Received: 23 September 2019 | Revised: 17 December 2019 | Accepted: 31 December 2019

DOI: 10.1002/widm.1356

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

OVERVIEW

Bias in data-driven artificial intelligence systems—An introductory survey

Eirini Ntoutsi[1] | Pavlos Fafalios[2] | Ujwal Gadiraju[1] | Vasileios Iosifidis[1] | Wolfgang Nejdl[1] | Maria-Esther Vidal[3] | Salvatore Ruggieri[4] | Franco Turini[4] | Symeon Papadopoulos[5] | Emmanouil Krasanakis[5] | Ioannis Kompatsiaris[5] | Katharina Kinder-Kurlanda[6] | Claudia Wagner[6] | Fariba Karimi[6] | Miriam Fernandez[7] | Harith Alani[7] | Bettina Berendt[8,9] | Tina Kruegel[10] | Christian Heinze[10] | Klaus Broelemann[11] | Gjergji Kasneci[11] | Thanassis Tiropanis[12] | Steffen Staab[1,12,13]

[1] L3S Research Center & Faculty of Electrical Engineering and Computer Science, Leibniz University Hannover, Hannover, Germany

[2]Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Greece

[3]TIB Leibniz Information Centre For Science and Tecnhnology, Hannover, Germany

[4]KDDLAB, Dipartimento di Informatica, Università di Pisa, Pisa, Italy

[5]Information Technologies Institute, The Centre for Research & Technology, Hellas (CERTH), Thessaloniki, Greece

[6]GESIS Leibniz Institute for the Social Sciences, Cologne, Germany

[7]Knowledge Media Institute, The Open University, Milton Keynes, UK

[8]Faculty of Electrical Engineering and Computer Science, TU Berlin, Berlin, Germany

[9]Department of Computer Science, KU Leuven, Leuven, Belgium

[10]Institute for Legal Informatics, Leibniz University of Hanover, Hanover, Germany

[11]Innovation Lab, SCHUFA Holding AG, Wiesbaden, Germany

[12]Electronics and Computer Science, University of Southampton, Southampton, UK

[13]Institute for Parallel and Distributed Systems, University of Stuttgart, Germany

Correspondence
Eirini Ntoutsi, L3S Research Center & Faculty of Electrical Engineering and Computer Science, Leibniz University Hannover, Hannover, Germany.
Email: ntoutsi@l3s.de

Funding information
European Commission, Grant/Award Number: 860630

Abstract

Artificial Intelligence (AI)-based systems are widely employed nowadays to make decisions that have far-reaching impact on individuals and society. Their decisions might affect everyone, everywhere, and anytime, entailing concerns about potential human rights issues. Therefore, it is necessary to move beyond traditional AI algorithms optimized for predictive performance and embed ethical and legal principles in their design, training, and deployment to ensure social good while still benefiting from the huge potential of the AI technology. The goal of this survey is to provide a broad multidisciplinary overview of the area of bias in AI systems, focusing on technical challenges and solutions as well as to suggest new research directions towards approaches well-grounded in a legal frame. In

this survey, we focus on data-driven AI, as a large part of AI is powered nowadays by (big) data and powerful machine learning algorithms. If otherwise not specified, we use the general term bias to describe problems related to the gathering or processing of data that might result in prejudiced decisions on the bases of demographic features such as race, sex, and so forth.

This article is categorized under:
Commercial, Legal, and Ethical Issues > Fairness in Data Mining
Commercial, Legal, and Ethical Issues > Ethical Considerations
Commercial, Legal, and Ethical Issues > Legal Issues

KEYWORDS

fairness, fairness-aware AI, fairness-aware machine learning, interpretability, responsible AI

1 | INTRODUCTION

Artificial Intelligence (AI) algorithms are widely employed by businesses, governments, and other organizations in order to make decisions that have far-reaching impacts on individuals and society. Their decisions might influence everyone, everywhere, and anytime, offering solutions to problems faced in different disciplines or in daily life, but at the same time entailing risks like being denied a job or a medical treatment. The discriminative impact of AI-based decision-making to certain population groups has been already observed in a variety of cases. For instance, the COMPAS system for predicting the risk of re-offending was found to predict higher risk values for black defendants (and lower for white ones) than their actual risk (Angwin, Larson, Mattu, & Kirchner, 2016) (*racial-bias*). In another case, Google's Ads tool for targeted advertising was found to serve significantly fewer ads for high paid jobs to women than to men (Datta, Tschantz, & Datta, 2015) (*gender-bias*). Such incidents have led to an ever increasing public concern about the impact of AI in our lives.

Bias is not a new problem rather "bias is as old as human civilization" and "it is human nature for members of the dominant majority to be oblivious to the experiences of other groups."[1] However, AI-based decision-making may magnify pre-existing biases and evolve new classifications and criteria with huge potential for new types of biases. These constantly increasing concerns have led to a reconsideration of AI-based systems towards new approaches that also address the fairness of their decisions. In this paper, we survey recent technical approaches on bias and fairness in AI-based decision-making systems, we discuss their legal ground[2] as well as open challenges and directions towards AI-solutions for societal good. We divide the works into three broad categories:

- *Understanding bias*. Approaches that help understand how bias is created in the society and enters our socio-technical systems, is manifested in the data used by AI algorithms, and can be modeled and formally defined.
- *Mitigating bias*. Approaches that tackle bias in different stages of AI-decision making, namely, preprocessing, in-processing, and post-processing methods focusing on data inputs, learning algorithms, and model outputs, respectively.
- *Accounting for bias*. Approaches that account for bias proactively, via bias-aware data collection, or retroactively, by explaining AI-decisions in human terms.

Figure 1 provides a visual map of the topics discussed in this survey.

This paper complements existing surveys that either have a strong focus on machine ethics, such as Yu et al. (2018), study a specific subproblem, such as explaining black box models (Atzmueller, 2017; Guidotti et al., 2019), or focus in specific contexts, such as the Web (Baeza-Yates, 2018), by providing a broad categorization of the technical challenges and solutions, a comprehensive coverage of the different lines of research as well as their legal grounds.

