AI and Ethics
https://doi.org/10.1007/s43681-023-00291-8

ORIGINAL RESEARCH

Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans

Sergio Genovesi[1] · Julia Maria Mönig[1] · Anna Schmitz[2] · Maximilian Poretschkin[2] · Maram Akila[2] · Manoj Kahdan[3] · Romina Kleiner[3] · Lena Krieger[4] · Alexander Zimmermann[4]

Received: 22 December 2022 / Accepted: 24 April 2023

Abstract

In the current European debate on the regulation of Artificial Intelligence there is a consensus that Artificial Intelligence (AI) systems should be fair. However, the multitude of existing indicators allowing an AI system to be labeled as "(un)fair" and the lack of standardized, application field specific criteria to choose among the various fairness-evaluation methods makes it difficult for potential auditors to arrive at a final, consistent judgment. Focusing on a concrete use case in the application field of finance, the main goal of this paper is to define standardizable minimal ethical requirements for AI fairness-evaluation. For the applied case of creditworthiness assessment for small personal loans, we highlighted specific distributive and procedural fairness issues inherent either to the computing process or to the system's use in a real-world scenario: (1) the unjustified unequal distribution of predictive outcome; (2) the perpetuation of existing bias and discrimination practices; (3) the lack of transparency concerning the processed data and of an explanation of the algorithmic outcome for credit applicants. We addressed these issues proposing minimal ethical requirements for this specific application field: (1) regularly checking algorithmic outcome through the conditional demographic parity metric; (2) excluding from the group of processed parameters those that could lead to discriminatory outcome; (3) guaranteeing transparency about the processed data, in addition to counterfactual explainability of algorithmic decisions. Defining these minimal ethical requirements represents the main contribution of this paper and a starting point toward standards specifically addressing fairness issues in AI systems for creditworthiness assessments aiming at preventing unfair algorithmic outcomes, in addition to unfair practices related to the use of these systems. As a final result, we indicate the next steps that can be taken to begin the standardization of the three use case-specific fairness requirements we propose.

Keywords Artificial intelligence · Data science · Fairness · Fairness metric · Standardization

✉ Sergio Genovesi
genovesi@uni-bonn.de

Julia Maria Mönig
moenig@uni-bonn.de

Anna Schmitz
anna.schmitz@iais.fraunhofer.de

Maximilian Poretschkin
maximilian.poretschkin@iais.fraunhofer.de

Maram Akila
maram.akila@iais.fraunhofer.de

Manoj Kahdan
kahdan@time.rwth-aachen.de

Romina Kleiner
kleiner@time.rwth-aachen.de

Lena Krieger
Lena.Krieger@din.de

Alexander Zimmermann
Alexander.Zimmermann@din.de

[1] University of Bonn, Bonn, Germany

[2] Fraunhofer IAIS, Sankt Augustin, Germany

[3] RWTH, University of Aachen, Aachen, Germany

[4] DIN, Berlin, Germany

Published online: 08 May 2023

1 Introduction

In the current European debate on the regulation of Artificial Intelligence there is a consensus that Artificial Intelligence (AI) systems should be developed in a human centered way and should be "trustworthy" [23, 24, 31, 99]. According to these documents, one value that constitutes "trustworthiness" is fairness. Many current publications on AI fairness predominantly focus on avoiding or fixing algorithmic discrimination of groups or individuals and on data-de-biasing, offering different metrics as tools to evaluate whether groups or individuals are treated differently [8, 71, 96]. Moreover, the International Standardization Organization (ISO)/International Electrotechnical Commission (IEC) TR 24028:2020, *Information technology—Artificial Intelligence—Overview of trustworthiness in Artificial Intelligence* lists fairness as an essential part for ensuring trustworthiness in AI (ISO/IEC 2020). However, the multitude of existing indicators allowing the labeling of an AI system as "(un) fair" and the lack of standardized, application field specific criteria to choose among the various fairness-evaluation methods makes it difficult for potential auditors to arrive at a final, consistent judgment [24, 96, 98]. The increasing need for standardized methods to assess the potential risks of AI systems is also highlighted by the draft for an "Artificial Intelligence Act" suggested by the European Commission in April 2021, which, in accordance with the so-called "New Legislative Framework," ascribes a major role to "Standards, conformity assessment, certificates [and] registration" (Chapter 5).

Focusing on a concrete use case in the application field of finance, the main goal of this paper is to define standardizable minimal ethical requirements for AI fairness evaluation. In Sect. 2, we explore different understandings of fairness from three perspectives and address the different vantage points of many stakeholders involved in the development, commercialization, and use of AI systems. In Sect. 3, we discuss the example of a risk scoring machine learning (ML) model for small personal loans. As a main contribution of the paper, we suggest ethical minimal requirements that should be complied with when evaluating fairness and highlight a preferred fairness metric for fairness-evaluation purposes in this specific application field. In Sect. 4, we investigate how to translate our research findings into standardization criteria to be used when assessing ML credit scoring systems for small personal loans.

2 Defining fairness

2.1 AI ethics

In the current AI ethics discussions, fairness is generally framed in accounts of distributive justice and is broadly referred to as unbiased distribution of access to services and goods—e.g., access to treatments in healthcare [39, 81] or access to credit [65]—and as absence of discrimination, understood as unjustified unequal treatment of groups or individuals [72, 76].[1]

Concerning distributive justice and non-discrimination as equal treatment, one of the primary contemporary philosophical references is the Rawlsian idea of equality of opportunities. This idea requires that citizens having the same talents and being equally motivated should receive the same educational and economic opportunities regardless of their wealth or social status [83] (p. 44). Since in the social praxis basic rights and liberties are neither accessible nor enjoyable in the same way for different citizens, society should take adequate measures in order for all citizens to enjoy their rights and liberties. Rawls develops on this intuition stating that "the worth of liberty to persons and groups depends upon their capacity to advance their ends within the framework the system defines. […] Some have greater authority and wealth, and therefore greater means to achieve their aims" [82] (p. 179). Consequently, to avoid the exaggeration of the unequal enjoyment of basic rights and liberties, a fair society must enact compensation mechanisms to maximize their worth to the least advantaged [82] (ibid). It is essential to avoid the development of vicious circles of (un)privilege-polarization in society due to the moral harm they produce. Therefore, removing the opportunities of the less privileged to truly benefit from their rights and liberties results in a harmful form of negative discrimination that amplifies economic inequalities and undermines the chances for the less advantaged to live an autonomous life and set self-determined goals. This is a form of disrespect toward their personhood [29, 68] and can amplify social resentment.

The philosophical debate about Rawls' theory and other forms of "egalitarianism" could help clarify current emerging issues concerning algorithmic fairness [11]. Egalitarianism in this sense means that "human beings are in some fundamental sense equal and that efforts should be made to avoid and correct certain forms of inequality" [11] (p. 2). Many approaches try to determine the kind of equality that

[1] Noble highlights intersectional questions of fairness underlining the adverse effects that "algorithms of oppression" have on black women. In general, feminist scholars have stressed that unfairness and injustices usually go hand in hand with domination and oppression [75, 75].

should be sought and which inequalities should be avoided in civil society to uphold the fundamental equality of human beings: among others, equality of preference-satisfaction [19], equality of welfare, income, and assets [28], and the equality of capabilities to achieve their goals [87]. However, regarding the application of these views on algorithmic decisions, defending an equal opportunity approach rather than an equal outcome is not always the most effective solution. If, for candidate selection or calculation of insurance, focusing on equal opportunity might lead to increase "economic justice," in other contexts, such as during airport security checks, equality of outcome in the form of "parity of impact" could help establish a sense of social solidarity avoiding the over-examination of certain groups [11] (p. 7). Thus, the choice of a specific approach to evaluate (in)equality depends on the specific application context. As Balayn and Gürses claim, the regulation of AI must go "beyond debiasing" [6]. Mere data-based or outcome-based solutions trying to solve local distributive issues of a system, such as trying to solve a racial bias in image recognition software by enlarging the data basis with pictures from people of all ethnic backgrounds, are not sufficient alone to address structural inequality issues at the root [60].[2] If decision-making processes that influence people's access to opportunities are biased, the intersection between algorithmic fairness and structural (in)justice requires investigation [51].

In addition, other fairness issues that are indirectly related with the algorithmic outcome, but rather with the entire system design and application processes, as well as with their consequences for society, can occur. These aspects of fairness also overlap with other human rights and societal values. For instance, a structural fairness issue of many ML systems is the phenomenon of "digital labor" [33], referring (among other things) to the precarious work conditions and the very low pay of many click workers generating training data for ML systems. In addition, the commodification of privacy related with the use of many digital services raises fairness issues since users are often kept unaware of the exact use of their data, and they are not always in the position to defend their right to privacy [74, 91, 95]—being so a disadvantaged stakeholder compared with the service providers. Finally, addressing the sustainability concerns that emerged during the so-called "third wave" of AI ethics, global and intergenerational justice can be highlighted as fairness issues [37, 93]. Considering intergenerational justice means to add a temporal, anticipatory dimension to our understanding of fairness and extend the claim for the equity of human living conditions—as, for instance, expressed in the UN's Sustainable Development Goals (SDG)—also to the future and not only limiting it to present generations. In the practice, fairness toward future generations means acting sustainably.

