model (how a decision was made) [7, 10]. In this paper, we loosely use the term interpretability to represent the ability of a human examiner to understand what the model did [30].

A purported trade-off between accuracy and interpretability has been established in the literature [2, 8]. Although interpretable models provide meaningful insight into the decision-making process, it is claimed that they lack the expressive power to capture the underlying relationship between input features and the output. Models that accommodate more complex functional relationships, such as black-box models, are claimed to have greater predictive power, but are often difficult to interpret [9, 13, 14].

1.4 Post-hoc explainability methods: insights

Post-hoc explainability takes a trained model as input and extracts the underlying relationships that the model had learned by querying the model [47] and constructing a white-box surrogate model [10]. Post-hoc explanations mimic model distillation [52] as they transfer the knowledge from a large, complex model [the black-box model] into a simpler, smaller one (the white-box surrogate model). In doing so, they represent an estimated explanation of what the larger, complex model is doing, but not exactly how or why it arrived at a prediction. They, therefore, only generate an approximation of the functioning of the black-box model. Although this approximate explanation is not an exact match, it is often thought to be close enough to be useful in understanding the black-box model's logic. It is important to note that post-hoc explainability methods do not place constraints on their black-box model counterparts [10, 21]). This means that they explain the output of black-box models without negatively impacting the underlying model's predictive accuracy [10, 21]) (Fig. 1).

Post-hoc explainability can traditionally generate two different 'types' of interpretability: global and local. Global interpretability explains the whole logic of a model and the reasoning behind all possible outcomes [32]. Global model interpretability explains a model through the most important rules learned from the training data and represents the explanation through the structure and parameters of a model [21]. Examples of global interpretability rules are the coefficients in a linear regression model or rules encoded by a path from the root node to the leaf nodes in a decision tree model.

Local interpretability explains model characteristics and the impact of input features for a specific prediction [1, 21, 32]. Because small sections of the model are more likely to be linear, local models expressed as a linear function of input features can be more accurate than global models [35]. To put this more plainly, local interpretability examines a particular aspect (or, rather, area) of a model and its working instead of the entirety of the model and its logic.

Post-hoc explainability can be applied in two ways. The first is model-specific interpretability, which refers to explanations exclusive to a particular model. It derives explanations using the given model's internal representation or learning process [1, 21, 61]. Therefore, model-specific interpretability needs to be inherently tailored for and applied to a particular model. Once achieved, it cannot be used to explain any other model. Model-agnostic explanatory methods approximate the behavior of underlying models to generate end-user explanations independent of the internal logic used to generate predictions and are standardized [56]. Model-agnostic explanatory methods are not model-specific explainers. Due to their potential to be applied to more models, they have a wide-scale application.

Post-hoc model-agnostic explanations fall into the broader category of explaining by removing [16]. These methods identify feature importance by applying different techniques of removing an input feature and, by doing so, examining the impact on the prediction. The more significant the impact on the prediction from the absence of a feature, the more that feature is considered important. A very popular local post-hoc explanation method is Local Model-Agnostic Explanations ("LIME"). The figure below illustrates LIME's output explaining an instance from a tabular dataset for a classifier predicting the probability of recidivism using the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset. The explanation provides visualization for the most important feature that contributed to a randomly selected instance prediction generated by LIME's default hyperparameters. The selected important features are assigned a coefficient that can positively or negatively indicate the direction of the relationship between the features and the predicted class. Coefficients values express the magnitude of feature contribution. The coefficient value represents the contribution of features to the underlying model's prediction (Fig. 2).

1.5 Model outcome bias: post-hoc explainability methods

According to Suresh and Guttag [67], model outcome discrimination (hereafter "bias") is manifested, if not intentionally, through five types of model development pipeline biases. They can be summarised as follows:

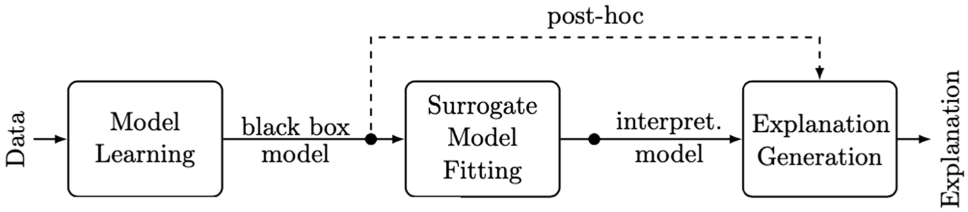

Fig. 1 Post-hoc interpretability. From: "Burkart, N., & Huber, M. F. (2020). A Survey on the Explainability of Supervised Machine Learning. arXiv preprint arXiv:2011.07876"