We are aware that the problems of bias and discrimination are not limited to AI and that the technology can be deployed (consciously or unconsciously) in ways that reflect, amplify or distort real world perception, and status quo. Therefore, as the roots to these problems are not only technological, it is also naive to believe that technological solutions will suffice. Rather, more than technical solutions are required including socially acceptable definitions of fairness and meaningful interventions to ensure the long-term well-being of all groups. These challenges require

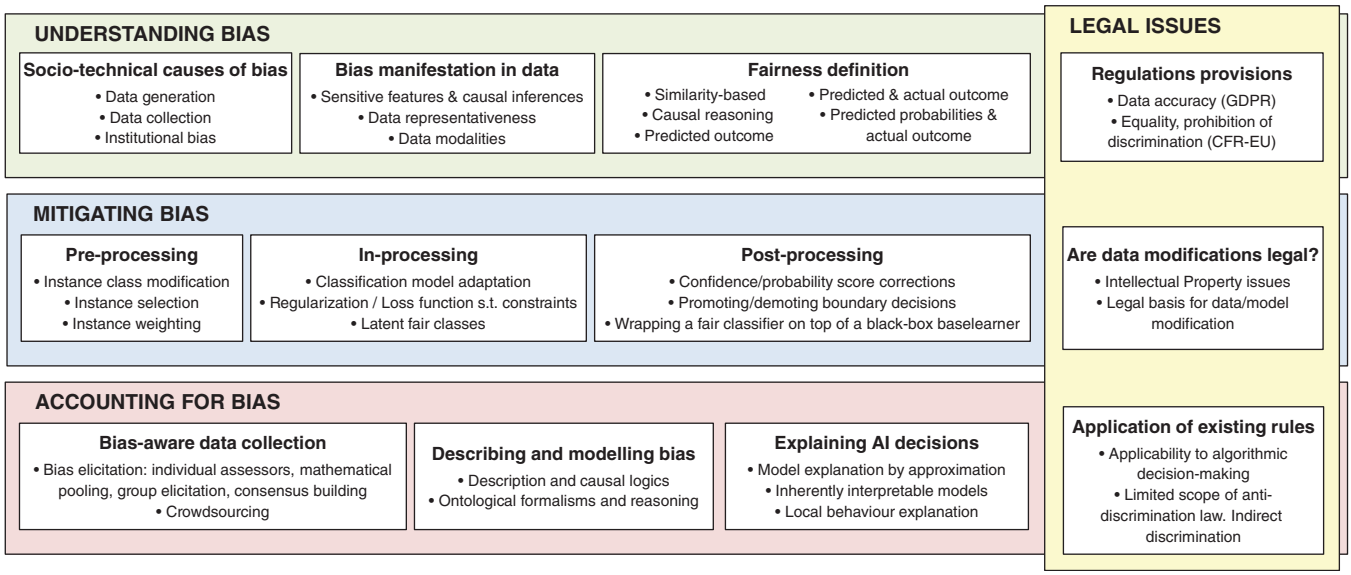

FIGURE 1 Overview of topics discussed in this survey

multidisciplinary perspectives and a constant dialogue with the society as bias and fairness are multifaceted and volatile. Nevertheless, as the AI technology penetrates our lives, it is extremely important for technology creators to be aware of bias and discrimination and to ensure responsible usage of the technology, keeping in mind that a technological approach on its own is not a panacea for all sorts of bias and AI problems.

2 | UNDERSTANDING BIAS

Bias is an old concept in machine learning (ML), traditionally referring to the assumptions made by a specific model (*inductive bias*) (Mitchell, 1997). A classical example is Occam's razor preference for the simplest hypothesis. With respect to human bias, its many facets have been studied by many disciplines including psychology, ethnography, law, and so forth. In this survey, we consider as bias the *inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair*. Given this definition, we focus on how bias enters AI systems and how it is manifested in the data comprising the input to AI algorithms. Tackling bias entails answering the question how to define fairness such that it can be considered in AI systems; we discuss different fairness notions employed by existing solutions. Finally, we close this section with legal implications of data collection and bias definitions.

2.1 | Socio-technical causes of bias

AI relies heavily on data generated by humans (e.g., user-generated content) or collected via systems created by humans. Therefore, whatever biases exist in humans enter our systems and even worse, they are amplified due to the complex sociotechnical systems, such as the Web.[3] As a result, algorithms may reproduce (or even increase) existing inequalities or discriminations (Karimi, Génois, Wagner, Singer, & Strohmaier, 2018). Within societies, certain social groups may be disadvantaged, which usually results in "institutional bias" where there is a tendency for the procedures and practices of particular institutions to operate in ways in which some social groups are being advantaged and others disadvantaged. This needs not be the result of conscious discrimination but rather of the majority following existing norms. Institutional racism and sexism are common examples (Chandler & Munday, 2011). Algorithms are part of existing (biased) institutions and structures, but they may also amplify or introduce bias as they favor those phenomena and aspects of human behavior that are easily quantifiable over those which are hard or even impossible to measure. This problem is exacerbated by the fact that certain data may be easier to access and analyze than others, which has caused, for example, the role of Twitter for various societal phenomena to be overemphasized (Tufekci, 2014). Once introduced, algorithmic systems encourage the creation of very specific data collection infrastructures and policies, for

example, they may suggest tracking and surveillance (Introna & Wood, 2004) which then change or amplify power relations. Algorithms thus shape societal institutions and potential interventions, and vice versa. It is currently not entirely clear, how this complex interaction between algorithms and structures plays out in our societies. Scholars have thus called for "algorithmic accountability" to improve understanding of the power structures, biases, and influences that algorithms exercise in society (Diakopoulos, 2015).

2.2 | How is bias manifested in data?

Bias can be manifested in (multimodal) data through sensitive features and their causal influences, or through under/over-representation of certain groups.

2.2.1 | Sensitive features and causal influences

Data encode a number of people characteristics in the form of feature values. Sensitive characteristics that identify grounds of discrimination or bias may be present or not. Removing or ignoring such sensitive features does not prevent learning biased models, because other correlated features (also known as *redundant encodings*) may be used as proxies for them. For example, neighborhoods in U.S. cities are highly correlated with race, and this fact has been used for systematic denial of services such as bank loans or same-day purchase delivery.[4] Rather, including sensitive features in data may be beneficial in the design of fair models (Zliobaite & Custers, 2016). Sensitive features may also be correlated with the target feature that classification models want to predict. For example, a minority's preference for red cars may induce bias against the minority in predicting accident rate if red cars are also preferred by aggressive drivers. Higher insurance premium may then be set for red car owners, which disproportionately impacts minority members. Simple correlation between apparently neutral features can then lead to biased decisions. Discovering and understanding causal influences among variables is a fundamental tool for dealing with bias, as recognized in the legal circles (Foster, 2004) and in medical research (Grimes & Schulz, 2002). The interested reader is referred to the recent survey on causal approaches to fairness in classification models (Loftus, Russell, Kusner, & Silva, 2018).

