Imbalanced data as risk factor of discriminating automated decisions: a measurement-based approach

Imbalanced data as risk factor of discriminating automated decisions

A measurement-based approach

by **Antonio Vetrò***

Abstract: Over the last two decades, the number of organizations -both in the public and private sector- which have automated decisional processes has grown notably. The phenomenon has been enabled by the availability of massive amounts of personal data and the development of software systems that use those data to optimize decisions with respect to certain optimization goals. Today, software systems are involved in a wide realm of decisions that are relevant for the lives of people and the exercise of their rights and freedoms. Illustrative examples are systems that score individuals for their possibility to pay back a debt, recommenders of the best candidates for a job or a house rent advertisement, or tools for automatic moderation of online debates. While advantages for using algorithmic decision making concern mainly scalability and economic affordability, on the other hand, several critical aspects have emerged, including systematic adverse impact for individuals belonging to minorities and disadvantaged groups. In this context, the terms data and algorithm bias have become familiar to researchers, industry leaders and policy makers, and much ink has been spelled on the concept of algorithm fairness, in order to produce more equitable results and to avoid discrimination. Our approach is different from the main corpus of research on algorithm fairness because we shift the focus from the outcomes of automated decision making systems to its inputs and processes. Instead, we lay the foundations of a risk assessment approach based on a measurable characteristic of input data, i.e. imbalance, which can lead to discriminating automated decisions. We then relate the imbalance to existing standards and risk assessment procedures. We believe that the proposed approach can be useful to a variety of stakeholders, e.g. producers and adopters of automated decision making software, policy makers, certification or audit authorities. This would allow for the assessment of the risk level of discriminations when using imbalanced data in decision making software. This assessment should prompt all the involved stakeholders to take appropriate actions to prevent adverse effects. Such discriminations, in fact, pose a significant obstacle to human rights and freedoms, as our societies increasingly rely on automated decision making. This work is intended to help mitigate this problem, and to contribute to the development of software systems that are socially sustainable and are in line with the shared values of our democratic societies.

Keywords: discrimination risk; data bias; algorithm fairness; digital policy; data ethics; data governance

Recommended citation: Antonio Vetrò, Imbalanced data as risk factor of discriminating automated decisions: A measurement-based approach 12 (2021) JIPITEC 272 para 1.

A. Background and Motivations

1 A large number of decisional processes -both in the public and private sector- are based on software elaborated recommendations, or they are completely automated, and it is likely that the phenomenon will further increase in the future [1] [2] [3]. This phenomenon has been enabled by the large availability of data and of the technical means in order to analyze them for building the predictive, classification and ranking models that are at the core of automated decision making (ADM) systems[1]. The decisions delegated or supported by these systems range from predicting debt repayment capability [4] to identifying the best candidates for a job position [5], from detecting social welfare frauds [6] to suggesting which university to attend [7], to name a few. While advantages for using ADM systems are evident and they concern mainly scalability of the operations and consequential economic efficiency, on the other hand, several critical aspects have emerged, including for instance transparency and accountability [8]. Yet another major controversy concerns discriminatory behavior, in terms of "unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category" [9]. This issue emerged from a large amount of evidence both in scientific literature [10] and journalistic investigations [11], which showed how ADM systems may systematically discriminate the weakest segments of society and exacerbate existing inequalities. Such problem would often occur as a result of imbalanced input datasets [12], which is the focus of this paper. Data imbalance is an unequal distribution of data between classes [13], which occurs when the number of data points available is very different among different classes. Causes of imbalance can be errors or limitations of the data collection design and operation, alternatively no other reason than disparities in the current reality that the data itself reproduce. Imbalance is between-class when only two classes are taken into consideration and one class is over-represented with respect to the other or multiclass when imbalances exist between multiple classes. In this paper, we focus on the more general case, i.e. multiclass imbalance.

2 Imbalanced data is known to be problematic in the machine learning domain since long [14], and is still relevant [15], especially because it can corrupt the performances of supervised learning algorithms in terms of heterogeneous accuracies across the classes of data. For example, consider an algorithm for predictive maintenance that labels a certain product component either as close to breakage or not close to breakage, and is trained with historical data from three different suppliers. A is a well-known company which sells several million pieces of that component per year. B is a company with a few thousand sales, and C is a company with less than a thousand sold components of that product. It is reasonable to expect that the algorithm trained with the historical data from the three companies could perform with higher prediction accuracy for components of supplier A and lower accuracy for products of suppliers B and C. In this fictitious example, imbalance in the input data could be the major cause for the disparate performance of the predictive algorithm, due to the fact that the model has been trained with significantly more data from Company A[2].

3 Now imagine a context where the objects of the prediction are not products but individuals, and an organization uses historical data on employees to predict which candidates' CVs most likely correspond to future successful software engineers. It comes as no surprise that the large majority of predicted candidates will be male, due to the disproportionate gender ratio in the sector. Indeed, this is not a fictitious example but rather a very blatant case of discrimination caused by data imbalance. Namely, the development of a software system by Amazon to evaluate the CVs of potential employees retrieved from the web [16]. The goal of the system was to find successful future employees, whereby the predictors were word patterns extracted from CVs of the past 10 years. According to the news agency report [16], the project started in 2014 and was stopped in 2017 because female profiles were systematically downgraded, regardless of a certain number of attempts to make technical adjustments. Here, the problem was that training data came mostly from men, since the majority of employees in the technology sector is male.

4 A similar unequal treatment due to gender imbalance in the input data has been found in a scientific

* Antonio Vetrò is a Senior Research Fellow at Nexa Center for Internet & Society and Assistant Professor at the Department of Control and Computer Engineering of Politecnico di Torino, Italy. ORCID: 0000-0003-2027-3308.

1 We follow the definition of Automated Decision Making provided by Algorithm Watch[1]: "Systems of automated decision-making (ADM) are always a combination of the following social and technological parts: i)a decision-making model ; ii) algorithms that make this model applicable in the form of software code ; iii) data sets that are entered into this software, be it for the purpose of training via Machine learning or for analysis by the software; iv) the whole of the political and economic ecosystems that ADM systems are embedded in (elements of these ecosystems include: the development of ADM systems by public authorities or commercial actors, the procurement of ADM systems, and their specific use)."

2 Due to the large difference of available data from the three companies, concurrent causes as incomplete data or different defectiveness ratios might play a minor role in explaining the divergence of performance measures.

experiment on the search engine Common Crawl [17]. The authors compared three techniques of machine learning for occupational classification with almost 400.000 collected biographies. In all cases, even without explicitly using gender indicators, the rate of correct classifications followed the existing gender imbalances of the occupational groups. In another study [18] it was reported that Facebook advertisements for employment opportunities were significantly skewed among ethnic and gender groups, leading to persistent discriminatory treatment and unequal job opportunities along the lifetime of the advertisements. This study was partially replicated by Algorithm Watch, with similar results [19]. For example, an advertisement for truck driver jobs was shown about ten times more to men than to women (4,864 times vs 386), which confirms that Facebook optimizes its target audience with past users' reactions to similar announcements, thus replicating imbalances in the data. The consequence of such a conservative mechanism is that people are deprived of opportunities based on gender, ethnic origin or other personal traits, in practice infringing Article 21 of the EU Charter of Human Rights [20]. In the United States (US), the discriminatory effect of the Facebook advertisement platform has been scrutinized by the Department of Housing and Urban Development. It sued Facebook in March 2019 for violating the Fair Housing Act, whereby the allegations were based on the evidence that housing advertisements were disproportionally targeted with respect to race, gender and other personal characteristics [21].