These considerations lead us to the following preliminary understanding of fairness in the context of an AI ethics assessment. First, focusing on the unbiased distribution of access to services and goods and on the absence of discrimination of groups or individuals, fairness means the equal treatment of people regardless of their sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation,[3] when it comes to granting or denying access to products, services, benefits, professional or educational opportunities, and medical treatments based on an automated evaluation and classification of individual or groups. In addition, a fair system should not involve work exploitation or the violation of human rights of any of the involved stakeholders during its life cycle. Moreover, the real-world application of the system should not create or amplify power unbalances between stakeholders, nor place specific stakeholders' groups in a disadvantaged position.

2.2 Data science

Fairness is discussed in the context of data and data-driven systems whose inherent patterns or statistical biases[4] can be interpreted as "unfair." Here, it is important to emphasize that the evaluation of whether certain patterns are "fair" or "unfair" transcends the specific expertise of data scientists and requires further legal, philosophical, political, and socio-economic considerations. What is being explored in data science under the term "fairness" are quantitative concepts to identify patterns or biases in data, in addition to technical methods to mitigate them.

Data analysis and data-based modeling of real-world relationships have progressed in recent years especially through Machine Learning (ML). ML is a subdiscipline of AI research in which statistical models are fitted to so-called training data, recognize patterns and correlations in this data, and generate predictions for new (input) data on this basis. ML methods have become a particular focus of fairness research, as they provide everyday applications using personal data, e.g., employment decisions, credit scoring,

[2] In addition, this cannot be the solution to this problem because it would feed even more data into the systems of the service providers and would therefore support their data hunger and business logic.

[3] These are the protected attributes listed in Article 21 (Non-discrimination) of the EU Charter of Fundamental Rights.

[4] While the term bias is often connoted negatively in other disciplines (with discriminatory effects, etc.), in this context (computer science) it merely means a statistical deviation from the standard [41] [7]. Whether this constitutes a case of discrimination is another question.

and facial recognition [71]. Furthermore, they pose the challenge that bias within the training data might lead to biased model results.

2.2.1 Short introduction to machine learning

ML-based applications have enabled technological progress which can particularly be attributed to the fact that their functionality is based on learning patterns from data. By this means, ML methods provide approaches to solving tasks that could not be effectively addressed by "traditional" software fully specified by human rules. In particular, deep neural networks, a type of ML method involving vast amounts of data, have significantly advanced areas, such as image [26] and speech recognition [13], in addition to predicting complex issues, for instance medical diagnostics [102] and predictive maintenance [17].

ML methods are designed to learn from data to improve performance on a given task [41]. A task can be viewed as finding a mapping that, for an input *x*, assigns an output *y* which is useful for a defined purpose. One ML task that is particularly relevant for fairness is classification. The purpose of classification is to identify to which of a set of categories a given input belongs, for instance, whether a person is creditworthy or not. ML is about finding such a model *f* that solves a task effectively by $y = f(x)$. To achieve this, a learning algorithm adjusts parameters within the model. The fitness of the model for the given task can be evaluated using quantitative measures. Such quantitative indicators of model or data properties are generally referred to as "metrics." For example, a typical performance metric for classification tasks is precision, which measures the proportion to which the classification to a certain category by the model was correct.

The data which ML methods use to build a model, called "training data," is a collection of input examples[5] that the model is expected to handle as part of the task. A single example in the data is called a datapoint. For classification tasks, a datapoint in the training data of a ML model contains, in addition to the example *x*, a "ground-truth" label that specifies how the ML model should process the respective input *x*. Following on the example of creditworthiness classification, the training data for a ML model addressing this task may be drawn from previous credit applications, and the individual datapoints could include features, such as income, age, or category of work activity (e.g., self-employed, employed). Moreover, each datapoint should also contain a ground-truth label that could be derived from manual processes or, if possible, from the observation whether in the given examples the loans were repaid in full.

When building a ML model, the training data is used to adjust the internal model parameters which determine the mapping through *f*. For instance, in a neural network, the weights assigned to the network's edges are adjusted by the learning algorithm during model building, a phase which is also called "training." Overall, ML is an optimization procedure that finds internal model parameters such that they optimize a defined performance metric on a training dataset. In this case, the performance metric specified as optimization objective is referred to as "loss function." For example, in a classification task, a quantitative measure of the distance between ground-truth and model output could be used as a loss function. Consequently, such a model would be generated in training, which optimally approximates the relationships between *x* and *y* provided in the training data.

Underlying the ML approach of fitting a model to training data is the idea that the model infers patterns which help produce valuable outputs when applied to new data. The term "generalizability" is used to describe the aim that the model performs well on data not seen during training. Thus, for model evaluation, an additional test dataset different from the training data is used. Given training and test data, according to Goodfellow et al. model quality is indicated by two quantities: (i) the training error measured by the loss function, and (ii) the difference between training and test error [41]. A model with a large training error is called "underfitting," while one with a low training error but large difference between training and test error is "overfitting" the training data.

2.2.2 Meaning and challenges of "fairness"

Data has a crucial impact on the quality of a ML model. In computer science, data quality has already been researched for "classical" information systems, where it is considered especially regarding large amounts of stored operational and warehousing data, e.g., a company's client database. Numerous criteria for data quality have been proposed, which can be mapped within four dimensions: "completeness, unambiguity, meaningfulness, and correctness" [100]. Only recently has the operationalization of data quality specifically for ML been explored [46]. The issue of data completeness, relating to the training and test data sufficiently capturing the application domain, is particularly relevant in this context. There is a high risk that an ML model has low statistical power on data either not included or statistically insignificant in its training set. To prevent strong declines in a model's productive performance, measures are being researched for dealing with missing or underrepresented inputs [46] as well as for detecting distribution skews, e.g., between training and production data [12].

[5] Reinforcement learning methods that learn from interaction with systems or humans are not considered in this description.

In certain tasks and application contexts, individuals are affected by the outputs of a ML model. For example, ML models are being used to support recruiting processes, decisions on loan approval, and facial recognition [71]. Consequently, it is essential that the model performs equally well for all individuals. The research direction in data science that addresses related issues from a technical perspective is referred to under the term "fairness." Clearly, the motivation for "fairness" in data or ML models does not derive from a technical perspective, nor does data science as a scientific discipline provide a sufficient basis for evaluating under which circumstances these should be classified as "fair." The approaches and methods researched in this area are usually neutral, as they can be applied to structurally similar scenarios that do not involve individuals.

Regarding model quality, the data are a central object of study in fairness from two perspectives. First, aspects of data quality should not differ regarding particular groups of people. Regarding the dimension of completeness, for instance, certain population groups could be underrepresented in the training data resulting in a lower performance of the ML model with respect to these groups [14]. Another example is that the ML model might infer biased patterns from the training data if their representativeness is compromised by non-random sampling, for example, if predominantly positive examples are selected from one population group but negative examples are selected from another. Second, even if data are of high quality from a technical perspective, they may (correctly) reflect patterns that one would like to prevent from being reproduced by the ML model trained on it. For instance, data might capture systemic bias rooted in (institutional) procedures or practices favoring or disadvantaging certain social groups [86]. The technical challenge that arises here is fitting a model to the training data but simultaneously preventing inferring certain undesirable patterns that are present. Overall, proceeding from the variety of biases[6] identified to date, both measures that "*detect*" and measures that "*correct*" (potentially unfair) patterns in datasets and models are being explored [48] (p. 1175).

2.2.3 Measures that "detect"

Aiming to "*detect*," one research direction is concerned with developing technical approaches to disclose and quantify biases in the first place. Numerous "fairness metrics" have been presented [96], particularly in light of providing statistical evidence for unequal treatment in classification tasks. Corresponding to the approach of identifying and comparing groups for identifying bias, so-called "group fairness metrics" constitute a large part of the fairness metrics presented to date. These metrics compare statistical quantities regarding groups defined on the basis of certain attributes in a dataset (e.g., a group could be defined by means of age, gender, or location if these attributes are provided in the data). Among the group fairness metrics, one can further distinguish between two types: i) metrics which compare the distribution of outputs, and ii) metrics which compare the correctness of the outputs with respect to different groups. An example of the first type is to measure the discrepancy to which a certain output is distributed by percentage among two different groups. This quantification approach is called "statistical parity," and Sect. 3.3. provides a detailed elaboration. The second type of metrics focus on model quality and compare performance-related aspects with respect to different groups (e.g., specific error rates or calibration). For instance, the metric "equal opportunity" [96] calculates the difference between the true-positive rates of a model on the respective data subsets representing two different groups. Such metrics can highlight model weaknesses by providing insight on where the model quality may be inconsistent.