2.2.2 | Representativeness of data

Statistical (including ML) inferences require that the data from which the model was learned be representative of the data on which it is applied. However, data collection often suffers from biases that lead to the over- or under-representation of certain groups, especially in big data, where many data sets have not been created with the rigor of a statistical study, but they are the by-product of other activities with different, often operational, goals (Barocas & Selbst, 2016). Frequently occurring biases include *selection bias* (certain individuals are more likely to be selected for study), often as *self-selection bias*, and the reverse *exclusion bias*; *reporting bias* (observations of a certain kind are more likely to be reported, which leads to a sort of selection bias on observations); and *detection bias* (a phenomenon is more likely to be observed for a particular set of subjects). Analogous biases can lead to under- or over-representations of properties of individuals, for example Boyd and Crawford (2012)). If the mis-represented groups coincide with social groups against which there already exists social bias such as prejudice or discrimination, even "unbiased computational processes can lead to discriminative decision procedures" (Calders & Zliobaite, 2013). Mis-representation in the data can lead to vicious cycles that perpetuate discrimination and disadvantage (Barocas & Selbst, 2016). Such "pernicious feedback loops" (O'Neil, 2016) can occur with both under-representation of historically disadvantaged groups, for example, women and people of color in IT developer communities and image datasets (Buolamwini & Gebru, 2018), and with over-representation, for example, black people in drug-related arrests (Lum & Isaac, 2016).

2.2.3 | Data modalities and bias

Data come in different modalities (numerical, textual, images, etc.) as well as in multimodal representations (e.g., audio-visual content). Most of the fairness-aware ML approaches refer to structured data represented in some

fixed feature space. Data modality-specific approaches also exist, especially for textual data and images. Bias in language has attracted a lot of recent interest with many studies exposing a large number of offensive associations related to gender and race on publicly available word embeddings (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016) as well as how these associations have evolved over time (Kutuzov, Øvrelid, Szymanski, & Velldal, 2018). Similarly for the computer vision community where standard image collections like MNIST are exploited for training, or off-the-shelf pretrained models are used as feature extractors, assuming the collections comprise representative samples of the real world. In reality, though, the collections can be biased as many recent studies have indicated. For instance, Buolamwini and Gebru (2018) have found that commercial facial recognition services perform much better on lighter male subjects than darker female ones. Overall, the additional layer of feature extraction that is typically used within AI-based multimodal analysis systems makes it even more challenging to trace the source of bias in such systems.

2.3 | How is fairness defined?

More than 20 different definitions of fairness have appeared thus far in the computer science literature (Verma & Rubin, 2018; Zliobaite, 2017); and some of these definitions and others were proposed and investigated in work on formalizing fairness from other disciplines, such as education, over the past 50 years (Hutchinson & Mitchell, 2019). Existing fairness definitions can be categorized into: (a) "predicted outcome," (b) "predicted and actual outcome," (c) "predicted probabilities and actual outcome," (d) "similarity based," and (e) "causal reasoning" (Verma & Rubin, 2018). "Predicted outcome" definitions solely rely on a model's predictions (e.g., *demographic parity* checks the percentage of protected and non-protected groups in the positive class). "Predicted and actual outcome" combine a model's predictions with the true labels (e.g., *equalized odds* requires false positive and negative rates to be similar among protected and non-protected groups). "Predicted probabilities and actual outcome" employ the predicted probabilities instead of the predicted outcomes (e.g., *good calibration* requires the true positive probabilities between protected and non-protected groups to be the same). Contrary to definitions (a)–(c) that only consider the sensitive attribute, "similarity based" definitions also employ non-sensitive attributes (e.g., *fairness through awareness* states that similar individuals must be treated equally). Finally, "causal reasoning" definitions are based on directed acyclic graphs that capture relations between features and their impact on the outcomes by structural equations (e.g., *counterfactual fairness* (Kusner, Loftus, Russell, & Silva, 2017) constructs a graph that verifies whether the attributes defining the outcome are correlated to the sensitive attribute).

Despite the many formal, mathematical definitions of fairness proposed over the last years the problem of formalizing fairness is still open as well as the discussion about the merits and demerits of the different measures is missing. Corbett-Davies and Goel (2018) show the statistical limitations of prevailing mathematical definitions of fairness and the (negative) effect of enforcing such fairness-measures on group well-being and urge the community to explicitly focus on consequences of potential interventions.

2.4 | Legal issues of bias and fairness in AI

Taking into account the variety of bias creation in AI systems and its impact on society, the question arises whether the law should provide regulations for non-discriminatory AI-based decision making. Generally speaking, existing EU regulation comes into play when (discriminatory) decisions have been taken, while provisions tackling the quality of selected data are rare. For the earlier, the control of discriminatory decisions, the principle of equality and the prohibition of discrimination (Art. 20, 21 EU Charter of Fundamental Rights, Art. 4 Directive 2004/113 and other directives) apply. However, these provisions only address discrimination on the basis of specific criteria and require prima facie evidence of a less favorable treatment on grounds of a prohibited criterion, which will often be difficult to establish (Hacker, 2018). For the latter, the control of the quality of the selected data, with respect to "personal data" Art. 5 (1) GDPR,[5] stipulates "the principle of data accuracy" which, however, does not hinder wrongful or disproportionate selection. With respect to automated decision-making (Art. 22 GDPR), recital 71 only points out that appropriate mathematical or statistical procedures shall be used and that discriminatory effects shall be prevented. While the effectiveness of Art. 22 GDPR is uncertain (Zuiderveen Borgesius, 2018) it provides some safeguards, such as restrictions on the use of automated decision-making, and, where it is used, a right to transparency, to obtain human intervention and to

contest the decision. Finally, some provisions in area-specific legislation can be found, for example, Art. 12 Regulation (EC) No 223/2009 for European statistics.

3 | MITIGATING BIAS

Approaches for bias mitigation can be categorized into: (a) preprocessing methods focusing on the data, (b) in-processing methods focusing on the ML algorithm, and (c) post-processing methods focusing on the ML model. We conclude the section with a discussion on the legal issues of bias mitigation.

3.1 | Preprocessing approaches

Approaches in this category focus on the data, the primary source of bias, aiming to produce a "balanced" dataset that can then be fed into any learning algorithm. The intuition behind these approaches is that the fairer the training data is, the less discriminative the resulting model will be. Such methods modify the original data distribution by altering class labels of carefully selected instances close to the decision boundary (Kamiran & Calders, 2009) or in local neighborhoods (Luong, Ruggieri, & Turini, 2011), by assigning different weights to instances based on their group membership (Calders, Kamiran, & Pechenizkiy, 2009) or by carefully sampling from each group. These methods use heuristics aiming to balance the protected and unprotected groups in the training set; however, their impact is not well controlled despite their efforts for minimal data interventions. Recently, Calmon, Wei, Vinzamuri, Ramamurthy, and Varshney (2017) proposed a probabilistic fairness-aware framework that alters the data distribution towards fairness while controlling the per-instance distortion and by preserving data utility for learning.