5 Amplifications of input data imbalance in software outputs have also been reported in general purpose search engines. A study by Kay et al. [22] on Google search results showed that in the occupational groups typically dominated by men, women were significantly under-represented, in comparison to the real gender ratio retrieved from the official employment statistics. The authors showed that such disproportion influences the perceptions of actual gender relations in occupations, with possible amplification effects on inequalities in jobs. Discrimination issues in the Google search engine are not a novel fact, as demonstrated in an empirical study from 2013 [23], which showed that advertisements for commercial products of arrest records were displayed with relevant different rates for names usually referred to non-Caucasian people than for names usually referred to Caucasian people. The opacity of the search algorithm did not allow the authors to isolate and validate the causes. However, they had confidence in reporting that the past clicks behavior of Google search users (used by Google AdSense service) might have played a major role and propagated a societal bias in the search algorithm results.

6 The negative effects illustrated in these cases could become worse or even life-altering in fields like justice or medicine, where the combined use of ADMs and historical data is rapidly increasing. The most famous case in the justice field is the investigation on COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), conducted by the non-profit organization Pro-Publica [24]. COMPAS is an algorithm used by judges to assess the probability of recidivism of defendants. The COMPAS algorithm was distorted in favor of white defendants, whereby those who were rearrested were nearly twice as likely to be misclassified as low risk than black defendants. Furthermore, the black defendants who did not get rearrested were nearly twice as likely to be misclassified as higher risk (false positive) than white defendants. The major cause was that the number of records in the dataset related to black defendants was much higher than the number of records of white defendants.

7 Regarding the medical field, a recent study [25] found evidence of ethnicity-based discrimination in a widely-used commercial system for deciding which patients should get into an intensive care program. Medical doctors applied risk scores generated by an algorithm trained on historical data about medical expenditure and the use of health services. In cases of an equivalent health status, white patients were significantly more likely than black patients to be assigned to the intensive care program. In fact, the risk score reflected more the expected cost of treatment than health conditions, with former being highly correlated to the economic wealth of the patients. In another empirical study in the medical field [26], the amount of data used for training classification algorithms in six different clinical disciplines showed that most of it came from only three geographic areas in the US, with no representation for the majority of states. Hence, automating a diagnosis on patients from states not included in the training data, would lead to wrong results and to missing health issues that are more common in the excluded geographic areas.

8 The cases summarized above, although exemplificative and not exhaustive, clearly show how imbalance in data can propagate and be reflected in the output ADM systems. When this occurs, it ceases to be a mere problem of data engineering and it becomes a socio-technical issue, particularly important when systems automate high stake decisions that can produce serious consequences for individuals. As our societies increasingly rely on ADMs, this phenomenon poses a significant challenge for the values on which our societies are based and for fundamental human rights and freedoms. The deployment of ADM systems embeds the risk to create an adverse impact for individuals belonging to minorities and marginalized groups, and to introduce or amplify distribu-

tive injustice [27] [28]. In this paper we face this important issue by focusing on the specific problem of data imbalance. We propose a measurement-based risk assessment approach, by measuring imbalance in input data, whereby we highlight the potential risk of discriminating automated decisions. We describe the theoretical foundations of the risk assessment approach, which resides in existing standards on software system quality and risk assessment. We identify three measures of imbalance and we apply them with an illustrative example.

9 The measures can be applied both before the deployment (i.e., during development) and after the deployment of ADM systems: for this reason, we believe that the proposed approach can be useful to a variety of stakeholders for assessing the risk of discriminations, including the creators or commissioners of the system, researchers, policymakers, regulators, certification or audit authorities. Assessments should prompt taking appropriate action to prevent adverse effects.

10 The paper is organized as follows: in Section B we lay the theoretical foundations of our proposal, followed in Section C by the explanation of three imbalance measures and an example of their application. In Section D we explain how this research contributes to the literature of algorithm bias and fairness, while in Section E we briefly report on the relations to the most recent policy efforts in Europe for regulating ADM systems. We conclude in Section F with a discussion of the limitations and share our roadmap for future work.

B. Data imbalance as risk factor of discriminations by automated decision making systems

11 The ADM systems described in the previous section systematically discriminate against certain groups of individuals because of imbalances in the input data. For this reason, we consider data imbalance as a risk factor and we propose measures to address it. This proposal has its foundations in software quality and risk management standards.

12 The cornerstone of the conceptual model is the series of standards ISO/IEC 25000:2014 Software Engineering — Software Product Quality Requirements and Evaluation (SQuaRE) [29]. SQuaRE includes quality modeling and measurements of software products,[3] data and software services. According to the philosophy and organization of this family of standards, quality is categorized into one or more quantifiable characteristics and sub-characteristics. For example, the standard ISO/IEC 25010:2011 formalizes the product quality model as composed of eight characteristics, which are further subdivided into sub-characteristics. Each (sub) characteristic relates to static properties of software and dynamic properties of the computer system[4]. An example of product quality characteristics is reliability, and one of its sub-characteristics is maturity[5]. Characteristics and sub-characteristics can be quantified by measurable properties of the software. For example, "failure" is a dynamic property of the software, and the number of failures is a quality measure element, which is used to measure maturity in terms of mean time between failures[6]. Reliability is quantified through the measures of its sub-characteristics.

13 Similar to product quality, data quality in ISO/IEC 25012:2008 is categorized into 15 characteristics, such as completeness, efficiency, recoverability. Each of these characteristics is quantifiable through measures of quality-related properties, defined in ISO/IEC 25024:2015. The characteristics can belong either to the "Inherent" point of view if dependent only on the data themselves, such as completeness. Alternatively, they can belong to the "System-dependent" point of view, such as recoverability. They can also belong to both, such as efficiency. Data imbalance is not a characteristic of data quality in ISO/IEC 25012:2008, however the SQuaRE standards have a structure which fits our purpose, and it defines a principle that is relevant in our context, which is the propagation principle. This principle entails that the quality of the software product, service and data would affect the quality in use and would thus have consequences for the users of a software system[7].

3 A software product is a "set of computer programs, procedures, and possibly associated documentation and data" as defined in ISO/IEC 12207:1998. In SQuaRE standards, software quality stands for software product quality.