Besides group fairness metrics, further measures have been developed to disclose biases. Two examples are "individual fairness" [27] and "counterfactual fairness" [63]. "Individual fairness" is based on comparing individuals. Therefore, a distance metric is defined that quantifies the similarity between two datapoints. The underlying idea of this approach is that similar model outputs should be generated for similar individuals. In addition, measurable indicators for an entire data set have been derived using such a distance metric, for example, "consistency" [104]. Similarly, inequality indices from economics such as the generalized entropy index have also been proposed as bias indicators for datasets [89], which require a definition of individual preferences. "Counterfactual fairness" considers individual datapoints, similar to "individual fairness"; however, it examines the effect of changing certain attribute values on model outputs. This can be used to uncover if the model would have generated a different output for an individual if they had a different gender, age, or ethnicity, for example. Many of the presented bias quantification and detection approaches have been implemented in (partially) open-source packages and tools [3, 9, 40] and are likewise applicable to input–output-mappings not based on ML.

Different fairness metrics might pursue different target states, e.g., balanced output rates between groups (statistical parity) versus balanced error rates (equal opportunity). Therefore, they also differ greatly in their potential conflict with other performance goals. For instance, consider a

[6] A variety of bias causes and types has been explored that cannot be fully mapped here. For a categorization, in line with the two viewpoints described, into computational as well as human and systemic bias, we refer to Schwartz et al. [86]. For a categorization of biases along the feedback cycle of data, algorithm and user interaction, see [71], and for a mapping of biases to the life cycle of AI applications, see [90].

dataset in which Group A contains 30% positive ground-truth labels and Group B contains 60%. If the model is to reach a low value for a fairness metric that measures the discrepancy in the distribution of positive labels across groups, its outputs must deviate from ground-truth. In addition to sacrificing accuracy, this could also result in unbalanced error rates. Thus, depending on the nature of the data, fairness metrics may be mutually exclusive [7].

2.2.4 Measures that "correct"

Another direction of research is striving to develop technical measures which can *"correct"* or mitigate detected bias. To this end, approaches along the different development stages of ML models are being explored [35]. The underlying technical issue, especially when facing systemic or historical bias, is to train a model by inferring correlations in data that performs well on a given task—but simultaneously preventing learning of certain undesirable patterns that are present in the data. An important starting point for addressing this apparent contradiction is the data itself. A basal pre-processing method that has been proposed is "Fairness through Unawareness," meaning that those (protected) attributes are removed from the data set for which correlation with model output values is to be avoided [63], or whose inclusion could be perceived as "procedurally unfair" [44]. However, this method alone is not recognized as sufficiently effective as correlated "proxies" might still be contained in the data [61], and many mitigation methods actively incorporate the protected attributes to factor out bias [63]. Further examples of pre-processing methods range from targeted reweighing, duplication, or deletion of datapoints to modifying the ground-truth [57] or creating an entirely new (synthetic) data representation [104]. The latter are usually based on an optimization in which the original datapoints are represented as a debiased combination of prototypes. While these methods primarily aim at equalizing ground-truth values among different groups, some optimization approaches for generating data representations also include aspects of individual fairness [64]. Furthermore, to mitigate representation or sampling bias, over-sampling measures to counteract class imbalance [16] are being researched [15]. In addition to algorithmic methods, documentation guidelines have been developed to support adherence to good standards, e.g., in data selection [36].

While the data centrally influences the model results and pre-processing methods offer the advantage that they can typically be selected independently of the model to be trained, research is also being conducted on so-called in- and post-processing measures. In-processing measures are those that intervene in modeling. This can be realized, for example, by supplementing the loss function with a regularization term that reduces the correlation between model output and certain attributes [59], or using optimization constraints to align certain error rates among different groups [103]. Another in-processing approach, which affects the entire model architecture, is to include an adversarial network in the training that attempts to draw an inference about protected attributes from the model outputs [105]. The model and its adversary are trained simultaneously, where the optimization goal of the original model is to keep the performance of the adversary as low as possible. In contrast, post-processing refers to those measures that are applied to fully trained models. For example, corresponding methods comprise calibration of outputs [78] and targeted threshold setting (with thresholds per group, if applicable) to equalize error rates [49]. Many of the in- and post-processing measures are researched primarily for classification tasks, and the methods developed are typically tailored to one type of model for technical reasons.

2.2.5 Outlook

In the fairness research field, a variety of approaches have been developed to better understand and control bias. Beyond these achievements, still, open research questions remain unaddressed from a technical and interdisciplinary perspective. Regarding the first, many metrics, in addition to the methods that work toward their fulfillment, are applicable to specific tasks only and impose strong assumptions. Here, one challenge is to adapt specific measures from one use case to another. Regarding the latter, a central issue is that different fairness metrics pursue different target states for an ML model (see "Measures to detect"), therefore a choice must be made when assessing fairness in practice. Furthermore, the concrete configuration of specific metrics, for example, how a meaningful similarity metric for assessing "individual fairness" should be defined, remains unresolved. The question of which fairness metrics and measures are desirable or useful in practice must be addressed in interdisciplinary discussion. A concrete example is provided in Sect. 3.3.

2.3 Management science

Management literature [4, 21, 38, 77, 88] divides fairness into four types: distributive, procedural, interpersonal, and informational fairness.

Distributive fairness refers to the evaluation of the outcome of an allocation decision [34]. Equity is inherent to distributive fairness [79]. Hence, to achieve distributive fairness, participants must be convinced that the expected value created by the organization is proportionate to their contributions [4]. Procedural fairness refers to the process of decision right allocation (i.e., how do the parties arrive at a decision outcome? [62]). To achieve procedural fairness,

Table 1 Procedural and distributive fairness measurement scales [22, 79]

Fairness Type	Components	Description
Procedural (Rules taken from [67, 92])	Process Control	Procedures provide opportunities for voice
	Decision Control	Procedures provide influence over outcomes
	Consistency	Procedures are consistent across persons and time
	Bias Suppression	Procedures are neutral and unbiased
	Accuracy	Procedures are based on accurate information
	Correctability	Procedures offer opportunities for appeals of outcomes
	Representativeness	Procedures consider concerns of subgroups
	Ethicality	Procedures uphold standards of morality
Distributive (Rules taken from [1, 66])	Equity	Outcomes are allocated according to contributions
	Equality	Outcomes are allocated equally
	Need	Outcomes are allocated according to need

a fair assignment of decision rights is required [4, 70]. The decision rights must ensure fair procedures and processes for future decisions that influence value creation [4, 70].

Interpersonal and informational fairness refer to interactional justice, which is defined by the interpersonal treatment that people experience in decision-making processes [10]. Interpersonal fairness reflects the degree of respect and integrity shown by authority figures in the execution of processes. Informational fairness is specified by the level of truthfulness and justification during the processes [20, 43].

To ensure a differentiated assessment of AI systems from a socioeconomic perspective, these four dimensions should be included in the evaluation of fairness. In particular, procedural and distributive fairness should be emphasized, as the credit scoring assessment concentrates primarily on the credit-granting decision process. Table 1 provides an overview of fairness measurement scales in management science based on Colquitt and Rodell's [22] and Poppo and Zhou's [79] work.

Within the data science perspective, inequality indices such as the generalized entropy (GE) index or the Gini coefficient are widely accepted [25]. Both measures aim to evaluate income inequality from an economic perspective. However, they differ in their meaning, with the GE index providing more detailed insights by collecting information on the impact of inequality across different income spectrums. The GE index is also used in an interdisciplinary context. For example, in computer science, it is used to measure redundancy in data, which is used to assess the disparity within the data. In addition to the economic level, approaches to fairness measurement also exist at the corporate level.

However, the operationalizability of the fairness types described, especially procedural and distributive fairness, remains unclear. This paper aims to address this issue. A typical instrument in management practice is price discrimination, aiming to exploit the market potential. Banks' business model is to spread risks and price risks according to their default risk in order to achieve the best possible return on investment. For example, banks price the default risk of loans variously and derive differentiated prices. Here, procedural fairness is crucial in the overall fairness assessment, since procedurally unfair price settings lead to higher overall price unfairness [32]. Ferguson, Ellen, and Bearden highlight that random pricing is assessed to be more unfair than possible cost-plus pricing (price is the sum of product costs and a profit margin) within the procedural fairness assessment [32]. Furthermore, they provide evidence that procedural and distributive fairness positively interact and thus, if implemented accordingly, can maximize the overall fairness. As described in Table 1, the presented six procedural components and three distributional components should be considered in the pricing process to achieve strong overall fairness. From an organizational perspective, financial institutions should ensure that their credit ratings are neutral and unbiased based on accurate information [30]. Regarding erroneous data, customers should be able to review and correct the data if necessary. Pricing should be consistent to avoid the impression of random pricing. Therefore, people with the same attribute characteristics should always receive the exact credit pricing. Furthermore, the possible use of algorithms should not disadvantage certain marginalized groups. Moreover, to maximize the overall fairness, banks should include the distributive components in their fairness assessment. Haws and Bearden emphasize that customers assess high prices with unfairness and vice versa [50]. Thus, Ferguson, Ellen, and Bearden argue that distributional fairness is given when customers receive an advantageous price [32]. Consequently, when pricing loans, banks should always adhere to market conditions to achieve a maximum overall fairness.