3.2 | In-processing approaches

In-processing approaches reformulate the classification problem by explicitly incorporating the model's discrimination behavior in the objective function through regularization or constraints, or by training on latent target labels. For example, Kamiran, Calders, and Pechenizkiy (2010) modify the splitting criterion of decision trees to also consider the impact of the split w.r.t. the protected attribute. Kamishima, Akaho, Asoh, and Sakuma (2012) integrate a regularizer to reduce the effect of "indirect prejudice" (mutual information between the sensitive features and class labels). Dwork, Hardt, Pitassi, Reingold, and Zemel (2012) redefine the classification problem by minimizing an arbitrary loss function subject to the *individual fairness-constraint* (similar individuals are treated similarly). Zafar, Valera, Gomez-Rodriguez, and Gummadi (2017) propose a constraint-based approach for *disparate mistreatment* (defined in terms of misclassification rates) which can be incorporated into logistic-regression and SVMs. In a different direction, Krasanakis, Xioufis, Papadopoulos, and Kompatsiaris (2018) assume the existence of latent fair classes and propose an iterative training approach towards those classes which alters the in-training weights of the instances. Iosifidis and Ntoutsi (2019) propose a sequential fair ensemble, AdaFair, that extends the weighted distribution approach of AdaBoost by also considering the cumulative fairness of the learner up to the current boosting round and moreover, it optimizes for balanced error instead of overall error to account for class imbalance.

While most of the in-processing approaches refer to classification, approaches for the unsupervised case have also emerged recently, for example, the fair-PCA approach of Samadi, Tantipongpipat, Morgenstern, Singh, and Vempala (2018) that forces equal reconstruction errors for both protected and unprotected groups. Chierichetti, Kumar, Lattanzi, and Vassilvitskii (2017) formulate the problem of fair clustering as having approximately equal representation for each protected group in every cluster and define fair-variants of classical k-means and k-medoids algorithms.

3.3 | Post-processing approaches

The third strategy is to postprocess the classification model once it has been learned from data. This consists of altering the model's internals (white-box approaches) or its predictions (black-box approaches). Examples of the white-box approach consist of correcting the confidence of CPAR classification rules (Pedreschi, Ruggieri, & Turini, 2009),

probabilities in Naïve Bayes models (Calders & Verwer, 2010), or the class label at leaves of decision trees (Kamiran et al., 2010). White-box approaches have not been further developed in recent years, being superseded by in-processing methods. Examples of the black-box approach aim at keeping proportionality of decisions among protected versus unprotected groups by promoting or demoting predictions close to the decision boundary (Kamiran, Mansha, Karim, & Zhang, 2018), by differentiating the decision boundary itself over groups (Hardt, Price, & Srebro, 2016), or by wrapping a fair classifier on top of a black-box base classifier (Agarwal, Beygelzimer, Dudík, Langford, & Wallach, 2018). An analysis of how to postprocess group-wise calibrated classifiers under fairness constraints is given in (Canetti et al., 2019). While the majority of approaches are concerned with classification models, bias post-processing has been deemed as relevant when interpreting clustering models as well (Lorimer, Held, & Stoop, 2017).

3.4 | Legal issues of mitigating bias

Pertinent legal questions involve whether modifications of data as envisaged by the pre- and in-processing approaches, as well as altering the model in the post-processing approach, could be considered lawful. Besides intellectual property issues that might occur, there is no general legal provision dealing with the way data is collected, selected or (even) modified. Provisions are in place mainly if such training data would (still) be personal data. Modifications (as well as any other processing) would need a legal basis. However, legitimation could derive from informed consent (provided that specific safeguards are met), or could rely on contract or legitimate interest. Besides, data quality could be relevant in terms of warranties, if a data provider sells data. A specific issue arises when "debiasing" involves sensitive data, as under Art. 9 GDPR special category data such as ethnicity often requires explicit consent (Kilbertus et al., 2018). A possible solution could be Art. 9(2)(g) GDPR which permits processing for reasons of substantial public interest, which arguably could be seen in 'debiasing'. The same grounds of legitimation apply when altering the model. However, contrary to data modification, data protection law would arguably not be applicable here, as the model would not contain personal data, unless the model is vulnerable to confidentiality attacks such as model inversion and membership inference (Veale, Binns, & Edwards, 2018).

4 | ACCOUNTING FOR BIAS

Algorithmic accountability refers to the assignment of responsibility for how an algorithm is created and its impact on society (Kaplan, Donovan, Hanson, & Matthews, 2019). In case of AI algorithms the problem is aggravated as we do not codify the solution, rather the solution is inferred via ML algorithms and complex data. AI accountability has many facets, we focus below on the most prominent ones that account for bias either *proactively*, via bias-aware data collection, or *retroactively* by explaining AI decisions in human terms; furthermore, we discuss the importance of describing and documenting bias by means of formalisms like ontologies.

4.1 | Proactively: bias-aware data collection

A variety of methods are adopted for data acquisition to serve diverse needs; these may be prone to introducing bias at the data collection stage itself, for example, Morstatter, Pfeffer, and Liu (2014). Proposals have been made for a structured approach to bias elicitation in evidence synthesis, including bias checklists and elicitation tasks that can be performed either by individual assessors and mathematical pooling, group elicitation and consensus building or hybrid approaches (Turner, Spiegelhalter, Smith, & Thompson, 2009). However, bias elicitations have themselves been found to be biased even when high quality assessors are involved and remedies have been proposed (Manzi & Forster, 2019).

Among other methods, crowdsourcing is a popular approach that relies on large-scale acquisition of human input for dealing with data and label scarcity in ML. Crowdsourced data and labels may be subject to bias at different stages of the process: task design and experimental setup, task decomposition and result aggregation, selection of workers, and the entailing human factors (Hube, Fetahu, & Gadiraju, 2019; Kamar, Kapoor, & Horvitz, 2015; Karger, Oh, & Shah, 2011). Mitigating biases in crowdsourced data becomes harder in subjective tasks, where the presence of varying ideological and cultural backgrounds of workers means that it is possible to observe biased labels with complete agreement among the workers.