4 A system is the "combination of interacting elements organized to achieve one or more stated purposes" (ISO/IEC 15288:2008), for example the aircraft system. It follows that a *computer system* is "a system containing one or more components and elements such as computers (hardware), associated software, and data", for example a conference registration system. An ADM system that determines eligibility for economic aid for paying drinking water bills is a software system.

5 Reliability is defined in ISO/IEC 25010:2011 as the degree to which a system, product or component performs specified functions under specified conditions for a specified period of time"; Maturity is defined in ISO/IEC 25010:2011 as "degree to which a system, product or component meets needs for reliability under normal operation".

6 Number of failures/average min-max duration.

7 In practice evaluating and improving product/service/

Figure 1 represents how this chain of effects is formalized in SQuaRE. In the realm of data quality, a simplification of this concept is the GIGO principle, which is the "garbage in, garbage out" principle. In other words, data that is outdated, inaccurate and incomplete make the output of the software unreliable.

14 We apply this principle to data imbalance because it can cause biased software outputs that negatively affect the final users, in the same way bad data quality affects the quality in use and thus has an impact on the final users. In fact, imbalanced datasets may lead to imbalanced results, which in the context of ADM means differentiation of products, information and services based on personal characteristics. In specific applications such as wages, insurance, education, working positions, tariffs, etc. such differentiations can lead to unjustified unequal treatment or discrimination. For this reason data imbalance shall be considered as a risk factor in all those ADM systems that rely on historical data and operate in relevant aspects of the lives of individuals.

15 The second conceptual pillar of the proposal is the ISO 31000:2018 standard [31] which identifies guiding principles for risk management. The proposal consists of a framework for integrating risk management into organizational contexts, and a process for managing risks at "strategic, operational, program or project levels". In the context of this discussion, data imbalance shall be explicitly taken into account within the risk management process, which we reproduce from the standard in Figure 2. Risk assessment is therefore at the center of our proposal. The process consists of risk identification, analysis and evaluation. Here, we briefly describe them and specify the relation with our approach.

- Risk identification refers to finding, recognizing and describing risks within a certain context and scope, and with respect to specific criteria defined prior to risk assessment. In our case, it is the risk associated with discriminating individuals or groups of individuals by operating ADM systems in contexts in which the impact on the lives of people would be relevant. Section A contains examples of these situations.

- Risk analysis aims to understand the characteristics of the risk and, when possible, its levels. This is the phase where measures of data imbalance are used as indicators for the risks of discrimination, due to the bias propagation effect previously described. In Section C we will introduce three measures and we will show them in action on a real dataset.

- Risk evaluation, as the last step, is a process in which the results of the analysis are taken into consideration in order to decide whether additional action is required. If affirmative, this process would then outline available risk treatment options and the need for conducting additional analyses. In addition, the process would define other types of required actions and the actors who would undertake those actions. In our case, the indicators of data imbalance should be analyzed in the context of the specific prediction/classification algorithms used, the social context, the legal requirements of the domain, etc.[8]

16 Figure 3 summarizes the approach and the connections with the international ISO/IEC standards used as reference frameworks. In the upper layer, we represent the elements of the SQuaRE series (2500n) which are most relevant for our scope. In the bottom layer, we report the main elements of the risk management process of ISO 31000. The constitutive elements of our approach - in the middle of Figure 3- are mapped to the concepts of SQuaRE and the phases of ISO 31000:

- the ADM systems constitute the context of use in terms of SQuaRE terminology, and they are specified in the context definition phase of the ISO 31000;

- the discrimination operated by ADM systems is the specific object of the risk identification process in ISO 31000 (given the context), and it decreases the quality in use of the software;

- data imbalance extends the SQuaRE data quality model because it is an inherent data characteristic: as such, i) it preserves the propagation principle and ii) it is measurable; the identified measures can be used as risk indicators in the risk analysis phase;

data quality is one mean of improving the system quality in use. It shall be clarified that in this text we refer only to the effects related to quality characteristics of the SQuaRE standards. However, the same principle can be applied to other aspects of software development that are treated in other standards, for instance the improvement of any of the lifecycle processes defined in ISO/IEC 12207:2008 and ISO/IEC 15288:2015 will determine an improvement of product quality, which in turn contributes to improving system quality in use and has a positive effect on final users (users can be direct and indirect). Although this aspect is out of our scope here, it could be relevant for techniques/procedures applied in software development processes to identify negative societal effects of software since its early development phase (for instance, in requirements definition [30]).

8 This part is not in the scope of this paper; however, we will provide some details for future work needed in this direction in Section F.

- the criteria for activating mitigation actions (e.g., thresholds for the indexes) and the mitigation actions are mapped respectively to the risk evaluation and risk treatment phases.