2.4 Summary

Fairness can be generally considered as the absence of unjustified unequal treatment. This broad understanding takes on different specific connotations that require consideration when evaluating AI systems. In our interdisciplinary overview, two main aspects of fairness were highlighted. Distributive fairness is one of these. It concerns how automated predictions impacting the access of individuals to products, services, benefits, and other opportunities are allocated. This algorithmic outcome can be analyzed through statistical tools to detect an eventual unequal distribution of certain predictions and to assess whether this is justified by the individual features of group members, or whether this is due to biases or other factors. Considering procedural fairness is also fundamental for the evaluation of AI systems. Therefore, it should be considered how a decision is reached for different stakeholders' groups and how members of these groups are treated in the different stages of the product life cycle.

3 Evaluating fairness

3.1 The use case: creditworthiness assessment scoring for small personal loans

Based on the empirical evidence of perpetuation of preexisting discriminatory bias and of the discrimination risks for specific demographic groups, recent literature on credit scoring algorithms investigated gender-related [96] and race-related [65] fairness issues of ML systems, looking for suitable tools to detect and correct discriminatory biases and unfair prediction outcomes. Here we consider the case of small personal loans. These are small volume credits to finance, for instance, the purchase of a vehicle or pieces of furniture, or to cover the costs of expenses such a wedding or a holiday. They typically range from 1.000 EUR to 80.000 EUR—in some cases they can be up to 100.000€,[7] which are granted without a comprehensive check-up by the credit institute.

During the credit application process, as a preliminary step, a bank requests customer information, such as address, income, employment status, and living situation, which it feeds into its own (simple) credit scoring algorithm.[8] As opposed to the application process for higher volume credits, extensive information on the overall assets and wealth of the applicant is not required. Regarding particularly small lending, account statements might not even be necessary.[9] In some cases, the authorization to conduct a solvency check through a credit check agency might be requested. If so, the credit check agency will process additional information concerning, among other things, the credit history of the applicant and other personal information to produce a credit rating.[10] Finally, based on the creditworthiness assessment, a bank clerk decides whether the small personal loan is granted. In some instances, the rates might be raised in order to compensate the credit institutes for the potential illiquidity of individual customers.[11]

In the European framework, guidelines to improve institutions' practices in relation to the use of automated models for credit-granting purposes have been produced [2, 5, 30]. In the report *Guidelines on loan origination and monitoring*, the European Banking Authority (EBA) recommends that credit institutions should "understand the quality of data and inputs to the model and detect and prevent bias in the credit decision-making process, ensuring that appropriate

[7] Different banks set different limits to the maximum amount of a small personal loan. For the purpose of this paper, we considered the five German top banks for number of customers (https://www.mobilebanking.de/magazin/banken-ranking-die-groessten-banken-deutschlands.html) and found the following ranges: Sparkasse, 1.000–80.000 EUR (https://www.skpk.de/kredit/privatkredit.html); Volksbank, 1.000–50.000 EUR (https://www.vr.de/privatkunden/unsere-produkte/kredite/privatkredit.html); ING, 5.000–75.000 EUR (https://www.ing.de/kredit/ratenkredit/); Postbank 3.000–100.000 EUR (https://www.postbank.de/privatkunden/produkte/kredite/privatkredit-direkt.html); Deutsche Bank 1.000–80.000 EUR (https://www.deutsche-bank.de/opra4x/public/pfb/privatkredit/#/page-2-0). Using online platforms to compare different lenders such as check24 (https://www.check24.de/) or verivox (https://www.verivox.de/kredit/kleinkredit/), we found out that no credit lenders in Germany offers more than 100.000 EUR.

[8] In the report *Guidelines on loan origination and monitoring*, the European Banking Authority lists a set of customer information that are admissible and, where applicable, recommended to collect for credit institutions [30]. These are: purpose of the loan, when relevant to the type of product; employment; source of repayment capacity; composition of a household and dependents; financial commitments and expenses for their servicing; regular expenses; collateral (for secured lending); other risk mitigants, such as guarantees, when available (85.a — 85.h). A more extensive list of possible private customer information that might be asked in different loan application scenarios is included in the Annex 2 of the same publication. The collection of all the listed information point is not mandatory.

[9] For instance, the Commerzbank don't require these for credit lending up to 15.000 EUR (https://www.commerzbank.de/kredit-finanzierung/produkte/ratenkredite/kleinkredit/).

[10] On its website, the German credit rating agency Schufa lists the following applicant features as impact factors for credit score: number and dates of relocations; number of credit cards and opening date of the credit cards' accounts; number and dates of online purchases on account; payment defaults; existing loans; number and opening dates of checking accounts; existing mortgage loans (https://www.schufa.de/scorechecktools/pt-einflussfaktoren.html).

[11] For instance, in the credit application form of the Deutsche Bank is stated that the interest rate depends on the credit rating of the credit check agency (https://www.deutsche-bank.de/opra4x/public/pfb/privatkredit/#/page-2-0).

safeguards are in place to provide confidentiality, integrity and availability of information and systems have in place" (53.e, see also 54.a and 55.a), take "measures to ensure the traceability, auditability, and robustness and resilience of the inputs and outputs" (54.b, see also 53.c), and have in place "internal policies and procedures ensuring that the quality of the model output is regularly assessed, using measures appropriate to the model's use, including backtesting the performance of the model" (54.c, see also 53.f and 55.b) [30]. In the white paper *Big data and artificial intelligence*, the German Federal Financial Supervisory Authority (BaFin) also recommends principles for the use of algorithms in decision-making processes. These include: preventing bias; ruling out types of differentiation that are prohibited by law; compliance with data protection requirements; ensuring accurate, robust and reproducible results; producing documentation to ensure clarity for both internal and external parties; using relevant data for calibration and validation purposes; putting the human in the loop; having ongoing validation, overall evaluation and appropriate adjustments [5].

The present work follows on these recommendations and contributes to the regulatory discussion by highlighting use case-specific operationalizable requirements that address the issues emphasized by European financial institutions. We focus specifically on small volume credit for two main reasons. First, the pool of potential applicants is significantly larger than the one for higher volume credits such as mortgage lending. While high volume credits usually require the borrower to pledge one or more assets as a collateral and to be able to make a down payment to cover a portion of the total purchase price of an expensive good, these conditions do not apply to small personal loans, making these also accessible for citizens without consistent savings or other assets. Since the overall personal wealth should not influence the decision outcome in small personal loans, this makes it a particularly interesting scenario to evaluate potential discrimination of individual belonging to disadvantaged groups that are not eligible for higher volume credits, but could be granted a small personal loan. Second, the amount of applicant information processed for creditworthiness assessment is significantly lower than in the case of higher volume credits, allowing a clearer analysis of the relevant parameters and their interplay.

3.2 Preliminary ethical analysis

Regarding credit access, structural injustice severely afflicts women and demographic minorities. Although in contemporary liberal democracies explicitly preventing credit access based on gender, race, disability or religion is illegal because it represents a violation of basic human rights,[12] for structural reasons, many individuals belonging to disadvantaged groups still struggle to access credit. The gender pay gap is a concrete example: a woman working full-time in the same position as a male colleague might earn less [42], and be less creditworthy from the bank's perspective.

Since ethics is supposed to influence shaping a fairer society, the definition of minimal ethical requirements for a fair ML system in a specific application field should consider whether and how technologies can help prevent unfairness, and aim at assisting disadvantaged groups. Considering our fairness understanding as the absence of unjustified unequal treatment of individuals or groups, the first question leading to a definition of minimal ethical requirements is the following: Are there individuals belonging to certain groups that are not granted loans although they share the same relevant parameters with other successful applicants belonging to other groups? This should not be mistaken with the claim that every group should have the same share of members being granted a loan (group parity), since it is not in the interest of the person applying for a loan to be granted one if they are unable to pay it back. This would also be unethical because it would further compromise the financial stability and creditworthiness of the person, causing legal trouble and moral harm. To answer this question, fairness metrics can be a useful tool to detect disparities among groups.

3.2.1 Which metric(s) to choose?

The choice of one metric in particular is not value-neutral. Several factors should be considered when investigating fairness metrics. Among others, there is an ethical multi-stakeholder consideration to be performed [47]. Certain metrics can better accommodate the businesses' needs and goals, while others will better safeguard the rights of those being ranked or scored by a software. For instance, while evaluating the accuracy of a credit scoring system among protected groups, financial institutes will be primarily interested in optimizing (to a minimal rate) the number of loans granted to people who will not repay the debt (false positive rate) for all groups. However, it is in the interest of solvent credit applicants to optimize to a minimal rate the number of credit applicants to whom credit is denied although they could have repaid the debt (false negative rate) for all groups. Therefore, if asked to choose a metric to evaluate fairness, the former could opt for a predictive equality fairness metric, measuring the probability of a subject in the negative class to have a positive predictive value [96]. However, someone representing the latter could rather choose the equal opportunity

[12] See, e.g., the EU Charter of Fundamental Right, §21, non-discrimination, and §23, equality between women and men.

metric, the probability of a subject in a positive class to have a negative predictive value [96]. Therefore, the choice of a specific metric is not value-neutral since it could better serve the interest of certain stakeholder groups. Among other things, the role of AI ethics and AI regulation should be to prevent a minority of advantaged stakeholders from receiving the majority of advantages at the expense of those who are less advantaged. In this specific use case, this goal should be reached by considering the equal treatment of all applicants as the actual chance of getting a loan when the financial requirements are met irrespectively of the applicant's demographic group.