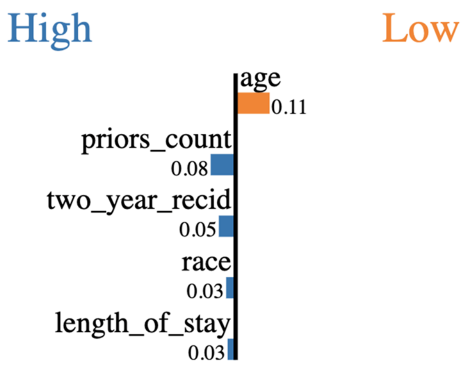

Fig. 2 Post-hoc explanation visualization by LIME

1. **Historical bias** occurs in the data processing stage and is rooted in the real world. It is not dependent on data processing or model building. These are systematic realities in the data and real world [67]. This is also referred to in relevant literature as "social bias" as opposed to "technical bias" [75],
2. **Representation bias** occurs in the data processing stage when the selected sample is not representative of the real population. Data might be skewed either towards or away from a certain group. This is the result of poor data governance or interrogation [18]. This would be reflective of "technical bias" as opposed to "social bias" [75],
3. **Measurement bias** occurs in the data processing stage when choosing or shortlisting the features of interest from the sample population. There can be bias in the measurement of these features of interest. This results from inadequate data exploration and/or interrogation [70]. This would be reflective—again—of "technical bias" as opposed to "social bias" [75],
4. **Evaluation bias** occurs in the model design and/or building stage(s) during a model's evaluation and iteration when the model's parameters are biased. This is the result of poor model design governance and/or decision-making [36, 66]. Again, this is "technical bias" as opposed to "social bias" [75], and
5. **Aggregation bias** arises from wrong assumptions about the population and can persist throughout data processing and model building stages. This results from poor model design, or development governance, or decision-making [23]. This would be reflective of "technical bias" as opposed to "social bias" [75].

The above biases are independent of each other and are not mutually exclusive. Therefore, multiple biases can exist at any one time.

Post-hoc model-agnostic explainability methods that produce feature importance can mediate measurement, evaluation, and aggregation bias and only do so when a human-in-the-loop [17] approach is adopted. As expressed, such methods visualize how important a particular feature is for a specific prediction.

Human-in-the-loop (i.e., human oversight) provides a sanity check to evaluate if the prediction is justified and why the model is making the prediction. Some of these biases are difficult to address by machine learning practitioners alone. Moreover, some of these biases are not taken into account by post-hoc explanations, like historical bias or measurement bias; i.e., post-hoc explanations cannot detect these. Therefore, domain relevant experts are needed. To give a hypothetical, hyperbolic example, if it is a norm in a particular geographic location that 95% of the candidates who become lawyers are male and only 5% female, then no matter how good qualitatively or quantitatively our data is, this bias will persist if models use this dataset to make predictive outcomes of the likelihood of a candidate to be a lawyer. This is an example of a historical bias that exists in society and is not caused by any data collection mistakes. Accordingly, if we train a model to approve candidates based on these data, a female candidate might receive a negative prediction, because they are female and are not showing similar characteristics to most lawyers who—in this hypothetical, hyperbolic example—are predominantly men. In this scenario, post-hoc explanations will fail to take this reason—i.e., historical bias—into account, because sensitive variable bias gender in this example is systemic to the training data instead of being representative of a particular feature per se.

1.6 Post-hoc explainability methods: limitations

Again, note that post-hoc explainability methods are broad in scope. They include methods supporting models based on different data types including tabular data [32], computer vision [45], and natural language text [52]. The output of post-hoc explanatory methods varies and text, visual explanations, and feature relevance/feature importance [6]. Given the broad scope of the field, an exhaustive review of post-hoc methods is beyond the scope of this paper. We make no claim that the limitations presented in this manuscript apply to all methods and all contexts; rather, our analysis is limited in scope to tabular data. The risks and limitations discussed in this section do not apply to all post-hoc methods.

Given the utility of post-hoc explainability methods, proponents of black-box models claim that, despite their complexity, sufficient interpretability can be generated to allow for human oversight post-model deployment. This, in turn, justifies their use for high-stakes decision-making. Any malfunctions, such as discrimination, can be detected and mitigated through renewed design. However, post-hoc explainability methods suffer pitfalls, which ought to substantively challenge this belief.