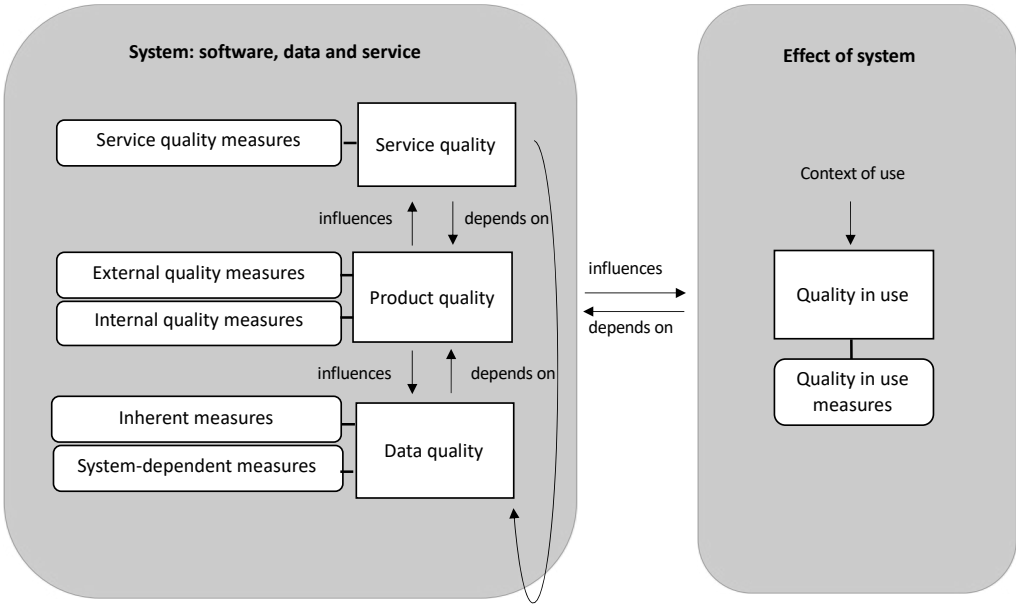


Figure 1. Quality effects in SQuaRE

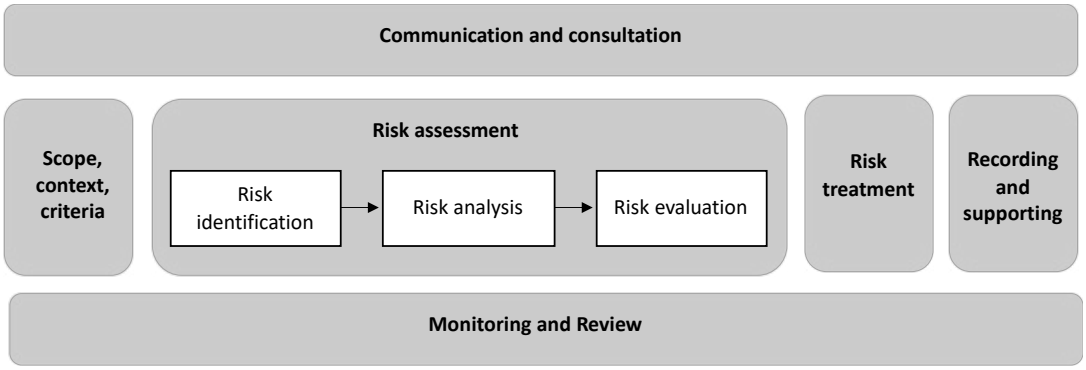


Figure 2. Risk management process in ISO 31000:2018

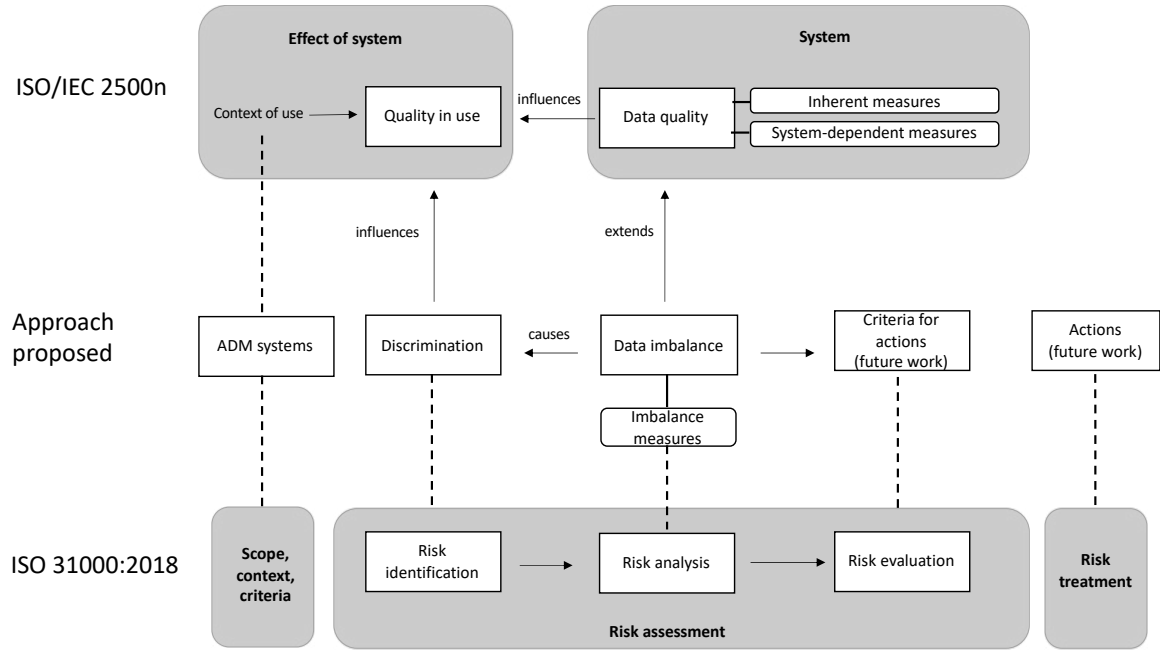


Figure 3. Approach proposed in relation to ISO standards of reference

C. Measures of imbalance for categorical data

17 According to our line of reasoning, imbalance in input data can propagate downstream to software output. As a consequence, measures of imbalance are interpreted as risk indicators.

18 Since imbalance is defined as an unequal distribution between classes [13], we focus on categorical data. In fact, most of the sensitive attributes are considered categorical data, such as gender, home town, marital status, and job. Alternatively, if they are numeric, they are either discrete and within a short range, such as family size, or they are continuous but often re-conducted to distinct categories, such as information on "age" which is often discretized into ranges such as "< 25", "25-40", "41-60". We identified three measures from the literature of social and natural sciences, where imbalance is known in terms of (lack of) heterogeneity and diversity: the identified measures are the Gini, Shannon and Simpson indexes. We provide details in Table 1, whereby we specify their formula and normalized versions, i.e. in the range 0-1, respectively in the second and third columns. In the fourth column, we provide notes for value interpretations.

19 We briefly comment on the measures here:

- Gini index measures how many different types are represented in a dataset. It has been conceived as a measure of heterogeneity, whereby it is used for different purposes in several disciplines, for example, to measure political polarization, market share in competition, ecological diversity, and racial discrimination. It increases if probabilities/frequencies become as equal as possible e.g. when different attributes would have similar representations.

- Shannon index has been proposed as a measure of diversity, and it provides information about community composition, taking the relative abundances of different classes into account. It is a concept widely employed in biology, phylogenetics, and ecology.

- Simpson index is another measure of diversity in ecology, which measures the probability that two individuals randomly selected from a sample belong to the same species or the same class/category. It has been used in ecology for measuring the diversity of living beings in a given place, as well as in social and economic sciences for measuring wealth, uniformity, and equity.

20 In order to show the three measures at work, we make an example with the widely used data from COMPAS as they are provided by the US based non-profit organization ProPublica [32]. The data contain variables used by the COMPAS algorithm in scoring criminal defendants in Broward County (Florida), along with their outcomes within two years of the decision. The original dataset includes 28 variables, eight of which are considered as protected attributes[9], such as last name, race, or marital status.

9 Protected attributes are qualities, traits or characteristics of individuals that, by law, cannot be discriminated against.

Table 1 Indexes of imbalance.

Index	Formula	Normalized formula	Notes
Gini	$G = 1 - \sum_{i=1}^{m} f_i^2$	$G_n = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} f_i^2\right)$	m is the number of classes f is the relative frequency of each class $f_i = \frac{n_i}{\sum_{i=1}^{m} n_i} f_i = \frac{n_i}{\sum_{i=1}^{m} n_i}$ n_i= absolute frequency The higher G and G_n, the higher is the heterogeneity: it means that categories have similar frequencies The lower the index, the lower is the heterogeneity: a few classes account for majority of instances
Shannon	$S = -\sum_{i=1}^{m} f_i \ln f_i$	$S_n = -\frac{1}{\ln m} \sum_{i=1}^{m} f_i \ln f_i$	For m, f, f_i and n_i check Gini Higher values of S and S_n indicate higher diversity in terms of similar abundances in classes The lower the index, the lower is the diversity, because a few classes account for most of the data
Simpson	$D = \frac{1}{\sum_{i=1}^{m} f_i^2}$	$D_n = \frac{1}{m-1}\left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$	For m, f, f_i and n_i check Gini Higher values of D and D_n indicate higher diversity in terms of probability of belonging to different classes The lower the index, the lower is the diversity, because frequencies are concentrated in a few classes

The identification of protected attributes can be related to the characteristics listed in Article 21 - Non- discrimination of the EU Charter of Human Rights [20].