In their paper, "Why fairness cannot be automated," Sandra Wachter, Brent Mittelstadt, and Chris Russel, highlight the "conditional demographic parity" as a standard baseline statistical measurement that aligns with the European Court of Justice "gold standard" for assessment of prima facie discrimination. Wachter, Mittelstadt, and Russel argue that, if adopted as an evidential standard, conditional demographic parity will help answer two key questions concerning fairness in automated systems:

1. Across the entire affected population, which protected groups could I compare to identify potential discrimination?
2. How do these protected groups compare to one another in terms of disparity of outcomes? [98]

Here we follow their general proposal and suggest to use this specific metric as an evaluation tool for the specific case of creditworthiness assessment for small personal loans. In Sect. 3.3., we show how this metric can be used to evaluate the algorithmic outcome in our application field.

3.2.2 De-biasing is not enough

A fairness metric alone is insufficient to address fairness issues. If it becomes clear that group inequality is motivated by a structural reason, both the algorithmic outcome and the parameters and steps behind the decision process, this process requires questioning [45]. Different examples of checklists addressing procedural aspects of AI systems' design, development and application can be found in recent reports, white papers, and standard proposals. The Assessment List for Trustworthy AI (ALTAI) for self-assessment by the High Level Expert Group for Artificial Intelligence of the EU includes "mechanisms to inform users about the purpose, criteria and limitations of the decisions generated by the AI system," "educational and awareness initiatives," "a mechanism that allows for the flagging of issues related to bias discrimination or poor performance of the AI system," and an assessment taking "the impact of the AI system on the potential end-users and/or subjects into account" [53]. The VDE SPEC 90012 (2022) "VCIO-based description of systems for AI trustworthiness characterization" recommends to audit working and supply chain conditions, data processing procedures, ecological sustainability, adequacy of the systems outcome's explanation to inform the affected persons [94]. NIST Special Publication "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" recommends considering "human factors, including societal and historic biases within individuals and organizations, participatory approaches such as human centered design, and human-in-the-loop practices" when addressing bias in AI [86]. Madaio et al. designed a check-list intended to guide the design of fair AI systems including "solicit input on definitions and potential fairness-related harms from different perspectives," "undertake user testing with diverse stakeholders," and "establish processes for deciding whether unanticipated uses or applications should be prohibited" among the to-dos [69].

In the credit lending scenario, certain applicants' groups have been structurally disadvantaged in their history of access to credit and could still experience obstacles in successfully participating in the application process. Consequently, it should be ensured that only parameters which are relevant to assess the applicant's ability to repay the loan are processed — e.g., bank statements or monthly income — and that parameters that may lead to direct or indirect discrimination and bias perpetuation — e.g., postal code, gender, or nationality — are excluded. On this point, we follow the privacy preserving principle of "data minimization" as expressed in Art. 5.1.(c) and 25.1. GDPR. Assuming that there are different computing methods to optimize the algorithmic outcome in order to avoid unjustified unequal treatment of credit applicants, those methods processing less data should be preferred over those requiring a larger dataset containing more information on additional applicant's attributes.

Moreover, to empower credit applicants from all groups, the decision process should be made explainable so that rejected applicants can understand why they were unsuccessful. This would prevent applicants from facing black box decisions that cannot be contested, therefore diminishing the bargaining power unbalance between applicants and credit institutes. The decision process can be questioned through "counterfactual" explanations stating how the world would have to be different for a desirable outcome to occur [73, 97]. As remarked by Wachter et al., in certain cases, knowing what is "the smallest change to the world that can be made to obtain a desirable outcome," is crucial for the discussion of counterfactuals and can help understand the logic of certain decisions [97] (p. 845). In our specific case, to provide applicants with this knowledge, the decisive parameters or parameter combination (e.g., insufficient income and/or being unemployed) that led to credit denial should be made transparent, and the counterfactual explanation should explain how these parameters should have differed in order for the credit application to be approved. This would provide the applicant the opportunity to contest the algorithmic decision, to provide supplementary information relevant to support their application or, eventually, to successfully reapply for a smaller loan.

This transparency requirement also relates to the issue concerning processing data that could result in direct or indirect discrimination since credit scoring might be performed based on data belonging to credit check agencies which are not made available for private citizens.

3.3 The "conditional demographic parity" metric

In order to conduct the statistical calculation concerning the potential existence of indirect discrimination, in the interdisciplinary literature, the so-called "conditional demographic parity" metric has been proposed [98] (p. 54 ff.). This metric mirrors a statistical approach, which can be applied to examine potential discrimination in the context of the European anti-discrimination laws [98]. This technique should not be confused with the second step, meaning the question of justification of a particular disadvantage. Instead, it only concerns the first step, which deals with the question whether a particular disadvantage within the meaning of the definition of indirect discrimination is present. From a computer science perspective, numerous approaches for measuring fairness have been presented which fit in the fundamental conceptions of "individual fairness" or "group fairness." Individual fairness relates to the idea of comparing two persons, which can be classified as similar apart from the sensitive attribute; it is infringed if these two persons are not treated correspondingly [48] (p. 1175). However, group fairness statistically compares two groups of persons [48] (p. 1175). A case of direct discrimination constitutes a breach of individual fairness; a case of indirect discrimination contravenes group fairness [48] (p. 1175).

Group fairness metrics generally compare statistical quantities regarding defined groups in a dataset, e.g., the set of data samples with annual income over 50.000€ and the group with income less or equal to 50.000€. "Statistical parity" metrics [96] in addition to "demographic parity" [98], which are equivalent under certain circumstances,[13] constitute basic representatives of group fairness metrics that compare the (distribution of) outputs. These metrics are best applied to scenarios where there is a commonly preferred output from the perspective of the affected individuals (e.g., "credit granted" in case of credit scoring or "applicant accepted" in case of automated processing of job or university applications), and they compare how this output is distributed. "Statistical parity" serves to compare the proportions to which different groups, defined by a sensitive/protected attribute, are assigned a (preferred) output. Let us illustrate this on the example of credit scoring: Denote c = 1 the prediction/outcome that a credit is granted (c = 0 if the credit is not granted), and S the sensitive attribute sex with S = m denoting a male applicant and S = f a female applicant (for now, we reduce this example to the binary case both for the output and the sensitive attribute). The "statistical parity" metric (with respect to the groups of female and male applicants) is defined as the difference between the proportion to which male applicants are granted a loan and the proportion to which female applicants are granted a loan. As a formula:

$$\frac{|applicants\ with\ c = 1\ and\ S = m|}{|applicants\ with\ S = m|} - \frac{|applicants\ with\ c = 1\ and\ S = f|}{|applicants\ with\ S = f|} \tag{1}$$

For instance, if 80% of male applicants and 60% of female applicants are granted a loan, the statistical (dis-)parity is |80%–60%|= 20%.

Other than contrasting the protected group (here meaning the people falling under the sensitive feature in question and being examined in the specific case) with the non-protected group (what the "statistical parity" metric does), one could also consider solely the protected group and compare group proportions along preferred and non-preferred outputs. "Demographic parity" as described by [98] follows the latter approach. This metric compares to what proportion the protected group is represented among those who received the preferred output and among those who received the non-preferred output. According to the description in [98], demographic disparity exists if a protected group is to a larger extent represented among those with non-preferred output than among those with preferred output.

Returning to our credit scoring example, "demographic parity" here measures the difference between the proportion of females within the group of persons whom a credit is granted, and the proportion of females within the group of persons whom a credit is not granted. As a formula, demographic disparity exists if

[13] Remark: Under the assumption that both the output and the sensitive attribute are binary, one can show that the concepts of "statistical parity" and "demographic parity" are equivalent, meaning that statistical parity is satisfied if and only if equality in demographic parity holds (for a proof under the above assumptions, see annex 1 in [98]). However, each sensitive attribute can be simplified to the binary case by considering the protected group (e.g., "female applicants") on the one hand and "all others" on the other (i.e., the male and all third genders are thrown together to form the second group). Although not comparing two specific groups anymore but the protected group with "all others," for this simplification, "statistical parity" and "demographic parity" are equivalent though.

$$\frac{|applicants\, with\, c = 0\, and\, S = f|}{|applicants\, with\, c = 0|} > \frac{|applicants\, with\, c = 1\, and\, S = f|}{|applicants\, with\, c = 1|} \quad (2)$$

For instance, if 36% of the applicants **being granted** a loan are female, but 51% of those **not being granted** a loan are female, demographic disparity exists with a discrepancy of |36%–51%|= 15%.

Both the "statistical parity" and "demographic parity" metrics provide a first indication of the "particular disadvantage" within the definition of indirect discrimination presented above. Moreover, they can be easily calculated independent of the (potentially biased) ground-truth data.

However, groups defined by only one sensitive attribute (e.g., sex) can be large. Thus, the metrics presented might be coarse and unable to capture potential disparity within a group. For example, within the group of females, single applicants from the countryside might have a far lower approval rate than the average female applicant, while married applicants from the city might be granted credits almost as often as men.