4.2 | Describing and modeling bias using ontologies

Accounting for bias not only requires understanding of the different sources, that is, data, knowledge bases, and algorithms, but more importantly, it demands the interpretation and description of the meaning, potential side effects, provenance, and context of bias. Usually unbalanced categories are understood as bias and considered as sources of negative side effects. Nevertheless, skewed distributions may simply hide features or domain characteristics that, if removed, would hinder the discovery of relevant insights. This situation can be observed, for instance, in populations of lung cancer patients. As highlighted in diverse scientific reports, for example, (Garrido et al., 2019), lung cancer in women and men has significant differences such as etiology, pathophysiology, histology, and risk factors, which may impact in cancer occurrence, treatment outcomes, and survival. Furthermore, there are specific organizations that collaborate in lung cancer prevention and in the battle against smoking; some of these campaigns are oriented to particular focus groups and the effects of these initiatives are observed in certain populations. All these facts impact on the gender distribution of the population and could be interpreted as bias. However, in this context, imbalance reveals domain specific facts that need to be preserved in the population, and a formal description of these uneven distributions should be provided to avoid misinterpretation. Moreover, as any type of data source, knowledge bases and ontologies can also suffer from various types of bias or knowledge imbalance. For example, the description of the existing mutations of a gene in a knowledge base like COSMIC,[6] or the properties associated with a gene in the Gene Ontology,[7] may be biased by the amount of research that has been conducted in the diseases associated with these genes. Expressive formal models are demanded in order to describe and explain the characteristics of a data source and under which conditions or context, the data source is biased.

Formalisms like description and causal logics, for example, (Besnard, Cordier, & Moinard, 2014; Dehaspe & Raedt, 1996; Krötzsch, Marx, Ozaki, & Thost, 2018; LeBlanc, Balduccini, & Vennekens, 2019), allow for measuring and detecting bias in data collections of diverse types, for example, online data sets (Pitoura et al., 2017) and recommendation systems (Serbos, Qi, Mamoulis, Pitoura, & Tsaparas, 2017). They also enable the annotation of statements with trustworthiness (Son, Pontelli, Gelfond, & Balduccini, 2016) and temporality (Ozaki, Krötzsch, & Rudolph, 2019), as well as causation relationships between them (LeBlanc et al., 2019). Ontologies also play a relevant role as knowledge representation models for describing universe of discourses in terms of concepts such as classes, properties, and subsumption relationships, as well as contextual statements of these concepts. NdFluents (Giménez-García, Zimmermann, & Maret, 2017) and Context Ontology Language (CoOL) (Strang, Linnhoff-Popien, & Frank, 2003), represent exemplar ontology formal models able to express and combine diverse contextual dimensions and interrelations (e.g., locality and vicinity). Albeit expressive, existing logic-based and ontological formalisms are not tailored for representing contextual bias or differentiating unbalanced categories that consistently correspond to instances of a real-world domain. Therefore, expressive ontological formalisms are demanded to represent the contextual dimensions of various types of sources, for example, data collections, knowledge bases, or ontologies, as well as annotations denoting causality and provenance of the represented knowledge. These formalisms will equip bias detection algorithms with reasoning mechanisms that not only enhance accuracy but also enable explainability of the meaning, conditions, origin, and context of bias. Thus, domain modeling using ontologies will support context-aware bias description and interpretability.

4.3 | Retroactively: explaining AI decisions

Explainability refers to the extent the internal mechanics of a learning model can be explained in human terms. It is often used interchangeably with interpretability, although the latter refers to whether one can predict what will happen given a change in the model input or parameters. Although attempts to tackle *interpretable* ML have existed for some time (Hoffman & Klein, 2017), there has been an exceptional growth of research literature in the last years with emerging keywords such as *explainable AI* (Adadi & Berrada, 2018) and *black box explanation* (Guidotti et al., 2019). Many papers propose approaches for understanding the *global* logic of a model by building an interpretable classifier able to mimic the obscure decision system. Generally, these methods are designed for explaining specific models, for example, deep neural networks (Montavon, Samek, & Müller, 2018). Only few are agnostic to the black box model (Henelius, Puolamäki, Boström, Asker, & Papapetrou, 2014). The difficulties in explaining black boxes and complex models ex post, have motivated proposals of transparent classifiers which are interpretable on their own and exhibit predictive accuracy close to that of obscure models. These include Bayesian models (Li & Huan, 2017), generalized additive

models (Lou, Caruana, Gehrke, & Hooker, 2013), supersparse linear models (Ustun & Rudin, 2016), rule-based decision sets (Lakkaraju, Bach, & Leskovec, 2016), optimal classification trees (Bertsimas & Dunn, 2017), model trees (Broelemann & Kasneci, 2019), and neural networks with interpretable layers (Zhang, Wu, & Zhu, 2018).

A different stream of approaches focuses on the *local* behavior of a model, searching for an explanation of the decision made for a specific instance (Guidotti et al., 2019). Such approaches are either *model-dependent*, for example, Taylor approximations (Kasneci & Gottron, 2016), saliency masks (the image regions that are mainly responsible for the decision) for neural network decisions (Ma, Yu, & Yue, 2015), and attention models for recurrent networks (Choi et al., 2016), or *model-agnostic*, such as those started by the LIME method (Ribeiro, Singh, & Guestrin, 2016). The main idea is to derive a local explanation for a decision outcome on a specific instance by learning an interpretable model from a randomly generated neighborhood of the instance. A third stream aims at bridging the local and the global ones by defining a strategy for combining local models in an incremental way (Pedreschi et al., 2019). More recent work has asked the fundamental question *What is an explanation?* (Mittelstadt, Russell, & Wachter, 2019) and reject such usage of the term "explanation," criticizing that it might be appropriate for a modeling expert, but not for a lay man, and that, for example, humanities or philosophy have an entirely different understanding of what explanations are.

We speculate that there are computational methods that will allow us to find some middle ground. For instance, some approaches in ML, statistical relational learning in particular (Raedt, Kersting, Natarajan, & Poole, 2016), take the perspective of knowledge representation and reasoning into account when developing ML models on more formal logical and statistical grounds. AI knowledge representation has been developing a rich theory of argumentation over the last 25 years (Dung, 1995), which recent approaches (Cocarascu & Toni, 2016) try to leverage for generalizing the reasoning aspect of ML towards the use of computational models of argumentation. The outcome are models of arguments and counterarguments towards certain classifications that can be inspected by a human user and might be used as formal grounds for explanations in the manner that Mittelstadt et al. (2019) called out for.

4.4 | Legal issues of accounting for bias

While data protection rules affect both the input (data) and the output (automated decision) level of AI decision-making, anti-discrimination laws, as well as consumer and competition rules, address discriminatory policies primarily from the perspective of the (automated) decision and the actions based on it. However, the application of these rules to AI-based decisions is largely unclear. Under present law and the principle of private autonomy, decisions by private parties normally do not have to include reasons or explanations. Therefore, a first issue will be how existing rules can be applied to algorithmic decision-making. Given that a decision will often not be reasoned (hence the reasons will be unknown), it will be difficult to establish that it was made on the basis of a biased decision-making process (Mittelstadt et al., 2019).