We chose the COMPAS dataset because it is probably the most known source in the scientific communities that study bias and fairness of algorithms. As we summarized previously, Pro Publica showed that the COMPAS algorithm classified black people with a much higher risk of recidivism than white people. Here, the probability of being predicted high risk was 47% for black people and 24% for white people, and a similar difference was observed in the false positives rate, i.e. 31% black people vs 14% white people. This occurred mainly because input data is highly imbalanced. In other words, not only black defendants in the dataset are many more than white defendants, with a 51% vs 34% ratio, but the ratio of black recidivist in the whole dataset was double the ratio of white recidivist with 27% against 13%. Similar, although less striking, considerations can be made for the gender attributes whereby women labeled high-risk got a much lower risk of recidivating than men classified as high-risk. The age attribute, on the other hand, was the stronger predictor of high score for violent recidivism (details are available in [32]).

21 Taking into considerations these problematic aspects, we make the following computations:

- We summarize the frequencies of ethnicity, gender, and age categories in Table 2 , both in terms of the overall percentage and as to the ratio of recidivists;
- We compute the imbalance measures on ethnicity and gender categories, both in the whole dataset and on recidivists only, and we report results in Table 3, embedding histograms.

22 We first look at the measurements in the whole dataset. There, all indexes are able to detect the imbalance in the three classes. However, each index has a different sensibility.

- Simpson values are much lower than Gini and Shannon and they point to ethnicity as the most imbalanced data, followed by sex and age categories. This index is more sensitive to the number of possible instances, i.e. six for ethnicity, two for gender and three for the age category;
- Shannon provides the same order of risk provided by Simpson, however, with higher values and shorter distances between rank positions, namely 0.08 between 1st-2nd and 0.19 between 2nd-3rd positions vs respectively 0.14 and 0.24 in Simpson, which results in more distinct values;
- Gini is different because it highlights a higher risk for the sex column, thus reflecting its strongest influence as a predictor, followed by ethnicity and age categories.

23 Looking at the column "percentage recidivists" in Table 2, we observe that measures are lower than the previous column, reflecting an even higher imbalance in the values of the three classes:

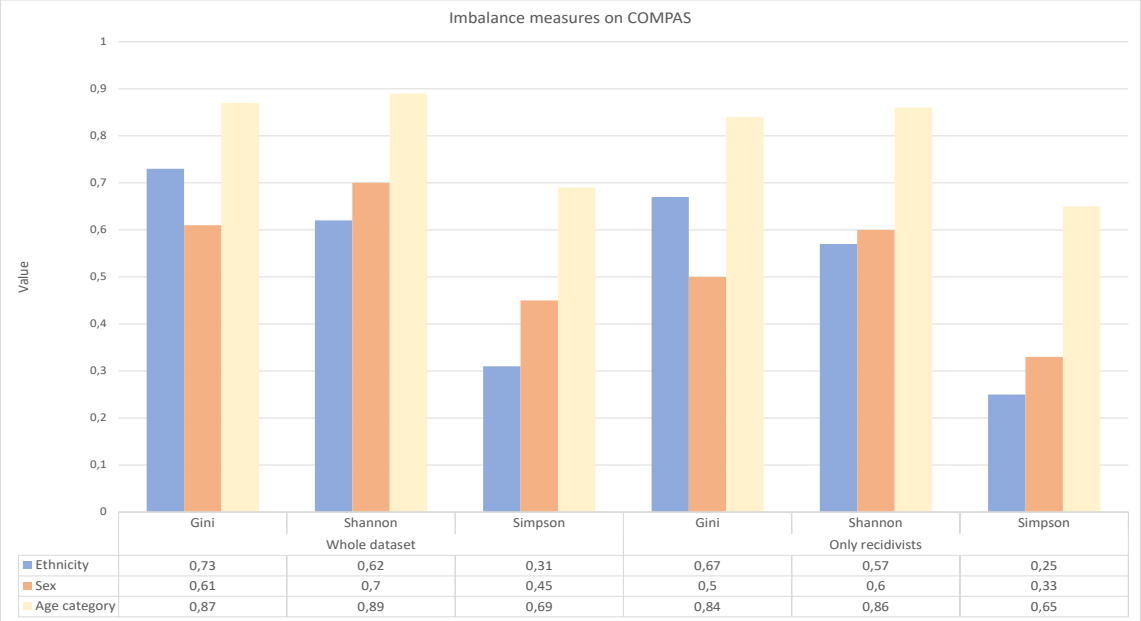
- Simpson preserves the previous rank, but the distance between ethnicity and sex is closer, while the age category has only a slight decrease;
- Shannon keeps being very similar to Simpson, however with higher values;
- Gini also preserves its rank of values, but the difference between the first and second positions is larger now.

24 The question is which index to use. Given that in COMPAS the most severe problem occurred with ethnicity, the answer for this specific dataset would be the Simpson index, due to the fact that it identifies the highest imbalance in a more distinct way. However, this is a consideration made *a posteriori*, on a well-known case with a well-established problem. In view of the future real cases, especially in the design and production phase of an ADM system where there is no information on how the system behaves in operation, a certain number of further considerations should be made, with the most relevant being how to handle a divergence of index values, how to choose meaningful severity thresholds for each index, and which actions to take after the risk is recognized as a relevant concern. To resolve these issues and to make the measures trustable as risk indicators, their reliability shall be extensively investigated, taking also into consideration different types of data and classification/prediction algorithms, the application domain and the groups of stakeholders who are potentially impacted. We will make a further mention to this future work in the last section of the manuscript.

Table 2 Frequency of occurrences for attributes in ethnicity, sex, age categories in COMPAS

ATTRIBUTE	ATTRIBUTE VALUE	OVERALL PERCENTAGE	PERCENTAGE RECIDIVISTS
ETHNICITY	African-American	51.4%	26.9%
	Caucasian	34.1%	13.3%
	Hispanic	8.2%	3.1%
	Asian	0.5%	0.1%
	Native American	0.2%	0.1%
	Other	5.6%	2.0%
SEX	Male	81.0%	38.8%
	Female	19.0%	6.7%
AGE CATEGORY	Less than 25	21.8%	12.2%
	Between 25 and 45	57.2%	26.6%
	Greater than 45	20.9%	6.7%

Table 3 Application of the indexes to COMPAS database



	Whole dataset			Only recidivists		
	Gini	Shannon	Simpson	Gini	Shannon	Simpson
Ethnicity	0,73	0,62	0,31	0,67	0,57	0,25
Sex	0,61	0,7	0,45	0,5	0,6	0,33
Age category	0,87	0,89	0,69	0,84	0,86	0,65