Considerations such as the previous, which aim to understand given statistical or demographic disparity more deeply (e.g., by finding correlated attributes to explain the existing bias), should be informed by statistical evidence. One approach to provide more granular information on potential biases is to include (a set of) additional attributes A, which do not necessarily need to be sensitive/protected. In particular, the statistical quantities which are subject to the metrics presented can be calculated on subgroups which are characterized by attributes A (additional to the sensitive attribute). This enables a comparison of (more homogeneous) subgroups. Following this approach, an extension of the demographic parity metric has been presented[14]:

"Conditional demographic parity" [98] is defined in the same way as "demographic parity" but restricted to a data subset characterized by attributes A. In other words, "conditional demographic parity" is violated if, for a (set of) attributes A, the protected group is to a larger extent represented among those with non-preferred output and attributes A than among those with preferred output and attributes A.

Returning to the credit scoring example, let $A =$ ("annual income" $<$ 50.000€) the attribute characterizing a person as having an annual income lower than 50.000€. For this configuration of A, c, and S, "conditional demographic parity" compares the proportion to which successful applicants satisfying A are female with the proportion to which unsuccessful applicants satisfying A are female. As a formula:

$$\frac{|applicants\, with\, c = 0, S = f\, and\, A|}{|applicants\, with\, c = 0\, and\, A|} > \frac{|applicants\, with\, c = 1, S = f\, and\, A|}{|applicants\, with\, c = 1\, and\, A|} \quad (3)$$

Let us assume that female loan applicants have an income under 50.000€ statistically more often than non-female applicants. Using fictitious numbers, let 90%[15] of the female applicants satisfy A but only 70% of the non-female applicants. "Conditional demographic parity" can now help us better understand whether this gender pay gap provides an explanation why female applicants are being granted a loan less frequently, or whether there is additional discrimination not resulting from unequally distributed income. Using fictitious numbers again, let us assume that 60% of the applicants who are **being granted** a loan and satisfy A are female and 62% of the **unsuccessful** applicants satisfying A are female. Thus, analyzing small income only (where female applicants represent a higher percentage), female and non-female applicants are not treated significantly differently. Complementary, one could examine whether there is unequal treatment in the high-income group. Let $B =$ ("annual income" $> =$ 50.000€). Following our example, 30% of the non-female applicants fall in category B but only 10% of the female applicants. Let us now apply conditional demographic parity with respect to "high-income." Using fictitious numbers, let 27% of the applicants who are **being granted** a loan and satisfy B be female and 25% of the **unsuccessful** applicants satisfying B be female. Again, the representation of females in the high-income group is fairly equal among successful and unsuccessful applicants. Overall, analyzing the subsets of applicants with small income and high-income separately, female and non-female applicants seem to be treated equally among these groups. Thus, our fictitious numbers indicate that the bias in overall acceptance rates results from female applicants being to a larger extent represented in the small income group than non-female applicants.

Regarding the examination under the European non-discrimination laws, the following remarks can be made with regard to the example above: In principle, trying to avoid the

[14] There is also an extension of the statistical parity metric called "conditional (non-)discrimination" [58], also referred to as "conditional statistical parity" [96], which is defined in the same way as "statistical parity" but is supposed to be calculated on a subset of the data characterized by attributes A. Due to the fact that this paper focuses on conditional demographic parity, this metric is not supposed to be further discussed in the following.

[15] These numbers are inspired by Einkommen von Frauen und Männern in Deutschland 2021 | Statista (https://de.statista.com/statistik/daten/studie/290399/umfrage/umfrage-in-deutschland-zum-einkommen-von-frauen-und-maennern/). According to this source, in 2021 in Germany, 91,1% of the women had a monthly net income between 0 and 2.500€ while this holds for 72,1% of the men.

non-repayment of a credit constitutes a legitimate aim of the banking institute [85] (p. 310). As to this, the financial capability of the applicant in question is decisive. In this respect, it is conceivable to consider the applicants' income — making it a suitable means to foster the legitimate aim. With this in mind one can see that when considering the criterion "income," the percentage of women among the unsuccessful and the successful applicants is more or less equal within the two construed income-(sub-)groups (yearly salary over and under 50.000€). Considering only the attribute (sex), the result might be that the percentage of unsuccessful female applicants is higher than the percentage of successful female applicants (demographic disparity). Comparing the two results, it is possible to make the assumption that income was crucial for the decision whether a credit is granted or not. If one considers the orientation toward the income as an indicator for the financial capability to be necessary and appropriate, the particular disadvantage might be justified.

While generally achieving "conditional demographic parity" seems unlikely given the variety of choice for A, calculating the "conditional demographic parity" metric for different configurations of A can still provide valuable evidence to assist in detecting the relevant "particular disadvantage" (pertaining to the first step in determining an indirect discrimination). Especially, being more fine-granular than the non-conditional metrics which measure bias only in the model's overall results, their extensions can be used to explain bias by analyzing additional attributes which might provide further relevant information.

3.4 Ethical minimal requirements

We claim that the following minimal requirements must be standardized to address discrimination and procedural fairness issues concerning the application of ML systems in credit scoring for small personal loans:

(1) Regular check of the algorithmic outcome through a fairness metric. We follow Wachter, Mittelstadt, and Russel in suggesting that the conditional demographic parity fairness metric should be used to detect unfair outcome [98]. For our simple case study, the conditionals will be income and employment status. If the bank requires an external credit rating, then other parameters that influence the rating such as past loan defaults or number of credit cards must also be considered. However, in our case, comprehensive information on living costs and total wealth and assets is not required. These can therefore not be considered as conditionals.

(2) Ensure the relevance of the chosen indicators. Parameters which are not directly relevant to assess the applicant's ability to repay the loan shall not be processed. These include attributes, such as postal code, nationality, marital status, gender, disability, age (within the fixed age limits to apply for a loan), and race.[16] Some of these, such as gender, race, and disability, are characteristics protected by national and international anti-discrimination acts such as the European anti-discrimination law or the German Equal Treatment Act (AGG). Others, even if not protected by anti-discrimination laws, might facilitate the deduction of one or more protected attributes.

(3) Provide transparency for credit applicants and other actors involved. The following shall be made transparent for the applicants:

o Which data are processed (no personal data is processed without informed consent of the applicant).

p Why an application is eventually rejected and what applicant features should be improved to obtain the loan. Therefore, the algorithmic decision must be counterfactually explainable, e.g., if the applicant had a higher income or if she/he was not unemployed, she/he would have received the loan.

This does not mean disclosing the entire computing process—which might be protected by trade secrets—but guaranteeing transparency regarding the criteria applicants need to fulfill.

4 Standardizing minimal ethical requirements to evaluate fairness

Standardization can connect different perspectives of "fairness" and can establish a universal understanding in the context of AI. It can erase trade barriers and support interoperability as well as foster the trust in a system or application. Within the realm of standardization, existing definitions for fairness are rather generic and currently not tailored for AI systems and applications; however, this may change soon with the development of new AI-dedicated standards.

Several documents are pushing in that direction:

- ISO/IEC TR 24028:2020, *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence* as it also lists fairness as an essential part for ensuring trustworthiness in AI [55].
- ISO/IEC TR 24027:2021, *Information technology — Artificial intelligence (AI) — Bias in AI systems*

[16] For instance, considering the "German Credit Dataset" (https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29) used by Verma and Rubin (2018) for their fairness metrics' analysis, it is possible to flag the following parameters according to this requirement: personal status and sex (attribute 9); age in years (attribute 13); foreign worker (attribute 20).

  and AI aided decision-making addresses bias in relation to AI systems [54].
- ISO/IEC TR 24368:2022, *Information technology — Artificial intelligence — Overview of ethical and societal concerns* aims to provide an overview of AI ethical and societal concerns, as well as International Standards that address issues arising from those concerns [56].

In addition, there are many other AI-specific projects published or under development within the ISO and the IEC on the topics of ML, AI system life cycle processes, functional safety, quality evaluation guidelines, explainability, data life cycle frameworks, concepts and terminology, risk management, bias in AI systems, AI aided decision-making, robustness assessment of neural networks, an overview of ethical and societal concerns, process management framework for big data analytics, and other topics. The focus of these projects is to develop a framework of requirements for the development and operation of safe, robust, reliable, explainable, and trustworthy AI systems and applications. Following the establishment of the general AI requirement framework, the focus may likely shift to more use case-specific standardization topics like "fairness," which is clearly needed in the standardization of AI, but cannot be generalized.

Based on our interdisciplinary analysis, the standardization of "fairness" in the context of AI with the aim to allow an assessment requires multiple relevant measurable and quantifiable parameters and/or attributes building state of the art use case-specific fairness metrics such as the above discussed conditional parity metric. Such fairness metrics can be developed and standardized with an independent consensus driven platform open to expertise from all use case-related stakeholders, including views from the perspectives of philosophy, industry, research, and legislation. This platform can be either a national standards body, ISO, IEC, or the European Standardization Organization (CEN); where the most appropriate option for this topic is the international joint committee between ISO and IEC, the ISO/IEC JTC 1/SC 42 "Artificial intelligence". To begin a standardization process for a use case-specific fairness metric, scope, outline, and justification of the proposed standardization project must be proposed to the respective standardization committee. To elevate the chances of approval, a first draft with the proposed fairness metric should also be included. The standardization process within national standards bodies, ISO, IEC, and CEN provides all participating members an equal right to vote, comment and work on a standardization project. When working internationally or in the European field, this means that all interested registered experts can work on the project; however, during mandatory voting (project proposal, drafts and finalization) each participating country (represented by delegated experts) has one vote to facilitate a fair consensus process. The outcome of this process is a recognized standard, enabling mutual understanding based on agreed requirements, thus fostering trade and the new development of quality AI products and services—either nationally, in Europe, or internationally depending on the used standardization platform.

