Bias and Discrimination in AI: A Cross-Disciplinary Perspective

Xavier Ferrer, Tom van Nuenen, Jose M. Such, Mark Coté, and Natalia Criado
King's College London, London WC2R 2LS, U.K.

Operating at a large scale and impacting large groups of people, automated systems can make consequential and sometimes contestable decisions. Automated decisions can impact a range of phenomena, from credit scores to insurance payouts to health evaluations. These forms of automation can become problematic when they place certain groups or people at a systematic disadvantage. These are cases of discrimination—which is legally defined as the unfair or unequal treatment of an individual (or group) based on certain protected characteristics (also known as protected attributes) such as income, education, gender, or ethnicity. When the unfair treatment is caused by automated decisions, usually taken by intelligent agents or other AI-based systems, the topic of digital discrimination arises. Digital discrimination is prevalent in a diverse range of fields, such as in risk assessment systems for policing and credit scores [1], [2].

Digital discrimination is becoming a serious problem, as more and more decisions are delegated to systems increasingly based on artificial intelligence (AI) techniques such as machine learning. Although a significant amount of research has been undertaken from different disciplinary angles to understand this challenge—from computer science to law to sociology— none of these fields have been able to resolve the problem on their own terms. For instance, computational methods to verify and certify bias-free data sets and algorithms do not account for socio-cultural or ethical complexities, and do not distinguish between bias and discrimination. Both of these terms have a technical inflection, but are predicated on legal and ethical principles [3].

In this article, we propose a synergistic approach that allows us to explore bias and discrimination in AI by supplementing technical literature with social, legal, and ethical perspectives. Through a critical survey of a synthesis of related literature, we compare and evaluate the sometimes contradictory priorities within these fields, and discuss how disciplines might collaborate to resolve the problem. We also highlight a number of interdisciplinary challenges to attest and address discrimination in AI.

Bias and discrimination

Technical literature in the area of discrimination typically refers to the related issue of bias. Yet, despite playing an important role in discriminatory processes, bias does not necessarily lead to discrimination. Bias means a deviation from the standard, sometimes necessary to identify the existence of some statistical patterns in the data or language used [4], [5]. Classifying and finding differences between instances would be impossible without bias.

In this article, we follow the most common definition of bias used in the literature and focus on the *problematic* instances of bias that may lead to discrimination by AI-based automated-decision making systems. Three main, well-known causes for bias have been distinguished [4] as follows:

Digital Object Identifier 10.1109/MTS.2021.3056293
Date of current version: 3 June 2021.

0278-0097/21©2021 IEEE

- *Bias in modeling:* Bias may be deliberately introduced, e.g., through smoothing or regularization parameters to mitigate or compensate for bias in the data, which is called *algorithmic processing bias,* or introduced while modeling in cases with the usage of objective categories to make subjective judgments, which is called *algorithmic focus bias*.
- *Bias in training*: Algorithms learn to make decisions or predictions based on data sets that often contain past decisions. If a data set used for training purposes reflects existing prejudices, algorithms will very likely learn to make the same biased decisions. Moreover, if the data do not correctly represent the characteristics of different populations, representing an *unequal ground truth*, it may result in biased algorithmic decisions.
- *Bias in usage*: Algorithms can result in bias when they are used in a situation for which they were not intended. An algorithm utilized to predict a particular outcome in a given population can lead to inaccurate results when applied to a different population—a form of *transfer context bias*. Further, the potential misinterpretation of an algorithm's outputs can lead to biased actions through what is called *interpretation bias*.