D. Relations to research in algorithmic bias and fairness

25 In recent years, much ink has been spilled on bias and fairness in algorithms. An impressive amount of scientific research has been carried out, especially in the machine learning communities, in order to elaborate strategies that would lead to more equitable results of ADM systems. Efforts mainly focus on techniques to detect systematic discriminations and mitigating them according to different definitions of fairness. Excellent references for getting an overall picture are the survey on bias and fairness in machine learning by Mehrabi et al. [33] , the comprehensive, and still ongoing, work on fairness in machine learning by Barocas et al. [34] and the review of discrimination measures for algorithm decision making by Žliobaitė [35]. A common limitation of these approaches is that mathematical formalizations of fairness cannot be simultaneously satisfied [36][37]. In other words, no universally accepted notion of fairness exists, since defining "fair impact" implicitly embodies political, economic or cultural visions [38]. The ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT[10]) has recognized this issue and has been designed and promoted not only for computer scientists working in the area, but also for scholars and practitioners from "law, social sciences and humanities to investigate and tackle issues in this emerging area". Our approach can be located in this space of inter-disciplinary discussion. It contributes to the main corpus of researches on algorithmic bias and fairness by moving the focus from the outcomes of ADM systems to their inputs, and by making a first step to fill a well-recognized existing gap in the literature, as reported in recent studies, such as in [39] ("There is a need to consider social-minded measures along the whole data pipeline") and in [40] ("returning to the idea of unfairness suggests several new areas of inquiry [...] a shift in focus from outcomes to inputs and processes").

26 In addition, we aim at reaching a higher generalizability of what we currently observe in the field of research. Namely, i) our approach can be applied to any ADM system which is data-based, and not only in machine learning; ii) we build our theoretical framework upon a series of international standards, which incorporate *by design* a multi-stakeholder perspective; iii) we look at data imbalance as risk factor and not as a technical fix, despite the fact that there are well-established techniques for reducing data imbalance in the field of data engineering, especially for machine learning, where the problem has been spotted since the beginning of the 2000's [13]). In this context, we think it is preferable to keep the ultimate responsibility in the realm of human agency. We believe that a risk approach is more suitable for the scope, as it creates space for active human considerations and interventions, rather than delegating the mitigation of the problem to yet another algorithm.

10 See <https://facctconference.org/>.

27 An approach similar to ours and with a wider scope is the work of Takashi Matsumoto and Arisa Ema [41], who proposed a risk chain model for risk reduction in Artificial Intelligence (AI) services, named RCM. By applying RCM in a given risk scenario, it can be proven that a propagation occurs from the technical components of AI systems (data and model) up to the user's understanding, behavior, and usage environment, passing through the service operation management and aspects related to the code of conduct of the service provider as well as the communication with users. The authors consider both data quality and data imbalance as risk factors, whereby they stress the importance of visualizing the relations between risk factors for the purpose of a better planned risk control. While our work is smaller in scope, we think that it can be easily plugged into the RCM framework, due to the fact that we offer a quantitative way to measure imbalance, backed by a structural relation to the ISO/IEC standards on software quality requirements and risk management. Furthermore, it shall be clarified that we did not address data quality as a risk factor given that data quality metrics are well-established in SQuaRE. Nevertheless, we recognize that specific studies would be necessary for selecting the types of measures for data quality that are suitable in the management of ADM system risks.

28 Other approaches which can be related to ours are in the direction of labeling datasets. Two of our previously published studies suggest i) the "Ethically and socially-aware labeling" (EASAL) [42] which aims at developing datasets metadata in order to raise the awareness of the risks of discriminative operations by ADM systems. And secondly, ii) an exploratory analysis of imbalance metrics on two datasets [43], on the basis of which we better specified the theoretical foundations of our approach, and extended the analysis to cover COMPAS. In the context of dataset labeling, the "The Dataset Nutrition Label Project"[11] has been an inspiring work for us. Similar to nutrition labels on food, this initiative aims to identify the "key ingredients" in a dataset such as provenance, populations, and missing data. The label takes the form of an interactive visualization that allows for exploring the previously mentioned aspects. Here, the ultimate goal is to avoid the fact that flawed, incomplete, skewed or problematic data would have a negative impact on automated decision systems, and to drive to the creation of more inclusive algorithms. Notably, our measures could be integrated in this project. Yet another labeling approach is "Datasheets for Datasets" [44]. With respect to other initiatives, this proposal consists of more discursive technical sheets for the purpose of encouraging an increasingly clear and comprehensive communication between users of a dataset and its creators. Eventually, it is worth mentioning the project called "DataTags - Share Sensitive Data with Confidence". [12]The aim of this project is to support researchers who are not legal or technical experts in investigating considerations about proper handling of human subjects' data, and to make informed decisions when collecting, storing, and sharing sensitive data.

E. Relations to European Union policy

29 We extensively reported on how and why bias (imbalance) in data used by ADM systems challenge a founding element of the rule of law of our democratic societies: the principle of non-discrimination [20]. The "Recommendation of the Committee of Ministers to member states on the human rights impacts of algorithmic systems" [45], published by the Council of Europe (CoE) on 8 April 2020, emphasizes the impact of algorithmic systems on human rights and the need for additional normative protections. Although the CoE cannot issue binding laws, it is the main organization for safeguarding human rights in the Europe, and for this reason the recommendation is of particular interest for our purposes. The document defines "high risk" in correspondence with "the use of algorithmic systems in processes or decisions that can produce serious consequences for individuals or in situations where the lack of alternatives prompts a particularly high probability of infringement of human rights, including by introducing or amplifying distributive injustice" (p.5). In these situations, "risk-management processes should detect and prevent the detrimental use of algorithmic systems and their negative impacts" (p.6). The recommended obligations for the states include a continuous review of algorithmic systems throughout their entire lifecycle. In terms of data management, bias in the data as risk factor for systematic discrimination is explicitly cited: "States should carefully assess what human rights and non-discrimination rules may be affected as a result of the quality of data that are being put into and extracted from an algorithmic system, as these often contain bias and may stand in as a proxy for classifiers such as gender, race, religion, political opinion or social origin" (p.7). The document adds that bias and discriminatory outputs should be properly tested since the analysis and modeling phase and even "discontinued if testing or deployment involves the externalization of risks or costs to specific individuals, groups, populations and their environments" (p.8). Precautionary measures should include risk assessment procedures to evaluate potential risks

11 It is a joint initiative of MIT Media Lab and Berkman Klein Center at Harvard University <https://datanutrition.org/>.

12 See <https://techscience.org/a/2015101601/>.

and minimize adverse effects, in cooperation with all relevant stakeholders. Similar obligations are recommended to the private sector.