The first is that post-hoc explainability methods only approximate their underlying models. They are, therefore, not faithful and suffer from low fidelity [62, 68]. This runs the risk that the interpretability generated might inaccurately reflect feature spaces of underlying models [62]. The surrogate models generated through post-hoc explainability methods might not accurately reflect their black-box counterparts. The accuracy to which post-hoc explainability methods represent their underlying models is currently under study and largely debated [12]. It is, therefore, an unknown. This means that any insights generated through post-hoc explainability methods should be taken with the preverbal "*grain of salt*" (read skepticism).

It is important to note that proponents who claim that post-hoc explainability methods completely and faithfully represent the computations of their black-box counterparts suffer from a categorical fault. Propose that post-hoc explainability methods completely and faithfully represented black-box models. These models would then be used instead of their "black-box" counterparts [62]. But, more importantly, this would mean that black-box models, themselves, would be interpretable, as post-hoc explainability methods could be used to make them so. As a result, they would not be black-boxes models, as black-box models are inherently opaque by definition [62].

Second, post-hoc explainability methods suffer from instability. This is best demonstrated in the presence of uncertainty in Local Model-Agnostic Explanations ("LIME") due to their randomness in sampling and procedure [81]. The figure below illustrates the lack of stability for a LIME analysis run through the eight input features from the COMPAS dataset. When assessing the top 5 important features for the COMPAS set, LIME generated two different explanations for the same instance on a Random Forest classifier we trained. Note that one explanation included race and the other explanation included gender. Additionally, the varying predictive contribution of each constant feature differed too. The result is that LIME explanations, like most post-hoc explainability methods, lack stability and produce different explanations for the same instance [73, 79]. Post-hoc explainability methods are, therefore, unstable (Fig. 3).

Furthermore, some post-hoc explainability methods are permutation-based and make the incorrect assumption of feature independence [44]. This means assuming that features are correlated but are not statistically dependent. This can generate misleading explanations [44]. Finally, post-hoc explainability methods are computationally expensive. In the case of LIME, the explanation requires generating a local model for each instance with a large number of samples [2, 64]. In the case of Shapley Values, feature importance requires identifying the average marginal contribution over all possible feature subsets, which can be very expensive over a large number of features. Hence, generating individual explanations for the entire dataset can be impractical [41]. This naturally limits their practical application.

1.7 Post-hoc explainability methods and non-discrimination law: analysis

Prima facie discrimination requires a claimant to "*adduce facts that are adequate and sufficient to raise a suspicion of discrimination*" [58]. This means, in essence, that a claimant must provide sufficient evidence to convince a court that it could (not that it should) conclude, failing an adequate explanation, that discrimination is present. The burden proof is lower than that of proving that, indeed, discrimination is present (or, rather, that it should conclude that discrimination is present). It is trite that the burden of proof in such cases lies with the claimant [69].

Statistics are often provided to overcome this burden. However, the proffering of statistics must meet certain conditions for success. In this regard, the CJEU noted, in *Enderby v Frenchay Health Authority and Secretary of State of Health*, that "[*i*]*t is for the national court to assess*

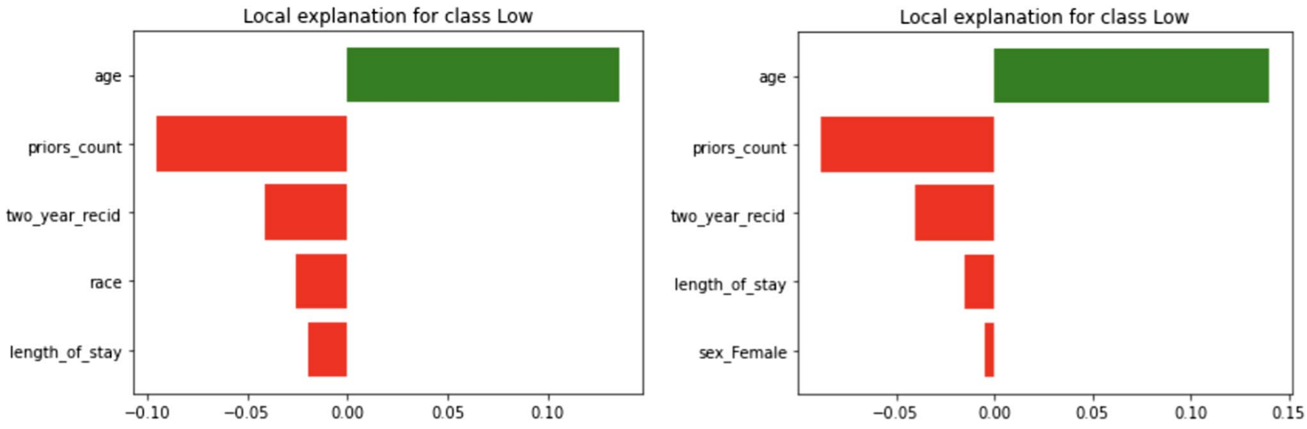

Fig. 3 The top 5 important features for the COMPAS dataset generated by LIME in two separate analyses