Even if bias can be proven, a second issue is the limited scope of anti-discrimination law. Under present law, only certain transactions between private parties fall under the EU anti-discrimination directives (Liddell & O'Flaherty, 2018). Moreover, in most cases AI decision-making instruments will not directly use an unlawful criterion (e.g., gender) as a basis for their decision, but rather a "neutral" one (e.g., residence) which in practice lead to a less favorable treatment of certain groups. This raises the difficult concept of indirect discrimination, that is, a scenario where an "apparently neutral rule disadvantages a person or a group sharing the same characteristics" (Liddell & O'Flaherty, 2018). Finally, most forms of differential treatment can be justified where it pursues a legitimate aim and where the means to pursue that aim are appropriate and necessary. It is unclear whether the argument that AI-based decision making systems produce decisions which are economically sound can be sufficient as justification.

5 | FUTURE DIRECTIONS AND CONCLUSIONS

There are several directions that can impact this field going forward. First, despite the large number of methods for mitigating bias, there are still no conclusive results regarding what is the state of the art method for each category, which of the fairness-related interventions perform best, or whether category-specific interventions perform better comparing to holistic approaches that tackle bias at all stages of the analysis process. We believe that a systematic evaluation of the existing approaches is necessary to understand their capabilities and limitations and also, a vital part of proposing new solutions. The difficulty of the evaluation lies on the fact that different methods work with different fairness notions

and are applicable to different AI models. To this end, benchmark datasets should be made available that cover different application areas and manifest real-world challenges. Finally, standard evaluation procedures and measures covering both model performance and fairness-related aspects should be followed, in accordance with international standards like the IEEE—ALGB-WG—Algorithmic Bias Working Group.[8]

Second, we recognize that "fairness cannot be reduced to a simple self-contained mathematical definition," "fairness is dynamic and social and not a statistical issue."[9] Also, "fair is not fair everywhere" (Schäfer, Haun, & Tomasello, 2015) meaning that the notion of fairness varies across countries, cultures and application domains. Therefore, it is important to have realistic and applicable fairness definitions for different contexts as well as domain-specific datasets for method development and evaluation. Moreover, it is important to move beyond the typical training-test evaluation setup and to consider the consequences of potential fairness-related interventions to ensure long-term wellbeing of different groups. Finally, given the temporal changes of fairness perception, the question of whether one can train models on historical data and use them for current fairness-related problems becomes increasingly pressing.

Third, the related work thus far focuses mainly on supervised learning. In many cases however, direct feedback on the data (i.e., as labels) is not available. Therefore alternative learning tasks should be considered, like unsupervised learning or reinforcement learning (RL) where only intermediate feedback is provided to the model. Recent works have emerged in this direction, for example, Jabbari, Joseph, Kearns, Morgenstern, and Roth (2017) examine fairness in the RL context where one needs to reconsider the effects of short-term actions on long-term rewards.

Fourth, there is a general trend in the ML community recently for generating plausible data from existing data using Generative Adversarial Networks in an attempt to cover the high data demand of modern methods, especially DNNs. Recently, such approaches have been used also in the context of fairness (Xu, Yuan, Zhang, & Wu, 2018), that is, how to generate synthetic fair data that are similar to the real data. Still however, the problem of representativeness of the training data and its impact on the representativeness of the generated data might aggravate issues of fairness and discrimination. In the same topic, recent work revealed that DNNs are vulnerable to adversarial attacks, that is, intentional perturbations of the input examples, and therefore there is a need for methods to enhance their resilience (Song et al., 2018).

Fifth, AI scientists and everyone involved in the decision making process should be aware of bias-related issues and the effect of their design choices and assumptions. For instance, studies show that representation-related biases creep into development processes because the development teams are not aware of the importance of distinguishing between certain categories (Buolamwini & Gebru, 2018). Members of a privileged group may not even be aware of the existence of (e.g., racial) categories in the sense that they often perceive themselves as "just people," and the interpretation of this as an unconscious default requires the voice of individuals from underprivileged groups, who persistently perceive their being "different." Two strategies appear promising for addressing this cognitive bias: try to improve diversity in development teams, and subject algorithms to outside and as-open-as-possible scrutiny, for example by permitting certain forms of reverse engineering for algorithmic accountability.

Finally, from a legal point of view, apart from data protection law, general provisions with respect to data quality or selection are still missing. Recently an ISO standard on data quality (ISO 8000) was published, though not binding and not with regard to decision-making techniques. Moreover, first important steps have been made, for example, the Draft Ethics Guidelines for trustworthy AI from the European Commission's high-level Expert group on AI or the European parliament resolution containing recommendations to the Commission on Civil Law Rules on Robotics. However, these resolutions are still generic. Further interdisciplinary research is needed to define specifically what is needed to meet the balance between the fundamental rights and freedoms of citizens by mitigating bias, while at the same time considering the technical challenges and economical needs. Therefore, any legislative procedures will require a close collaboration of legal and technical experts. As already mentioned, the legal discussion in this paper refers to the EU where despite the many recent efforts, there is still no consensus for algorithmic fairness regulations across its countries. Therefore, there is still a lot of work to be done on analyzing the legal standards and regulations at a national and international level to support globally legal AI designs.

To conclude, the problem of bias and discrimination in AI-based decision-making systems has attracted a lot of attention recently from science, industry, society and policy makers, and there is an ongoing debate on the AI opportunities and risks for our lives and our civilization. This paper surveys technical challenges and solutions as well as their legal grounds in order to advance this field in a direction that exploits the tremendous power of AI for solving real world problems but also considers the societal implications of these solutions. As a final note, we want to stress again that biases are deeply embedded in our societies and it is an illusion to believe that the AI and bias problem will be

eliminated only with technical solutions. Nevertheless, as the technology reflects and projects our biases into the future, it is a key responsibility of technology creators to understand its limits and to propose safeguards to avoid pitfalls. Of equal importance is also for the technology creators to realize that technical solutions without any social and legal ground cannot thrive and therefore multidisciplinary approaches are required.