30 Looking at the Institutions of the European Union (EU), the problem of biased ADM systems is widely recognized, as acknowledged by the words of Margrethe Vestager[13] : "If they're trained on biased data then they can learn to repeat those same biases. Sadly, our societies have such a history of prejudice that you need to work very hard to get that bias out" [46]. The words of M. Vestager should be considered in the context of the ongoing efforts of the EU to redefine the markets rules in response to the rapid technological advancements related to the emergence of automated decision making processes. As a matter of fact, we report the "Resolution on automated decision making processes and consumer protection" [47] which was approved by the EU Parliament on 6 February 2020. The document is relevant because it comes from the highest legislative Institution in the EU and because therein, we find explicit references to the two foundational elements of our proposals. More precisely, the Parliament stresses:

- "the need for a risk-based approach to regulation, in light of the varied nature and complexity of the challenges created by different types and applications of AI and automated decision-making systems" (p. 4);
- "the importance of using only high-quality and unbiased data sets in order to improve the output of algorithmic systems and boost consumer trust and acceptance" (p.11-12).

31 Although the general context of the Resolution is market surveillance, it is still within the ambit of the EU Charter of Fundamental Rights, and in particular Article 38 on consumer protection [48]. It is worth reminding that the European Commission acknowledged the problem of biased ADM since the publication of its communication "Artificial Intelligence for Europe" [49] on 25 April 2018 by stipulating "Whilst AI clearly generates new opportunities, it also poses challenges and risks, for example [...] bias and discrimination" (p.15). Notwithstanding the non-binding value of the document, this communication paved the way to several other policy documents[14]. In the given policy document examples, the term "risk management" recurred often and hitherto it is indicated as the more suitable approach for regulating algorithmic systems[15].

32 This short overview of the most recent efforts on regulating algorithmic systems in Europe, although not exhaustive, defines a further perspective from which our proposal should be derived. In fact, we showed that the risk-based approach is a cornerstone element of the European approach to regulating algorithmic systems, which is currently under redefinition. As a consequence, our proposal can potentially cross this path, whereby balance measures can be suitable risk indicators of propagation (or even amplification) of bias in the input data of ADM systems. In addition, they can be used for certification and labeling purposes, as our notes in the preceding section highlighted.

F. Conclusions: limitations and future work

33 This study faces a problem of wide impact, but it has a well limited boundary of applicability. We take action concerning the problem of systematic discriminations caused by the use of ADM systems, and we focus on a very specific cause, i.e., the imbalance in the data used as input. We propose a metric-based approach in order to evaluate imbalance in a given dataset as a risk factor of discriminatory output of ADM systems. This approach has its foundations on the ISO standards on software quality and risk management. We identify three measures for categorical data, and we run an illustrative example on three columns of the COMPAS dataset, a well-recognized and widely debated case, where imbalance was the main cause of discriminative software output. The example shows that all the indexes detect imbalance, however with different severity and with little variation in the rank of risks. The example, and the study in general, falls short in defining how to effectively manage the risk after the identification. This is a structural, albeit temporary, limitation of the proposal. In fact, in order to derive criteria for action, a systematic investigation is necessary to assess the reliability of the indexes, to identify how their sensibility to imbalance changes in correspondence with different types of data and algorithms used, and to find meaningful thresholds of risks in relation to the context of use and the severity of the impact on individuals. We are working in this direction and we

13 Margrethe Vestager is the Executive Vice President of the European Commission for A Europe Fit for the Digital Age since December 2019 and European Commissioner for Competition since 2014.

14 Including the Coordinated Plan on Artificial Intelligence, the Strategy for Artificial Intelligence, and the very recent (15 December 2020) Digital Services Act draft proposal of the European Commission.

15 Risk management is also a cornerstone element of the AI regulation proposal by the European Commission, which was intentionally left out of the scope of the policy overview because subject to numerous future negotiations.

will be able to elaborate the first guidelines in the following months, thus increasing the internal validity of the present study. Extensive analyses on real systems and replications from third parties will be necessary in order to improve the external validity. We will therefore try to engage researchers in a community effort for testing the measures and to build an open benchmark.

34 We conclude remarking that a much wider number of technical and societal risk factors connected to the deployment of ADM systems exist. For the reader who would like to get an overarching vision, we recommend policy and research reports which investigate the impact of ADM systems, including AI systems, on human rights[16]. For all other readers who stumbled upon this manuscript, we hope that the proposal, despite its current limitations, provided useful insights as a valuable contribution in the common effort of building and regulating algorithmic decision making in a socially sustainable way. More importantly, such is aimed in the direction of protecting individual and collective rights, as well as the promotion of freedoms and the flourishing of our democratic societies.

Acknowledgments

We would like to thank Prof. Marco Ricolfi for his careful review and precious suggestions to the text. We are also grateful to Eleonora Bassi and Giovanni Battista Gallus for inspiring the risk management approach, to Prof. Juan Carlos De Martin for his recommendations on data labeling, and to Prof. Marco Torchiano for his contributions to the data quality aspect of the approach. A special mention also to Elena Beretta and Mariachiara Mecati, who are working on their PhD on translating the concepts described here into practice.

16 Some illustrative but not exhaustive references: [50] [51] [52] [53].

Bibliography

[1] F. Chiusi, S. Fischer, N. Kayser-Bril, and M. Spielkamp, "Automating Society Report 2020," Berlin, Oct. 2020. Accessed: Nov. 10, 2020. [Online]. Available: https://automatingsociety.algorithmwatch.org.

[2] B. Reese, *The fourth age: Smart robots, conscious computers, and the future of humanity*. Simon and Schuster, 2018.

[3] E. Brynjolfsson and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, Reprint edition. New York London: W. W. Norton & Company, 2016.

[4] I. I. Makrygianni and A. P. Markopoulos, "Loan Evaluation Applying Artificial Neural Networks," in *Proceedings of the SouthEast European Design Automation, Computer Engineering, Computer Networks and Social Media Conference*, New York, NY, USA, Sep. 2016, pp. 124–128, doi: 10.1145/2984393.2984407.

[5] Z. Siting, H. Wenxing, Z. Ning, and Y. Fan, "Job recommender systems: A survey," in *2012 7th International Conference on Computer Science Education (ICCSE)*, Jul. 2012, pp. 920–924, doi: 10.1109/ICCSE.2012.6295216.

[6] D. Abu Elyounes, "'Computer Says No!': The Impact of Automation on the Discretionary Power of Public Officers," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3692792, Sep. 2020. [Online]. Available: https://papers.ssrn.com/abstract=3692792.

[7] S. Kanoje, D. Mukhopadhyay, and S. Girase, "User Profiling for University Recommender System Using Automatic Information Retrieval," *Procedia Comput. Sci.*, vol. 78, pp. 5–12, Jan. 2016, doi: 10.1016/j.procs.2016.02.002.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[9] B. Friedman and H. Nissenbaum, "Bias in Computer Systems," *ACM Trans Inf Syst*, vol. 14, no. 3, pp. 330–347, Jul. 1996, doi: 10.1145/230538.230561.

[10] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2477899, 2016. Accessed: Jul. 16, 2019. [Online]. Available: https://papers.ssrn.com/abstract=2477899.