whether it may take into account those statistics, that is to say, whether they cover enough individuals, whether they illustrate purely fortuitous or short-term phenomena, and whether, in general, they appear to be significant." In light of the above, issues surrounding statistical validity and/or their methodological approaches are bound to arise [58]. These latter concerns—regarding post-hoc explainability methods—are the principal focus of this paper.

The first consideration is the technical appropriateness of post-hoc explainability methods. As alluded to above, non-discrimination law focuses predominately on outcome parity [74]: specifically, the disparity in outcomes for protected classes and/or sub-groups. Conversely, post-hoc explainability methods are not concerned with outcome parity per se but rather the reasons for such. In examining feature importance, post-hoc explainability methods are focused on the "how" and "why." In doing so, post-hoc explainability methods apply different techniques for removing a feature and, consequently, estimating its impact on the model outcome. The more significant the impact that is identified on the outcome from the absence of a feature, the more that feature is considered important. For example, the feature of gender-female (if binary) or, rather, race might be removed to demonstrate how much it contributes to the model outcome. Feature importance established through selective feature removal techniques can establish a correlation between features and model outcomes, but they cannot establish causality. This means that the focal points of post-hoc explainability methods and non-discrimination law are conceptually different, which has a direct bearing on the appropriateness of these methods. It seems inapt to use post-hoc explainability methods to examine non-discrimination law when the latter is not, per se, interested the "how" and "why," but rather the mere presence of outcome parity—at least in terms of proving *prima facie* discrimination. Therefore, the use of post-hoc explainability methods to demonstrate *prima facie* discrimination seems ill-advised.

Non-discrimination law's focus on outcome parity is merit-worthy and rational. Due to the impractical evidential burden of demonstrating discrimination (specifically the systemic and tacit nature of indirect discrimination), non-discrimination law has experienced a doctrinal reorientation from requiring clear discriminatory causation to, rather, discriminatory outcome [69]. The manifestation of this shift is two-fold: (a) no longer is intent needed to prove discrimination [24], and (b) legal causality is imputed once *prima facie* discrimination—outcome parity—has been demonstrated [69]. Given the limited utility of post-hoc explainability methods in exploring the "how" and "why" of models, instead of model outcomes, they appear ill-suited for such instances.

Despite this doctrinal evolution of non-discrimination law, the procedural requirement for claimants to demonstrate *prima facie* discrimination is impractical for instances of machine learning. As aptly identified by Tischbirek [69], "*the court is summoned to retrospectively assess a specific incident that has occurred between the litigants. Classic civil procedure mechanisms of generating and digesting the necessary extra-legal knowledge are narrowly tailored to serve this purpose*". The result is that the procedural requirements of non-discrimination law are orientated towards case—or rather individual—specific facts, specifically by placing the evidentiary *onus* on litigant parties and, more so, the claimant. However, the modern practice of machine learning is often technically obscure (as articulated above). Moreover, it is highly unlikely that claimants will ever have access to information about the inner workings of machine learning. Naturally, information surrounding model inputs, design, exploration, development, and operationalization are protected by trade secrecy laws [77]. This makes overcoming the evidentiary *onus* difficult, if not near impossible. To labour this point further, the modern practice of applied commercial machine learning more conceptually, it has been proposed, is "*aimed at deriving population-level insights from data subjects for population-level applicability, not individual-level insights*" [72, p. 2]. As a result, the outcomes of models are hard to detect and/or appreciate for non-technical individuals who are located within the broader context of society and are effected by them [25]. This is more so given the subtle nature in which models are operationalized [25, 77], although new draft European laws look to change this (Proposal for Regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (artificial Intelligence Act) and amending union legislative Acts, COM/2021/206, Art. 52(1)). The result is the procedural requirements associated with of non-discrimination law are further hurdles for claimants.