A significant amount of literature focuses on forms of bias that may or may not lead to discriminatory outcomes, i.e., the relationship between bias and discrimination is not always clear or understood. Most literature assumes that systems free from biases do not discriminate, hence, reducing or eliminating biases reduces or eliminates the potential for discrimination. However, whether an algorithm can be considered discriminatory or not depends on the context in which it is being deployed and the task it is intended to perform. For instance, consider a possible case of algorithmic bias in usage, in which an algorithm is biased toward hiring young people. At first glance, it can be considered that the algorithm is discriminating against older people. However, this (biased) algorithm should only be considered to discriminate if the context in which it is intended to be deployed does not justify hiring more young people than older people. Therefore, statistically reductionist approaches, such as estimating the ratio between younger and older people hired, are insufficient to attest whether the algorithm is discriminating without considering this socially and politically fraught context; it remains ethically unclear where we need to draw the line between biased and discriminating outcomes. Therefore, AI and technical researchers often: 1) use discrimination and bias as equivalent or 2) focus on measuring biases without actually attending to the problem of whether or not there is discrimination. Our aim, in the below, is to disentangle some of these issues.

Measuring biases

To assess whether an algorithm is free from biases, there is a need to analyze the entirety of the algorithmic process. This entails first confirming that the algorithm's underlying assumptions and its modeling are not biased; second, that its training and test data do not include biases and prejudices; and finally, that it is adequate to make decisions for that specific context and task. More often than not, however, we do not have access to this information. A number of issues prevent such an analysis. The data used to train a model, for instance, is typically protected since it contains personal information, rendering the task of attesting training bias impossible. Access to the algorithm's source code might also be restricted to the general public, removing the possibility of identifying modeling biases. This is common as algorithms are valuable private assets of companies. Third, the specifics of where and how the algorithm will be deployed might be unknown to an auditor. Depending on what is available, different types of bias attesting might be possible, both in terms of the process and in terms of the metrics used to measure it.

Procedural versus relational approaches

We can distinguish between two general approaches to measure bias: 1) procedural approaches, which focus on identifying biases in the decision-making the process of an algorithm [6] and 2) relational approaches, which focus on identifying (and preventing) biased decisions in the data set or algorithmic output. Although ensuring unbiased outcomes is useful to attest whether a specific algorithm has a discriminatory impact on a population, focusing on the algorithmic process itself can help yield insights about the reason why it happened in the first place.

Procedural approaches focus on identifying biases in the algorithmic "logic." Such ante hoc interventions are hard to implement for two main reasons: 1) AI algorithms are often sophisticated and complex since, in addition to being trained on huge data sets, they usually make use of unsupervised learning structures that might prove difficult to trace and understand (e.g., neural networks) and 2) the source code of the algorithm is rarely available. Procedural approaches will become more beneficial with further progress in explainable AI [6].

Being able to understand the process behind an algorithmic discriminatory decision can help us understand possible problems in the algorithm's code and behavior, and thus act accordingly toward the creation of nondiscriminatory algorithms. As such, current literature on nondiscriminatory AI promotes the introduction of explanations into the model itself, e.g., through inherently interpretable models such as decision trees, association rules, causal reasoning, or counterfactual explanations which provide coarse approximations of how a system behaves by explaining the weights and relationships between variables in (a segment of) a model [7]–[11]. Note, however, that attesting that an algorithmic process is free from biases does not ensure a nondiscriminatory algorithmic output, since discrimination can arise as a consequence of biases in training or in usage [12].

While procedural approaches attend to the algorithmic process, relational approaches measure biases in the data set and the algorithmic output. Such approaches are popular in the literature, as they do not require insights into the algorithmic process. Besides evaluating biases in the data itself, where it is available (e.g., by looking at statistical parity), implementations can compare the algorithmic outcomes obtained by two different subpopulations in the data set [13], or make use of counterfactual or contrastive explanations [8], [11], [14], which have shown promising results in aiding the provision of interpretable models and make the decisions of inscrutable systems intelligible to developers and users, by asking questions such as "What if X instead of Y?"

Bias, here, is only located at testing time. One example is the post-hoc approach of local interpretable model-agnostic explanations (LIME), which makes use of adversarial learning to generate counterfactual explanations [6]. Other approaches evaluate the correlation between algorithmic inputs and biased outputs, to identify those features that may lead to biased actions that affect protected subpopulations [15]. Since implementations often ignore the context in which the algorithm will be deployed, the decision of whether a biased output results in a case of discrimination is often left to the user to assess [9].