ACKNOWLEDGMENT

This work is supported by the project "NoBias - Artificial Intelligence without Bias," which has received funding from the European Union's Horizon 2020 research and innovation programme, under the Marie Skłodowska-Curie (Innovative Training Network) grant agreement no. 860630.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Eirini Ntoutsi: Conceptualization; investigation; methodology; resources; validation; writing-original draft; writing-review and editing. **Pavlos Fafalios:** Conceptualization; project administration; writing-original draft; writing-review and editing. **Ujwal Gadiraju:** Investigation; resources; writing-original draft. **Vasileios Iosifidis:** Investigation; resources; writing-original draft. **Wolfgang Nejdl:** Conceptualization. **Maria-Esther Vidal:** Resources; writing-original draft. **Salvatore Ruggieri:** Conceptualization; investigation; methodology; resources; writing-original draft; writing-review and editing. **Franco Turini:** Conceptualization; methodology; resources; writing-original draft; writing-review and editing. **Symeon Papadopoulos:** Conceptualization; methodology; resources; visualization; writing-original draft; writing-review and editing. **Emmanouil Krasanakis:** Writing-original draft. **Ioannis Kompatsiaris:** Conceptualization. **Katharina Kinder-Kurlanda:** Methodology; resources; writing-original draft. **Claudia Wagner:** Resources; writing-original draft. **Fariba Karimi:** Resources; writing-original draft. **Miriam Fernandez:** Resources; writing-original draft. **Harith Alani:** Resources; writing-original draft; writing-review and editing. **Bettina Berendt:** Conceptualization; methodology; resources; writing-original draft; writing-review and editing. **Tina Kruegel:** Investigation; resources; writing-original draft; writing-review and editing. **Christian Heinze:** Resources; writing-original draft; writing-review and editing. **Klaus Broelemann:** Methodology; resources; writing-original draft; writing-review and editing. **Gjergji Kasneci:** Methodology; resources; writing-original draft; writing-review and editing. **Thanassis Tiropanis:** Methodology; resources; writing-original draft; writing-review and editing. **Steffen Staab:** Conceptualization; methodology; resources; writing-original draft; writing-review and editing.

ORCID

Eirini Ntoutsi https://orcid.org/0000-0001-5729-1003
Pavlos Fafalios https://orcid.org/0000-0003-2788-526X
Salvatore Ruggieri https://orcid.org/0000-0002-1917-6087
Franco Turini https://orcid.org/0000-0001-6789-5476
Symeon Papadopoulos https://orcid.org/0000-0002-5441-7341
Emmanouil Krasanakis https://orcid.org/0000-0002-3947-222X
Ioannis Kompatsiaris https://orcid.org/0000-0001-6447-9020
Katharina Kinder-Kurlanda https://orcid.org/0000-0002-7749-645X
Miriam Fernandez https://orcid.org/0000-0001-5939-4321
Bettina Berendt https://orcid.org/0000-0002-8003-3413

ENDNOTES

[1] Fei-Fei Li, Chief-Scientist for AI at Google and Professor at Stanford (http://fortune.com/longform/ai-bias-problem/).

[2] The legal discussion in this paper refers primarily to the EU situation. We acknowledge the difficulty of mapping the territory between AI and the law as well as the extra complexity that country-specific legislation brings upon and therefore, we believe this is one of the important areas for future work.

[3] The Web Science Manifesto—By Web Scientists. Building Web Science. For a Better World, https://www.webscience.org/manifesto/

[4] Amazon does not consider the race of its customers. Should It? https://www.bloomberg.com/graphics/2016-amazon-same-day/

[5] The General Data Protection Regulation (EU) 2016/679 (GDPR), https://gdpr-info.eu/

[6] https://cancer.sanger.ac.uk/cosmic

[7] http://geneontology.org/

[8] https://standards.ieee.org/project/7003.html

[9] https://www.wired.com/story/ideas-joi-ito-insurance-algorithms/

RELATED WIRES ARTICLES

Causability and explainability of artificial intelligence in medicine

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. M. (2018). A reductions approach to fair classification. *ICML*, *80*, 60–69.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*.

Atzmueller, M. (2017). Declarative aspects in explicative data mining for computational sensemaking. In *Declarative programming and knowledge management* (pp. 97–114). Springer.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039–1082.

Besnard, P., Cordier, M., & Moinard, Y. (2014). Arguments using ontological and causal knowledge. In *FoIKS Lecture notes in computer science* (Vol. 8367, pp. 79–96). Springer.

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS* (pp. 4349–4357).

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communications Society*, *15*(5), 662–679.

Broelemann, K., & Kasneci, G. (2019). A gradient-based split criterion for highly accurate and transparent model trees. In *IJCAI* (pp. 2030–2037). AAAI Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT, Proceedings of machine learning research* (Vol. 81, pp. 77–91). PMLR.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *ICDM workshops* (pp. 13–18). IEEE Computer Society.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292.

Calders, T., & Zliobaite, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society Studies in applied philosophy, epistemology and rational ethics* (Vol. 3, pp. 43–57). Springer.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *NIPS* (pp. 3992–4001).

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., & Smith, A. D. (2019). From soft classifiers to hard decisions: How fair can we be? In *FAT* (pp. 309–318). ACM.

Chandler, D., & Munday, R. (2011). *A dictionary of media and communication*. OUP Oxford.

Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. In *NIPS* (pp. 5029–5037).

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. F. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS* (pp. 3504–3512).

Cocarascu, O., & Toni, F. (2016). Argumentation for machine learning: A survey. In *COMMA Frontiers in artificial intelligence and applications* (Vol. 287, pp. 219–230). IOS Press.

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. arXiv preprint arXiv: 1808.00023.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Privacy Enhancing Technologies*, *2015*(1), 92–112.

Dehaspe, L., & Raedt, L. D. (1996). DLAB: A declarative language bias formalism. In *ISMIS Lecture notes in computer science* (Vol. 1079, pp. 613–622). Springer.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77*(2), 321–358.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. In *ITCS* (pp. 214–226). ACM.

Foster, S. R. (2004). Causation in antidiscrimination law: Beyond intent versus impact. *Houston Law Review*, *41*(5), 1469.

Garrido, P., Viñolas, N., Isla, D., Provencio, M., Majem, M., Artal, A., ... Felip, E. (2019). Lung cancer in spanish women: The WORLD07 project. *European Journal of Cancer Care*, *28*(1), e12941.

Giménez-García, J. M., Zimmermann, A., & Maret, P. (2017). NdFluents: An ontology for annotated statements with inference preservation. In *ESWC (1) Lecture notes in computer science* (Vol. 10249, pp. 638–654).

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *Lancet*, *359*, 248–252.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93:1–93:42.

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. *Common Market Law Review*, *55*(4), 1143–1185.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS* (pp. 3315–3323).

Henelius, A., Puolamäki, K., Boström, H., Asker, L., & Papapetrou, P. (2014). A peek into the black box: Exploring classifiers by randomization. *Data Mining and Knowledge Discovery*, *28*(5–6), 1503–1529.

Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems*, *32*(3), 68–73.

Hube, C., Fetahu, B., & Gadiraju, U. (2019). Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *CHI* (p. 407). ACM.

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In *FAT* (pp. 49–58). ACM.

Introna, L., & Wood, D. (2004). Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society*, *2*(2/3), 177–198.