Nonetheless, it might still be proposed that post-hoc explainability methods serve an important function in tandem with outcome parity tests. They help explain the "how" and "why" when outcomes parities are detected. Therefore, they serve a utility in demonstrating and/or investigating *prima facie* discrimination, specifically via helping to (a) exemplify a legal *nexus* between (i) unfavourable outcomes and (ii) the application of a criterion that either directly or indirectly appeals to a prohibited ground of discrimination; and, thereby, (b) affirming that this legal *nexus* is not "*purely fortuitous or short-term phenomena*", but rather significant. This might be done, for example, in cases of investigations and/or audits by regulators. The proposed investigatory and/or auditing role of regulators in machine learning oversight has been widely advocated for [31, 51, 76, 78]. Although at first logical, the proposed use of post-hoc explainability methods in this capacity suffers from two shortcomings. The first is the technical shortcomings—or rather instability—of post-hoc explainability methods; the second is the limited insights post-hoc explainability methods actually offer.

The technical shortcomings of post-hoc explainability methods significantly hamper their application. As discussed above, post-hoc explainability methods' that distil and mimic the nature of their black-box counter parts suffer from low fidelity, such as LIME [62, 63]. In approximating underlying models, these post-hoc explainability methods might not accurately reflect feature spaces. Due to the nature of black-box modeling, the ability to faithfully test the fidelity and/or accuracy of these given post-hoc explainability method remains impractical. This means that the insights generated through these post-hoc explainability methods can be beleaguered with legal claims of inaccuracy, which cannot largely be disputed. The result is that using such methods to justify a legal *nexus* is susceptible to constant skepticism. Depending on the convictions of presiding officers, such skepticism might be legally fateful.

The instability of some post-hoc explainability methods such as LIME further calls into question their use. As demonstrated above, due to their randomness in sampling and procedure, the outcomes generated from post-hoc explainability methods are not consistent. This makes reliance on their use questionable, especially in legal proceedings where the certainty of the facts at hand is central in decision-making [59, 60]. Opponents might claim that—due to the law of large numbers—although complete stability and/or faithfulness can never be obtained, through numerous iterations of post-hoc explainability modeling, somewhat stable results can be derived and relied on. This is true; however, the computational expense of post-hoc explainability methods is a hurdle to this. Although perhaps not detrimental, the instability surrounding post-hoc explainability methods does limit their practical application, especially for large complex models.

In finality, the insights generated from post-hoc explainability methods are limited. Post-hoc explainability methods can only demonstrate specific types of model biases, traditionally evaluation and aggregation bias more than measurement bias. The confined use of post-hoc explainability methods to assess these bias types is by no means a detriment to the methods themselves. However, advocates of post-hoc explainability methods and their possible role within the Law must remain mindful of their limited use. Post-hoc explainability methods can only reveal a particular part of a model's workings, namely measurement, evaluation, and aggregation bias. Post-hoc explainability methods cannot be used to demonstrate historical bias—or, rather, "social bias." This means that they will be particularly limited in unearthing indirect systemic discrimination due to its persuasive and tacit nature in datasets and dataset collection [40, 75].

2 Conclusion

The above demonstrates the limitations of post-hoc explainability methods in demonstrating *prima facie* discrimination. Post-hoc explainability methods lack the orientation towards illustrating outcome parity, which is essential for EU non-discrimination law. Moreover, their technical shortcomings mean that they are in some cases unstable and suffer from low fidelity. Subsequently, they cannot faithfully demonstrate the absence of discrimination (the null hypothesis). Finally, the limited bias types unearthed through post-hoc explainability methods mean that their use must be confined and contextually appreciated. The utility of post-hoc explainability methods is useful, especially in model design and development, but they are possibly limited for regulatory or legal use. They, therefore, cannot be championed as silver bullets and/or can longer be appreciated alone in a void ignorant of broader fairness metrics.

If post-hoc explainability methods cannot *prima facie* prove discrimination, the substantive legal weight they might be able to carry does not bode well. Accordingly, if one cannot guarantee the insights and/or inner workings of a black-box model, they ought not to use them in instances where its decisions can have long-lasting and/or dramatic effects.

Our goal is to encourage legal practitioners and compliance officers to embrace a more holistic view of the inherit risks involved in deploying machine learning model in high-stakes decision-making and to recognize the insufficiency of some post-hoc explanation methods as the sole mechanism in achieving fairness, accountability, and transparency, where issues of non-discrimination ought to be of principal concern.