Bias metrics

The metrics for measuring bias can be organized in three different categories: 1) statistical measures; 2) similarity-based measures; and 3) causal reasoning. While reviews such as [16] offer an extensive description of some of these metrics, we will discuss the intuition behind the most common types of metrics used in the literature below.

Statistical measures to attest biases represent the most intuitive notion of bias, and focus on exploring the relationships or associations between the algorithm's predicted outcome for the different (input) demographic distributions of subjects, and the actual outcome that is achieved. These measures include, first, *group fairness* (also named *statistical parity*), which requires that an equal quantity of each group of distinct individuals should receive each possible algorithmic outcome. For instance, if four out of five applicants of the advantaged group were given a mortgage, the same ratio of applicants from the protected group should obtain the mortgage as well. Second, *predictive parity* is satisfied if both protected and unprotected groups have equal positive predictive value—that is, the probability of an individual to be correctly classified as belonging to the positive class. Finally, the principle of *well calibration* states that the probability estimates provided by the decision-making algorithm should be properly adjusted with the real values. Despite the popularity of statistical metrics, it has been shown that statistical definitions are insufficient to estimate the absence of biases in algorithmic outcomes, as they often assume the availability of verified outcomes necessary to estimate them, and often ignore other attributes of the classified subject than the sensitiveones [17].

Similarity measures, on the other hand, focus on defining a similarity value between individuals. *Causal discrimination* is an example of such measures, stating that a classifier is not biased if it produces the same classification for any two subjects with the same nonprotected attributes. A more complex bias metric based on a similarity measure

between individuals is *fairness through awareness* [17], which states that, for fairness to hold, the distance between the distributions of outputs for individuals should *at most* be the distance between the two individuals as estimated by means of a similarity metric. The complexity in using this metric consists in accurately defining a similarity measure that correctly represents the complexity of the situation in question, which is often an impossible task to generalize. Moreover, the similarity measure between individuals can suffer from the implicit biases of the expert, resulting in a biased similarity estimator.

Finally, definitions based on causal reasoning assume bias can be attested by means of a directed causal graph. In the graph, attributes are presented as nodes joined by edges which, by means of equations, represent the relations between attributes [10]. By exploring the graph, the effects that the different protected attributes have on the algorithm's output can be assessed and analyzed. Causal fairness approaches are limited by the assumption that a valid causal graph able to describe the problem can be constructed, which is not always feasible due to the sometimes unknown and complex relations between attributes and the impact they have on the output.

Attesting and addressing discrimination

The first step explored in the related literature to identify discriminatory outputs is determining the groups whose algorithmic outputs are going to be compared. Technical approaches to select the subpopulations of interest vary, either: 1) they consider subpopulations as already defined [9], [18] or 2) they are selected by means of a heuristic that aggregates individuals that share one or more protected or proxy attributes (protected groups), as in *FairTest's* framework[1] for detecting biases in data sets. *Protected attributes* are encoded in legislation (see the "Legal perspective" section) and usually include attributes such as sex, gender, and ethnicity, while *proxy attributes* are attributes strongly correlated with protected attributes, e.g., weightlifting ability (strongly correlated with gender). However, the process of selecting individuals or groups based on these attributes is nontrivial since groups often result from the intersection of multiple protected and proxy attributes (see the "Social perspective" section).

Once the protected and the potentially advantaged groups have been selected, implementations apply different bias metrics (see the "Bias metrics" section) to compare and identify relevant differences in the algorithm's outcomes for the different groups. If these differences are a consequence of protected attributes, it is likely that the algorithm's decision can be considered discriminatory.

To alleviate the contextual problem of whether an algorithmic outcome may form a case of discrimination, approaches often incorporate explanatory attributes: attributes, such as gender or age, on which in specific contexts is deemed acceptable to differentiate, even if this leads to apparent discrimination on protected attributes [18]. Some relevant approaches are the open-source IBM AI Fairness 360 toolkit,[2] which contains techniques developed by IBM and the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle, and Google's Whatif-tool,[3] which offers an interactive visual interface that allows researchers to investigate model performances for a range of features in the data set and optimization strategies.