A standard can be used for a quality assessment in order to promote a product's or service's quality, trustworthiness, and user acceptability. In the assessment process of an AI system or application, the related standardized fairness metric can be used to attest the system's or application's ability to execute fair decisions. Consequently, a fairness-related attestation based on corresponding standards (e.g., certification) can increase the user acceptability and trustworthiness of the AI system or application, which can result in increased sales figures.

5 Conclusion

Evaluating the fairness of an AI system requires analyzing an algorithmic outcome and observing the consequences of the development and application of the system on individuals and society. Regarding the applied case of creditworthiness assessment for small personal loans, we highlighted specific distributive and procedural fairness issues inherent either to the computing process or to the system's use in a real-world scenario: (1) the unjustified unequal distribution of predictive outcome; (2) the perpetuation of existing bias and discrimination practices; (3) the lack of transparency concerning the processed data and of an explanation of the algorithmic outcome for credit applicants. We addressed these issues proposing ethical minimal requirements for this specific application field: (1) regularly checking algorithmic outcome through the conditional demographic parity metric; (2) excluding from the group of processed parameters those that could lead to discriminatory outcome; (3) guaranteeing transparency about the processed data, in addition to counterfactual explainability of algorithmic decisions. Defining these minimal ethical requirements represents a starting point toward standards specifically addressing fairness issues in AI systems for creditworthiness assessments. These requirements aim to prevent unfair algorithmic outcomes, as well as unfair practices related to the use of these systems.

Funding Open Access funding enabled and organized by Projekt DEAL. Our research is funded by the Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes NRW (MWIDE NRW) in the framework of the project "Zertifizierte KI" ("Certified AI"); funding number 005–2011-0050.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Adams, J.S.: Inequity in social exchange. Adv. Exp. Soc. Psychol. **2**, 267–299 (1965)
2. Aggarwal, N.: Machine learning, big data and the regulation of consumer credit markets: the case of algorithmic credit scoring. In: Aggarwal, N., Eidenmüller, H., Enriques, L., Payne, J., van Zwieten, K. (eds.) Autonomous Systems and the Law, pp. 37–44. Beck, München (2018)
3. Ahn, Y., Lin, Y.R.: Fairsight: visual analytics for fairness in decision making. IEEE Trans. Vis. Comput. Graph. **26**, 1086–1095 (2020)
4. Ariño, A., Ring, P.S.: The role of fairness in alliance formation. Strat. Manag. J. **31**, 1054–1087 (2010)
5. BaFin, Big data and artificial intelligence: Principles for the use of algorithms in decision-making processes, https://www.bafin.de/SharedDocs/Downloads/EN/Aufsichtsrecht/dl_Prinzipienpapier_BDAI_en.html (2021). Accessed 3 March 2023.
6. Balayn, A., Gürses, S., Beyond Debiasing: Regulating AI and its Inequalities. Accessed 16 December 2022
7. Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. Nips Tutorial. 1/2 (2017)
8. Barocas, S., Hardt, M., Narayanan, A. Fairness and machine learning. https://fairmlbook.org/ (2019). Accessed 30 August 2022
9. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**, 1–15 (2019)
10. Bies, R.J., Moag, J.F.: Interactional Justice: Communication Criteria of Fairness Research on Negotiation in Organizations, 1st edn., pp. 43–55. JAI Press, Greenwich (1986)
11. Binns, R.: Fairness in machine learning: lessons from political philosophy. Proc. Mach. Learn. Res. **81**, 1–11 (2018)
12. Breck, E., Polyzotis, N., Roy, S., Whang, S., Zinkevich, M.: Data validation for machine learning. Proceedings of the 2nd SysML conference, Palo Alto, CA, USA (2019)
13. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language models are few–shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020)
14. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, 77–91 (2018)
15. Chakraborty, J., Majumder, S., Menzies, T.: Bias in machine learning software: why? how? what to do? In: Proceedings of the 29th ACM joint meeting on European software engineering Conference and Symposium on the. foundations of software engineering, 429–440 (2021)
16. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over–sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
17. Çınar, Z.M., Abdussalam Nuhu, A., Zeeshan, Q., Korhan, O., Asmael, M., Safaei, B.: Machine learning in predictive maintenance toward sustainable smart manufacturing in industry 4.0. Sustainability. **12**, 8211 (2020)
18. Coeckelbergh, M.: AI for climate: freedom, justice, and other ethical and political challenges. AI Ethics **1**, 67–72 (2021). https://doi.org/10.1007/s43681-020-00007-2
19. Cohen, G.A.: On the currency of egalitarian justice. Ethics **99**, 906–944 (1989)
20. Colquitt, J.A.: On the dimensionality of organizational justice: a construct validation of a measure. J. Appl. Psychol. **86**, 386–400 (2001)
21. Colquitt, J.A., Rodell, J.B.: Justice, trust, and trustworthiness: a longitudinal analysis integrating three theoretical perspectives. Acad. Manag. J. **54**, 1183–1206 (2011)
22. Colquitt, J.A., Rodell, J.B.: Measuring justice and fairness. In: Cropanzano, R.S., Ambrose, M.L. (eds.) The Oxford Handbook of Justice in the Workplace, pp. 187–202. OUP, Oxford (2015)
23. Cremers, A.B. et al.. Trustworthy Use of Artificial Intelligence. Priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of Artificial Intelligence, https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_Thrustworthy_AI.pdf (2019). Accessed 28 January 2022
24. Datenetikkommission der Bundesregierung. Gutachten der Datenethikkommission der Bundesregierung, https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6 (2019). Accessed 28 January 2022
25. De Maio, F.G.: Income inequality measures. J. Epidemiol. Community Health. **61**, 849–852 (2007)
26. Druzhkov, P.N., Kustikova, V.D.: A survey of deep learning methods and software tools for image classification and object detection. Pattern Recognit. Image Anal. **26**, 9–15 (2016)
27. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, 214–226 (2012)
28. Dworkin, R.: What is equality? Part 1: Equality of welfare. Philos. Publ. Aff. **2**, 185–246 (1981)
29. Eidelson, B.: Discrimination and Disrespect. OUP, Oxford (2015)
30. European Banking Authority, Final Report – Guidelines on Loan Origination and Monitoring, https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Guidelines/2020/Guidelines%20on%20loan%20origination%20and%20monitoring/884283/EBA%20GL%202020%2006%20Final%20Report%20on%20GL%20on%20loan%20origination%20and%20monitoring.pdf (2020). Accessed 02 March 2023
31. European Commission. Proposal for A regulation of the European Parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (2021). Accessed 18 March 2022