References

1. Adadi, A., Berrada, M.: Peeking inside the Black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/access.2018.2870052
2. Ahmad, M.A., Eckert, C., Teredesai, A.:. Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. https://doi.org/10.1145/3233547.3233667 (2018)
3. Alom, M., Taha, T., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M., Asari, V., et al.: The history began from alexnet: a comprehensive survey on deep learning approaches. https://arxiv.org/abs/1803.01164 (2018)
4. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. Retrieved from ProPublica: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
5. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Muller, K.: How to explain individual classification decisions. J Mach Learn Res 1803–1831. https://dl.acm.org/doi/pdf/10.5555/1756006.1859912 (2010)

6. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
7. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.:. Benchmarking and survey of explanation methods for black box models. https://arxiv.org/pdf/2102.13076.pdf (2021)
8. Bratko, I.: Machine learning: between accuracy and interpretability. In: Della Riccia, G., Lenz, H.-J., Kruse, R. (eds.) Learning, Networks and Statistics. ICMS, vol. 382, pp. 163–177. Springer, Vienda (1997)
9. Breiman, L.: Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci (2001). https://doi.org/10.1214/ss/1009213726
10. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. J Artif Intell Res **70**, 245–317 (2021). https://doi.org/10.1613/jair.1.12228
11. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. **3**(1), 205395171562251 (2016). https://doi.org/10.1177/2053951715622512
12. Camburu, O., Giunchiglia, E., Foerster, J., Lukasiewicz, T., Blunsom, P.: Can I trust the explainer? Verifying post-hoc explanatory methods. https://arxiv.org/abs/1910.02065 (2019)
13. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning Interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019). https://doi.org/10.3390/electronics8080832
14. Choi, E., Bahadori, M., Kulas, J., Schuetz, A., Stewart, W., Sun, J.: RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. Adv Neural Inf Process Syst 3504–3512 (2016). https://arxiv.org/abs/1608.05745
15. Council of Europe: European Court of Human Rights: Handbook on European non-discrimination law. Council of Europe: European Court of Human Rights, Strasburg (2018)
16. Covert, I., Lundberg, S., Lee, S.: Explaining by removing: a unified framework for model explanation. https://arxiv.org/abs/2011.14878 (2020)
17. Cranor, L.: A framework for reasoning about the human in the loop. https://www.usenix.org/legacy/event/upsec/tech/full_papers/cranor/cranor.pdf (2008)
18. Deng, J., Dong, W., Socher, R., Li, L., Kai, L., Li, F.-F.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2009.5206848 (2009)
19. Douglas-Scott, S.: The European Union and human rights after the treaty of Lisbon. Hum. Rights Law Rev. **11**(4), 645–682 (2011). https://doi.org/10.1093/hrlr/ngr038
20. Doyle, O.: Direct discrimination, indirect discrimination and autonomy. Oxf. J. Leg. Stud. **27**(3), 537–553 (2007). https://doi.org/10.1093/ojls/gqm008
21. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Commun. ACM **63**(1), 68–77 (2019). https://doi.org/10.1145/3359786
22. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on-ITCS '12. https://doi.org/10.1145/2090236.2090255 (2012)
23. Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M.: Decoupled classifiers for fair and efficient machine learning. https://arxiv.org/abs/1707.06613 (2017)
24. Ellis, E., Watson, P.: Key concepts in EU anti-discrimination law. EU Anti-Discrimination Law (2012). https://doi.org/10.1093/acprof:oso/9780199698462.003.0004
25. Ernst, C.: Artificial intelligence and autonomy: self-determination in the age of automated systems. Regulat. Artif. Intell. (2019). https://doi.org/10.1007/978-3-030-32361-5_3
26. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2783258.2783311 (2015)
27. Floridi, L., Chiriatti, M.: GPT-3: its nature, scope, limits, and consequences. Mind. Mach. **30**(4), 681–694 (2020). https://doi.org/10.1007/s11023-020-09548-1
28. Foster, K.R., Koprowski, R., Skufca, J.D.: Machine learning, medical diagnosis, and biomedical engineering research - commentary. Biomed. Eng. Online **13**(1), 94 (2014). https://doi.org/10.1186/1475-925x-13-94
29. Gerards, J., Xenidis, R.: Algorithmic discrimination in Europe: Challenges and opportunities forgender equality and non-discrimination law. Publications Office of the European Union (2021)
30. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). https://doi.org/10.1109/dsaa.2018.00018 (2018)
31. Girasa, R.: AI US policies and regulations. Intell. Disrupt. Technol Artif (2020). https://doi.org/10.1007/978-3-030-35975-1_3
32. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2019). https://doi.org/10.1145/3236009
33. Guiraudon, V.: Equality in the making: implementing European non-discrimination law. Citizsh. Stud. **13**(5), 527–549 (2009). https://doi.org/10.1080/13621020903174696
34. Hall, P., Gill, N., Schmidt, P.: Proposed guidelines for the responsible use of explainable machine learning. https://arxiv.org/abs/1906.03533 (2019)
35. Hall, P., Gill, N., Kurka, M., Phan, W.: Machine learning interpretability with H20 driverless AI. Mountain View: H20. https://www.h2o.ai/wp-content/uploads/2017/09/MLI.pdf (2017)
36. Hand, D.J.: Classifier technology and the illusion of progress. Stat. Sci. (2006). https://doi.org/10.1214/088342306000000060
37. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems. https://arxiv.org/abs/1610.02413 (2016)
38. Kantola, J., Nousiainen, K.: The European Union: Initiator of a New European Anti-DiscriminationRegime? In: Krizsan A, Skjeie H, Squires J (eds) Institutionalizing Intersectionality: The ChangingNature of European Equality Regimes. Palgrave Macmillan (2012)
39. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the COMPAS recidivism algorithm. ProPublica 1–16 (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
40. Laugel, T., Lesot, M., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: unjustified counterfactual explanations. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2019/388 (2019)
41. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9
42. Mair, J.: Direct discrimination: limited by definition? Int. J. Discrim. Law **10**(1), 3–17 (2009). https://doi.org/10.1177/135822910901000102