Despite these efforts in parameterizing context uncertainty in technical implementations, the interpretive dimension that separates bias and discrimination remains a challenge. As a response, some approaches base their implementations on various antidiscrimination laws that focus on the relationships between protected attributes and decision outcomes. For instance, the U.S. *fourth-fifth court rule* and the *Castaneda rule* are used as a general, and often arguably adequate, *prima facie* evidence of discrimination—see the "Legal perspective" section for more details on these rules.

Approaches that intervene on problematic biases focus on: 1) removing protected attributes from the data, as an attempt to impede the algorithm from using these protected attributes to make discriminatory decisions (*fairness through blindness* [12], [17]), or on 2) debiasing algorithms' outputs [19]. An issue here is that removing protected attributes from the input data often results in a significant loss of accuracy in the algorithm [17]. Moreover, excluded attributes can often be correlated with *proxy attributes* that remain in the data set, meaning bias may still be present (i.e., certain residential

[1]https://github.com/columbia/fairtest

[2]https://github.com/IBM/AIF360

[3]https://pair-code.github.io/what-if-tool/

areas have specific demographics that play the role of proxy variables for ethnicity). These approaches can also be criticized because they alter the model of the world that an AI makes use of, instead of altering how that AI perceives and acts on bias [17].

On a broader level, debiasing an algorithm's output requires a specific definition of its context and, as such, is difficult to achieve from a technical perspective only. A myriad of lingering questions remains to be answered: how *much* bias does an algorithm need to encode to consider its outputs discriminating? How can we reflect on the peculiarity of the data on which these algorithms are operating—data which often reflects the inequities of its time? In short, a clearer definition of the relation between algorithmic biases and discrimination is needed. We argue that such a definition can only be provided by a cross-disciplinary approach that takes legal, social, and ethical considerations into account. In response, in the next sections, we will engage critically with related work from legal, social, and ethical perspectives.

Legal perspective

Legislation designed to prevent discrimination against particular groups of people that share one or more protected attributes—namely *protected groups*—receives the name of antidiscrimination law. Antidiscrimination laws vary across countries. For instance, European antidiscrimination legislation is organized in directives, such as Directive 2000/43/EC against discrimination on grounds of race and ethnic origin, or Chapter 3 of the EU Charter of fundamental rights. Antidiscrimination laws in the US are described in the *Title VII of the Civil Rights Act of 1964* and other federal and state statutes, supplemented by court decisions. For instance, Title VII prohibits discrimination in employment on the basis of race, sex, national origin, and religion; and *The Equal Pay Act* prohibits wage disparity based on sex by employers and unions.

The main issues in trials related to discrimination consist of determining [20]: 1) the relevant population affected by the discrimination case, and to which groups it should be compared; 2) the discrimination measure that formalizes group under-representation, e.g., *disparate treatment* or *disparate impact* [18], [21]; and 3) the threshold that constitutes prima facie evidence of discrimination. Note that the three issues coincide with the problems explored in the technical approaches presented earlier. With respect to the last point, no strict threshold has been laid down by the European Union. In the U.S., the *fourth–fifth rule* from the Equal Employment Opportunity Commission (1978), which states that a job selection rate for the protected group of less than 4/5 of the selection rate for the unprotected group, is sometimes used a prima facie evidence of an adverse impact. The *Castaneda rule*, which states that the number of people of the protected group selected from a relevant population cannot be smaller than 3 standard deviations the number expected in a random selection, is also used [21]. Although such laws can relieve discriminatory issues, more complex scenarios can arise. For instance, Hildebrandt and Koops [22] mention the legally gray area of price discrimination, where consumers in different geographical areas can be offered different prices based on differences in average income.