[11] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Reprint edition. New York: Broadway Books, 2017.

[12] G. Ristanoski, W. Liu, and J. Bailey, "Discrimination aware classification for imbalanced datasets," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, New York, NY, USA, Oct. 2013, pp. 1529–1532, doi: 10.1145/2505515.2507836.

[13] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[14] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[15] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.

[16] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 10, 2018. https://reut.rs/2Od9fPr (accessed Nov. 10, 2020).

[17] M. De-Arteaga *et al.*, "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2019, pp. 120–128, doi: 10.1145/3287560.3287572.

[18] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke, "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, p. 199:1-199:30, Nov. 2019, doi: 10.1145/3359301.

[19] N. Kayser-Bril, "Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery," *AlgorithmWatch*. https://algorithmwatch.org/en/story/automated-discrimination-facebook-google/ (accessed Nov. 10, 2020).

[20] European Union, "EU Charter of Fundamental Rights - Article 21 - Non-discrimination," *European Union Agency for Fundamental Rights*, 2007. https://fra.europa.eu/en/eu-charter/article/21-non-discrimination (accessed Nov. 10, 2020).

[21] T. Jan and E. Dwoskin, "Facebook is sued by HUD for housing discrimination," *The Washington Post*. https://www.washingtonpost.com/business/2019/03/28/hud-charges-facebook-with-housing-discrimination (accessed Nov. 10, 2020).

[22] M. Kay, C. Matuszek, and S. A. Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2015, pp. 3819–3828, doi: 10.1145/2702123.2702520.

[23] L. Sweeney, "Discrimination in online ad delivery," *Commun. ACM*, vol. 56, no. 5, pp. 44–54, May 2013, doi: 10.1145/2447976.2447990.

[24] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed Nov. 10, 2020).

[25] Z. Obermeyer and S. Mullainathan, "Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2019, p. 89, doi: 10.1145/3287560.3287593.

[26] A. Kaushal, R. Altman, and C. Langlotz, "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms," *JAMA*, vol. 324, no. 12, p. 1212, Sep. 2020, doi: 10.1001/jama.2020.12067.

[27] "World stumbling zombie-like into a digital welfare dystopia, warns UN human rights expert," *OHCHR | United Nations Human Rights - Office of the High Commissioner*, Oct. 17, 2019. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25156&LangID=E (accessed Nov. 10, 2020).

[28] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, 2018.

[29] International Organization for Standardization, "ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE," *ISO-International Organization for Standardization*, 2014. https://www.iso.org/standard/64764.html (accessed Nov. 10, 2020).

[30] L. Duboc, C. McCord, C. Becker, and S. I. Ahmed, "Critical Requirements Engineering in Practice," *IEEE Softw.*, vol. 37, no. 1, pp. 17–24, Jan. 2020, doi: 10.1109/MS.2019.2944784.

[31] International Organization for Standardization, "ISO 31000:2018 Risk management — Guidelines," *ISO - International Organization for Standardization*, 2018. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65694.html (accessed Nov. 10, 2020).

[32] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed Nov. 10, 2020).

[33] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ArXiv190809635 Cs*, Sep. 2019, [Online]. Available: http://arxiv.org/abs/1908.09635.

[34] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019.

[35] I. Žliobaitė, "Measuring discrimination in algorithmic decision making," *Data Min. Knowl. Discov.*, vol. 31, no. 4, pp. 1060–1089, Jul. 2017, doi: 10.1007/s10618-017-0506-1.

[36] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im)possibility of fairness," *ArXiv160907236 Cs Stat*, Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.07236.

[37] J. Kleinberg, "Inherent Trade-Offs in Algorithmic Fairness," in *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, New York, NY, USA, Jun. 2018, p. 40, doi: 10.1145/3219617.3219634.

[38] E. Beretta, A. Santangelo, B. Lepri, A. Vetrò, and J. C. De Martin, "The Invisible Power of Fairness. How Machine Learning Shapes Democracy," in *Advances in Artificial Intelligence*, Cham, 2019, pp. 238–250.

[39] E. Pitoura, "Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias," *J. Data Inf. Qual.*, vol. 12, no. 3, p. 12:1-12:8, Jul. 2020, doi: 10.1145/3404193.

[40] B. Hutchinson and M. Mitchell, "50 Years of Test (Un)fairness: Lessons for Machine Learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2019, pp. 49–58, doi: 10.1145/3287560.3287600.

[41] T. Matsumoto and A. Ema, "RCModel, a Risk Chain Model for Risk Reduction in AI Services," *ArXiv200703215 Cs*, Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.03215.

[42] E. Beretta, A. Vetrò, B. Lepri, and J. C. De Martin, "Ethical and Socially-Aware Data Labels," in *Information Management and Big Data*, Cham, 2019, pp. 320–327.

[43] M. Mecati, F. E. Cannavò, A. Vetrò, and M. Torchiano, "Identifying Risks in Datasets for Automated Decision–Making," in *Electronic Government*, Cham, 2020, pp. 332–344, doi: 10.1007/978-3-030-57599-1_25.

[44] T. Gebru *et al.*, "Datasheets for Datasets," *ArXiv180309010 Cs*, Mar. 2020, [Online]. Available: http://arxiv.org/abs/1803.09010.

[45] Council of Europe, "Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems." Apr. 08, 2020, [Online]. Available: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016809e1154.

[46] M. Vestager, "Algorithms and democracy - AlgorithmWatch Online Policy Dialogue." Oct. 30, 2020, [Online]. Available: https://ec.europa.eu/commission/commissioners/2019-2024/vestager/announcements/algorithms-and-democracy-algorithmwatch-online-policy-dialogue-30-october-2020_en.

[47] P. De Sutter, "Motion for a Resolution - on automated decision-making processes: ensuring consumer protection and free movement of goods and services." Feb. 06, 2020, [Online]. Available: https://www.europarl.europa.eu/doceo/document/B-9-2020-0094_EN.pdf.

[48] European Union, "EU Charter of Fundamental Rights - Article 38 - Consumer protection," *European Union Agency for Fundamental Rights*, 2007. https://fra.europa.eu/en/eu-charter/article/38-consumer-protection.

[49] European Commission, "Artificial Intelligence for Europe." Apr. 25, 2018, [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625.

[50] G. Noto La Diega, "Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3188080, May 2018. [Online]. Available: https://papers.ssrn.com/abstract=3188080.

[51] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3518482, Jan. 2020. doi: 10.2139/ssrn.3518482.

[52] A. Mantelero, "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment," *Comput. Law Secur. Rev.*, vol. 34, no. 4, pp. 754–772, Aug. 2018, doi: 10.1016/j.clsr.2018.05.017.

[53] D. Allison-Hope, "Artificial Intelligence: A Rights-Based Blueprint for Business," *BSR*, Aug. 28, 2018.

https://www.bsr.org/en/our-insights/report-view/artificial-intelligence-a-rights-based-blueprint-for-business.