43. Maliszewska-Nienartowicz, J.: Direct and indirect discrimination in European Union Law—how to draw a dividing line? Int. J. Soc. Sci. 41–55 (2014). https://www.iises.net/download/Soubory/soubory-puvodni/pp041-055_ijoss_2014v3n1.pdf
44. Molnar, C., Konig, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio G., Grosse-Wentrup M., Bischl, B.: Pitfalls to avoid when interpreting machine learning models. https://arxiv.org/abs/2007.04131 (2020)
45. Meske, C., Bunde, E.: Transparency and trust in human–AI-interaction: the role of model-agnostic explanations in computer vision-based decision support. Artif. Intell. HCI (2020). https://doi.org/10.1007/978-3-030-50334-5_4
46. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. **65**, 211–222 (2017). https://doi.org/10.1016/j.patcog.2016.11.008
47. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. **116**(44), 22071–22080 (2019). https://doi.org/10.1073/pnas.1900654116
48. Narayanan, A.: Translation tutorial: 21 fairness definitions and their politics. In: Proceedings of the Conference on. Fairness Accountability Transparency. https://fairmlbook.org/tutorial2.html (2018)
49. Nie, L., Wang, M., Zhang, L., Yan, S., Zhang, B., Chua, T.: Disease inference from health-related questions via sparse deep learning. IEEE Trans. Knowl. Data Eng. **27**(8), 2107–2119 (2015). https://doi.org/10.1109/tkde.2015.2399298
50. Onishi, T., Saha, S.K., Delgado-Montero, A., Ludwig, D.R., Onishi, T., Schelbert, E.B., Schwartzman, D., Gorcsan, J.: Global longitudinal strain and global circumferential strain by speckle-tracking echocardiography and feature-tracking cardiac magnetic resonance imaging: comparison with left ventricular ejection fraction. J. Am. Soc. Echocardiogr. **28**(5), 587–596 (2015). https://doi.org/10.1016/j.echo.2014.11.018
51. O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., Holzinger, K., Holzinger, A., Sajid, M.I., Ashrafian, H.: Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. Int. J. Med. Robot. Comput. Assist. Surg. **15**(1), e1968 (2019). https://doi.org/10.1002/rcs.1968
52. Qian, K., Danilevsky, M., Katsis, Y., Kawas, B., Oduor, E., Popa, L., Li, Y.: XNLP: A living survey for XAI research in natural language processing. In: 26th International Conference on Intelligent User Interfaces. https://doi.org/10.1145/3397482.3450728 (2021)
53. Pasquale, F.: The black box society, the secret algorithms that control money and information. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061 (2015)
54. Pasquale, F.: Toward a fourth law of robotics: preserving attribution, responsibility, and explainability in an algorithmic society. Ohio State Law J. https://ssrn.com/abstract=3002546 (2017)
55. Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., Turini, F.: Meaningful explanations of black box AI decision systems. Proc. AAAI Conf. Artif. Intell. **33**, 9780–9784 (2019). https://doi.org/10.1609/aaai.v33i01.33019780
56. Ribeiro, M., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. https://arxiv.org/abs/1606.05386 (2016)
57. Ribeiro, M., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778 (2016)
58. Ringelheim, J.: The burden of proof in antidiscrimination proceedings. A focus on Belgium, France and Ireland. Eur. Equal. Law Rev. (2019). https://ssrn.com/abstract=3498346
59. Rissland, E.: AI and legal reasoning. In: Proceedings of the 9th International Joint Conference on Artificial Intelligence. https://dl.acm.org/doi/abs/10.5555/1623611.1623724 (1985)
60. Rissland, E.L., Ashley, K.D., Loui, R.: AI and law: a fruitful synergy. Artif. Intell. **150**(1–2), 1–15 (2003). https://doi.org/10.1016/s0004-3702(03)00122-x