Iosifidis, V., & Ntoutsi, E. (2019). AdaFair: Cumulative fairness adaptive boosting. In *CIKM*. ACM.

Jabbari, S., Joseph, M., Kearns, M. J., Morgenstern, J., & Roth, A. (2017). Fairness in reinforcement learning. In *ICML Proceedings of machine learning research* (Vol. 70, pp. 1617–1626). PMLR.

Kamar, E., Kapoor, A., & Horvitz, E. (2015). Identifying and accounting for task-dependent bias in crowdsourcing. In *HCOMP* (pp. 92–101). AAAI Press.

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. In *Computer, control and communication* (pp. 1–6). IEEE Computer Society.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *ICDM* (pp. 869–874). IEEE Computer Society.

Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2018). Exploiting reject option in classification for social discrimination control. *Information Sciences*, *425*, 18–33.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD (2) Lecture notes in computer science* (Vol. 7524, pp. 35–50). Springer.

Kaplan, R., Donovan, J., Hanson, L., & Matthews, J. (2019). *Algorithmic accountability: A primer*. Retrieved from https://datasociety.net/output/algorithmic-accountability-a-primer/

Karger, D. R., Oh, S., & Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. In *NIPS* (pp. 1953–1961).

Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. *Scientific Reports*, *8*.

Kasneci, G., & Gottron, T. (2016). LICON: A linear weighting scheme for the contribution of input variables in deep artificial neural networks. In *CIKM* (pp. 45–54). ACM.

Kilbertus, N., Gascon, A., Kusner, M., Veale, M., Gummadi, K., & Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. In *ICML* (pp. 2630–2639). PMLR.

Krasanakis, E., Xioufis, E. S., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *WWW* (pp. 853–862). ACM.

Krötzsch, M., Marx, M., Ozaki, A., & Thost, V. (2018). Attributed description logics: Reasoning on knowledge graphs. In *IJCAI* (pp. 5309–5313). AAAI Press.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *NIPS* (pp. 4066–4076).

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *COLING* (pp. 1384–1397). ACL.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *KDD* (pp. 1675–1684). ACM.

LeBlanc, E. C., Balduccini, M., & Vennekens, J. (2019). Explaining actual causation via reasoning about actions and change. In *JELIA Lecture notes in computer science* (Vol. 11468, pp. 231–246). Springer.

Li, X., & Huan, J. (2017). Constructivism learning: A learning paradigm for transparent predictive analytics. In *KDD* (pp. 285–294). ACM.

Liddell, R., & O'Flaherty, M. (2018). *Handbook on European non-discrimination law*. European Union Agency for Fundamental Rights (FRA).

Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). *Causal reasoning for algorithmic fairness*. CoRR, abs/1805.05859

Lorimer, T., Held, J., & Stoop, R. (2017). Clustering: How much bias do we need? *Philosophical Transactions of the Royal Society A*, *375*, 20160293.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *KDD* (pp. 623–631). ACM.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. In *KDD* (pp. 502–510). ACM.

Ma, R., Yu, Y., & Yue, X. (2015). Survey on image saliency detection methods. In *CyberC* (pp. 329–338). IEEE Computer Society.

Manzi, G., & Forster, M. (2019). Biases in bias elicitation. *Communications in Statistics – Theory and Methods*, 48(18), 4656-4674.

Mitchell, T. M. (1997). *Machine learning* (1st ed.). New York, NY: McGraw-Hill.

Mittelstadt, B. D., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *FAT* (pp. 279–288). ACM.

Montavon, G., Samek, W., & Müller, K. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased? Assessing the representativeness of twitter's streaming API. In *WWW (companion volume)* (pp. 555–556). ACM.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.

Ozaki, A., Krötzsch, M., & Rudolph, S. (2019). Temporally attributed description logics. In *Description logic, theory combination, and all that Lecture notes in computer science* (Vol. 11560, pp. 441–474). Springer.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. In *AAAI* (pp. 9780–9784). AAAI Press.

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. In *SDM* (pp. 581–592). SIAM.

Pitoura, E., Tsaparas, P., Flouris, G., Fundulaki, I., Papadakos, P., Abiteboul, S., & Weikum, G. (2017). On measuring bias in online information. *SIGMOD Record*, *46*(4), 16–21.

Raedt, L. D., Kersting, K., Natarajan, S., & Poole, D. (2016). Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *10*(2), 1–189.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *KDD* (pp. 1135–1144). ACM.

Samadi, S., Tantipongpipat, U. T., Morgenstern, J. H., Singh, M., & Vempala, S. (2018). The price of fair PCA: One extra dimension. In *NeurIPS* (pp. 10999–11010).

Schäfer, M., Haun, D. B., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological Science*, *26*(8), 1252–1260.

Serbos, D., Qi, S., Mamoulis, N., Pitoura, E., & Tsaparas, P. (2017). Fairness in package-to-group recommendations. In *WWW* (pp. 371–379). ACM.

Son, T. C., Pontelli, E., Gelfond, M., & Balduccini, M. (2016). Reasoning about truthfulness of agents using answer set programming. In *KR* (pp. 605–608). AAAI Press.

Song, C., Cheng, H., Yang, H., Li, S., Wu, C., Wu, Q., ... Li, H. (2018). MAT: A multi-strength adversarial training method to mitigate adversarial attacks. In *ISVLSI* (pp. 476–481). IEEE Computer Society.

Strang, T., Linnhoff-Popien, C., & Frank, K. (2003). Cool: A context ontology language to enable contextual interoperability. In *DAIS Lecture notes in computer science* (Vol. 2893, pp. 236–247). Springer.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM*. The AAAI Press.

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S., & Thompson, S. G. (2009). Bias modelling in evidence synthesis. *Journal of of the Royal Statistical Society A*, *172*(1), 21–47.

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349–391.

Veale, M., Binns, R., & Edwards, L. (2018). Algorithms that remember: Model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180083.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *FairWare@ICSE* (pp. 1–7). ACM.

Xu, D., Yuan, S., Zhang, L., & Wu, X. (2018). FairGAN: Fairness-aware generative adversarial networks. In *BigData* (pp. 570–575). IEEE.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. In *IJCAI* (pp. 5527–5533). AAAI Press.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW* (pp. 1171–1180). ACM.

Zhang, Q., Wu, Y. N., & Zhu, S. (2018). Interpretable convolutional neural networks. In *CVPR* (pp. 8827–8836). IEEE Computer Society.

Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089.

Zliobaite, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artifical Intelligence and Law*, *24*(2), 183–201.

Zuiderveen Borgesius, F. (2018). *Discrimination, artificial intelligence, and algorithmic decision-making*.

How to cite this article: Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining Knowl Discov*. 2020;10:e1356. https://doi.org/10.1002/widm.1356