32. Ferguson, J.L., Ellen, P.S., Bearden, W.O.: Procedural and distributive fairness: Determinants of overall price fairness. J. Bus. Ethics. **121**, 217–231 (2014)
33. Fisher, E., Fuchs, C.: Reconsidering Value and Labor in the Digital Age. Palgrave Macmillan, Basingstoke (2015)
34. Folger, R., Konovsky, M.A.: Effects of procedural and distributive justice on reactions to pay raise decisions. Acad. Manag. J. **32**, 115–130 (1989)
35. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness–enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, 329–338 (2019)
36. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, I.: Datasheets for datasets. Commun. ACM. **64**, 86–92 (2021)
37. Genovesi, S., Mönig, J.M.: Acknowledging sustainability in the framework of ethical certification for AI. Sustainability. **14**, 4157 (2022)
38. Gilliland, S.W.: The perceived fairness of selection systems – an organizational justice perspective. Acad. Manag. Rev. **18**, 694–734 (1993)
39. Giovanola, B., Tiribelli, S.: Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine learning algorithms. AI & Soc, 1–15 (2022)
40. Gomez, O., Holter, S., Yuan, J., Bertini, E.: ViCE: visual counterfactual explanations for machine learning models. In: Proceedings of the 25th international conference on intelligent user interfaces, 531–535 (2020)
41. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, Cambridge (2016)
42. Gould, E., Schieder, J., Geier, K., What is the gender pay gap and is it real? Economic Policy Institute. https://files.epi.org/pdf/112962.pdf (2016). Accessed 05 September 2022
43. Greenberg, J., Cropanzano, R.: The social side of fairness: interpersonal and informational classes of organizational justice. In: Cropanzano, R. (ed.) Justice in the Workplace: Approaching Fairness in Human Resource Management, pp. 79–103. Lawrence Erlbaum Associates, Hillsdale (1993)
44. Grgić–Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The case for process fairness in learning: feature selection for fair decision making. NIPS symposium on machine learning and the law (2016)
45. Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A.: Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. AAAI. Proceedings of the AAAI conference on artificial intelligence 32. 32 (2018)
46. Gudivada, V., Apon, A., Ding, J.: Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. Int. J. Adv. Softw. **10**, 1–20 (2017)
47. Häberlein, L., Mönig, J.M., Hövel, P.: Mapping stakeholders and scoping involvement. A guide for HEFRCs, Deliverable 3.1 of the H2020–project ETHNA System. https://ethnasystem.eu/wp-content/uploads/2021/10/ETHNA_2021_d3.1-stakeholdermapping_2110011.pdf (2021) Accessed 10 May 2021
48. Hacker, P.: Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law (April 18, 2018). Common Mark. Law Rev. **55**, 1143–1186 (2018)
49. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. **29**, 2 (2016)
50. Haws, K.L., Bearden, W.O.: Dynamic pricing and consumer fairness perceptions. J. Consum. Res. **33**, 304–311 (2006)
51. Herzog, L.: Algorithmic bias and access to opportunities. In: Véliz, C. (ed.) The Oxford Handbook of Digital Ethics. Oxford Academic, Oxford (2021). https://doi.org/10.1093/oxfordhb/9780198857815.013.21
52. HLEG. HLEG on AI. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2019). Accessed 10 May 2022
53. HLEG. Assessment List for Trustworthy AI (Altai) for self–assessment. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (2020). Accessed 01 December 2022
54. ISO/IEC TR 24027:2021, Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making. https://www.iso.org/standard/77607.html (2021). Accessed 22 December 2022
55. ISO/IEC TR. 24028:2020, Information technology — Artificial intelligence — Overview of trustworthiness. https://www.iso.org/standard/77608.html (2020). Accessed 22 December 2022
56. ISO/IEC TR 24368:2022, Information technology — Artificial intelligence — Overview of ethical and societal concerns. https://www.iso.org/standard/78507.html (2022). Accessed 22 December 2022
57. Kamiran, F., Calders, T.: Data pre–processing techniques for classification without discrimination. Knowl. Inf. Syst. **33**, 1–33 (2012)
58. Kamiran, F., Žliobaitė, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl. Inf. Syst. **35**, 613–644 (2013)
59. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness–aware classifier with prejudice remover regularizer. In: Flach, P.A., Bie, T., Cristianini, N. (eds.) Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 35–50. Springer, Berlin (2012)
60. Kasirzadeh, A. Algorithmic fairness and structural injustice. Insights from Feminist Political Philosophy. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society 8, 349–356 (2022)
61. Kilbertus, N., Rojas, C.M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. Adv. Neural Inf. Process. Syst. **30**, 656–666 (2017)
62. Konovsky, M.A.: Understanding procedural justice and its impact on business organizations. J. Manag. **26**, 489–511 (2000)
63. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Adv. Neural Inf. Process. Syst. **30**, 2 (2017)
64. Lahoti, P., Gummadi, K.P., Weikum, G.: ifair: learning individually fair data representations for algorithmic decision making. 35th international conference on data engineering (icde), IEEE Publications, 1334–1345 (2019)
65. Lee, M.S.A., Floridi, L.: Algorithmic fairness in mortgage lending: from absolute conditions to relational trade–offs. Minds Mach. **31**, 165–191 (2021)
66. Leventhal, G.S.: The distribution of rewards and resources in groups and organizations. Adv. Exp. Soc. Psychol. **9**, 91–131 (1976)
67. Leventhal, G.S.: What should be done with equity theory? In: Gergen, K.J., Greenberg, M.S., Willis, R.H. (eds.) Social exchange, pp 27–56, Springer, Boston (1980)
68. Lippert-Rasmussen, K.: Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination. OUP, New York (2013)
69. Madaio, M.A., Stark, L., Wortman Vaughan, J.W., Wallach, H.: Co– designing checklists to understand organizational challenges and opportunities around fairness in AI. In: Proceedings of the 2020 CHI conference on human factors in computing systems, 1–14 (2020)
70. Mathur, P., Sarin Jain, S.S.: Not all that glitters is golden: the impact of procedural fairness perceptions on firm evaluations

and customer satisfaction with favorable outcomes. J. Bus. Res. **117**, 357–367 (2020)

71. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**, 1–35 (2021)
72. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. (2016). https://doi.org/10.1177/2053951716679679
73. Molnar, C. Interpretable machine learning. A guide for making black box models explainable, https://christophm.github.io/interpretable-ml-book/ (2022). Accessed 22 December 2022
74. Mönig, J.M.: Privatheit als Luxusgut in der Demokratie? In: Grimm, P., Zöllner, O. (eds.) Demokratie und Digitalisierung, pp. 105–114. Steiner, Stuttgart (2020)
75. Myers West, S.: Redistribution and recognition. A feminist critique of algorithmic fairness. Catalyst. (2020). https://doi.org/10.28968/cftt.v6i2.33043
76. Noble, S.U.: Algorithms of oppression: how search engines reinforce racism. New York University Press, New York (2018)
77. Pillai, R., Schriesheim, C.A., Williams, E.S.: Fairness perceptions and trust as mediators for transformational and transactional leadership: A two–sample study. J. Manag. **25**, 897–933 (1999)
78. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Adv. Neural Inf. Process. Syst. **30**, 2 (2017)
79. Poppo, L., Zhou, K.Z.: Managing contracts for fairness in buyer–supplier exchanges. Strat. Manag. J. **35**, 1508–1527 (2014)
80. Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A.B., Hecker, D., Houben, S., Mock, M., Rosenzweig, J. et al. Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (2021). Accessed 10 March 2021
81. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H.: Ensuring fairness in machine learning to advance health equity. Ann. Intern. Med. **169**, 866–872 (2018)
82. Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge MA (1999)
83. Rawls, J.: Justice as Fairness: A Restatement. Harvard University Press, Cambridge MA (2001)
84. Rohde, F. et al.: Nachhaltigkeitskriterien für künstliche Intelligenz. Entwicklung eines Kriterien– und Indikatorensets für die Nachhaltigkeitsbewertung von KI–Systemen entlang des Lebenszyklus. Schriftenr. IÖW. 220/21. https://www.ioew.de/fileadmin/user_upload/BILDER_und_Downloaddateien/Publikationen/2021/IOEW_SR_220_Nachhaltigkeitskriterien_fuer_Kuenstliche_Intelligenz.pdf (2021). Accessed 18 January 2022
85. Schürnbrand, J.: Organschaft im Recht der privaten Verbändecht. Mohr Siebeck, Tübingen (2007)
86. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P. Toward a standard for identifying and managing bias in artificial intelligence, National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf (2022). Accessed 21 December 2022
87. Sen, A.: Inequality Reexamined. Clarendon Press, Oxford (1992)
88. Skarlicki, D.P., Folger, R., Tesluk, P.: Personality as a moderator in the relationship between fairness and retaliation. Acad. Manag. J. **42**, 100–108 (1999)
89. Speicher, T., Heidari, H., Grgić–Hlača, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: A unified approach to quantifying algorithmic unfairness: measuring individual &group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2239–2248 (2018)
90. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. Equity Access Algor. Mech. Optim. **17**, 1–9 (2021)
91. Taylor, L.: What is data justice? The case for connecting digital rights and freedoms globally. Big Data Soc. (2017). https://doi.org/10.1177/2053951717736335
92. Thibaut, J.W., Walker, L.: Procedural Justice: A Psychological Analysis. Erlbaum Associates, Hillsdale (1975)
93. van Wynsberghe, A.: Sustainable AI: AI for sustainability and the sustainability of AI. AI Ethics. **1**, 213–218 (2021)
94. VDE SPEC 90012:2022 VCIO based description of systems for AI trustworthiness characterization. https://www.vde.com/resource/blob/2176686/a24b13db01773747e6b7bba4ce20ea60/vde-spec-vcio-based-description-of-systems-for-ai-trustworthiness-characterisation-data.pdf (2022). Accessed 01 December 2022
95. Véliz, C.: Privacy is Power: Why and How You Should Take Back Control of Your Data. Bantam Press, London (2020)
96. Verma, S., Rubin, J.: Fairness definitions explained. Proceedings of the international workshop on software fairness, 1–7 (2018)
97. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. SSRN J. **31**, 841–888 (2018)
98. Wachter, S., Mittelstadt, B., Russell, C.: Why fairness cannot be automated: bridging the gap between EU non–discrimination law and AI. Comput. Law Secur. Rev. (2021). https://doi.org/10.1016/j.clsr.2021.105567
99. Wahlster, W., Winterhalter, C.: Deutsche Normungsroadmap. Künstliche Intelligenz. DIN. https://www.dke.de/resource/blob/2019482/0c29125fa99ac4c897e2809c8ab343ff/nr-ki-deutsch---download-data.pdf (2020). Accessed 21 December 2022
100. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM. **39**, 86–95 (1996)
101. Young, I.M.: Justice and the Politics of Difference. Princeton University Press, Princeton (1990)
102. Yu, K.H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. Nat. Biomed. Eng. **2**, 719–731 (2018)
103. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. Proceedings of the 26th international conference on world wide web, 1171–1180 (2017)
104. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. International Conference on Machine Learning, 325–333 (2013)
105. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335–340 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.