61. Robnik-Šikonja, M., Bohanec, M.: Perturbation-based explanations of prediction models. Hum. Mach. Learn. (2018). https://doi.org/10.1007/978-3-319-90403-0_9
62. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
63. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: a review of methods and applications. Proc. IEEE **109**(3), 247–278 (2021). https://doi.org/10.1109/JPROC.2021.3060483
64. Schwab, P., Karlen, W.: CXPlain: causal explanations for model interpretation under uncertainty. In: Advances in Neural Information Processing Systems. https://arxiv.org/abs/1910.12336 (2019)
65. Selbst, A.D., Barocas, S.: The intuitive appeal of explainable machines. SSRN Electron. J. (2018). https://doi.org/10.2139/ssrn.3126971
66. Suresh, H., Gong, J.J., Guttag, J.V.: Learning tasks for multitask learning. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3219819.3219930 (2018)
67. Suresh, H., Guttag, J.: A framework for understanding unintended consequences of machine learning. https://arxiv.org/abs/1901.10002 (2019)
68. Tan, S., Caruana, R., Hooker, G., Lou, Y.: Distill-and-compare. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3278721.3278725 (2018)
69. Tischbirek, A.: Artificial intelligence and discrimination: discriminating against discriminatory systems. Regul. Artif. Intell. (2019). https://doi.org/10.1007/978-3-030-32361-5_5
70. VanderWeele, T.J., Hernan, M.A.: Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. Am. J. Epidemiol. **175**(12), 1303–1310 (2012). https://doi.org/10.1093/aje/kwr458
71. Verma, S., Rubin, J.: Fairness definitions explained. Proc. Int. Workshop Softw. Fairness (2018). https://doi.org/10.1145/3194770.3194776
72. Viljoen, S.: Democratic data: a relational theory for data governance. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3727562
73. Visani, G., Bagli, E., Chesani, F., Poluzzi, A., Capuzzo, D.: Statistical stability indices for LIME: obtaining reliable explanations for machine learning models. J. Oper. Res. Soc. (2021). https://doi.org/10.1080/01605682.2020.1865846
74. Wachter, S., Mittelstadt, B., Russell, C.: Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3547922
75. Wachter, S., Mittelstadt, B., Russell, C.: Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. SSRN Electron. J. (2021). https://doi.org/10.2139/ssrn.3792772
76. Wang, W., Siau, K., Keng, S.: Artificial intelligence: a study on governance, policies, and regulations. Association for Information Systems AIS Electronic Library. http://aisel.aisnet.org/mwais2018/40 (2018)
77. Wischmeyer, T.: Artificial intelligence and transparency: opening the black box. Regul. Artif. Intell. (2019). https://doi.org/10.1007/978-3-030-32361-5_4

78. Wischmeyer, T., Rademacher, T.: Regulating Artificial Intelligence. International Springer Publications, New York City (2020). https://doi.org/10.1007/978-3-030-32361-5
79. Zafar, M., Khan, N.: DLIME: a deterministic local interpretable model-agnostic explaination approach for computer-aided diagnosis systems. https://arxiv.org/abs/1906.10263 (2019)
80. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333. PMLR. https://proceedings.mlr.press/v28/zemel13.html (2013)
81. Zhang, Y., Song, S., Sun, Y., Tan, S., Udell, M.: "Why Should You Trust My Explaination?" Understanding uncertainty in LIME explanations. https://arxiv.org/abs/1904.12991 (2019)
82. Zuiderveen Borgesius, F.J.: Strengthening legal protection against discrimination by algorithms and artificial intelligence. Int. J. Hum. Rights **24**(10), 1572–1593 (2020). https://doi.org/10.1080/13642987.2020.1743976
83. Zuiderveen Borgesius, F.J.: Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe, Directorate General of Democracy. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (2018)