More recent regulations, such as the general data protection regulation (GDPR), have been offered as a framework to alleviate some of the enforcement problems of antidiscrimination law, and include clauses on automated decision-making related to procedural regularity and accountability, introducing a right of explanation for all individuals to obtain meaningful explanations of the logic involved when automated decision making takes place. However, these solutions often assume white-box scenarios, which, as we have seen, may be difficult to achieve technically, and even when they are achieved, they may not necessarily provide the answers sought to assess whether discrimination is present or not. Generally, current laws are badly equipped to address algorithmic discrimination [21]. Leese [23], for instance, notes that antidiscrimination frameworks typically follow the establishment of a causal chain between indicators on the theoretical level (e.g., sex or race) and their representation in the population under scrutiny. Data-driven analytics, however, create aggregates of individual profiles, and as such are prone to the production of arbitrary categories instead of real communities. As such, even if data subjects are granted procedural and relational explanations, the question remains at which point potential biases can reasonably be considered forms of discrimination.

Social perspective

Digital discrimination is not only a technical phenomenon regulated by law but one that also needs to be considered from a socio-cultural perspective to be rigorously understood. Defining what constitutes discrimination is a matter of understanding the particular social and historical conditions and ideas that inform it, and needs to be reevaluated according to its implementation context. Bias in usage, as defined above, forms a challenge to any kind of generalist AI solution.

One complication highlighted by a social perspective is the potential of digital discrimination to reinforce existing social inequalities. This point becomes increasingly pressing when multiple identities and experiences of exclusion and subordination start interacting—a phenomenon called intersectionality [24]. One example is formed by the multiple ways that race and gender interact with class in the labor market, effectively generating new identity categories. From a legislation perspective, antidiscrimination laws can be applied when discrimination is experienced by a population that shares one or more protected attributes. However, this problem can exponentially grow in complexity when also considering proxy variables and the intersection of different features [15].

On a cultural and ideological level, the call for ever-expanding transparency of AI systems needs to be seen as an *ideal* as much as a form of "truth production" [25]. Furthermore, no standard evaluation methodology exists among AI researchers to ethically assess their bias classifications, as the explanation of classification serves different functions in different contexts [14], and is arguably assessed differently by different people (e.g., the way a data set is defined and curated, for instance, depends on the assumptions and values of the creator) [26]. Conducting a set of experimental studies to elicit people's responses to a range of algorithmic decision scenarios and explanations of these decisions, Binns et al. [27] find a strong split in their respondents: some find the general idea of algorithmic discrimination immoral, others resist imputing morality to a computer system altogether "*the computer is just doing its job*" [27].

Although algorithmic decision-making implicates dimensions of justice, its claim to objectivity may also preclude the public awareness of these dimensions. Given the differing stances on discrimination in society, providing explanations to the public targeted by algorithmic decision-making systems is key, as it allows individuals to make up their own minds about their evaluations of these systems. Hildebrand and Koops [22], for instance, call for smart transparency by designing the socio-technical infrastructures responsible for decision-making in a way that allows individuals to anticipate and respond to how they are profiled. In this context of public evaluation, it also becomes important to question which moral standards can or should be encoded in AI, and which considerations of discrimination can be expected to be most readily shared by a widely differing range of citizens [28]. Although such frameworks can always be criticized as reductionist approaches to the complexity of social values, keeping into account what kinds of values are important in society can go some way in helping to establish *how* discrimination can be defined.

Ethical perspective

Finally, we need to bring in an ethical perspective; as Tasioulas argues, discrimination does not need to be unlawful to be unfair [29]. Yet, moral standards are historically dynamic, and continuously evolving due to technological developments. This explains why law and encoded social morality often lag behind technical developments. In light of the discriminatory risks (and benefits) that AI might pose, moral standards need to be reassessed to enable new definitions of discriminatory impact. It says that one of the famous attempts to address this question in robotics derives from fiction: Isaac Asimov's Three Laws of Robotics. More recently, the AI community has attempted to codify ethical principles for AI, such as the Asilomar AI Principles.[4] However, these principles are criticized as being vague, mainly due to their level of abstraction, making them not necessarily helpful [29].

More grounded and detailed frameworks for AI ethics have recently been proposed, such as the standards being defined by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,[5] which aim to provide an incubation space for new solutions relevant to the ethical implementation of intelligent technologies. Another noteworthy

[4]https://futureoflife.org/ai-principles/

[5]https://ethicsinaction.ieee.org/

contribution is presented in [29], stating that the ethical questions related to the usage of AI can be organized into three interconnected levels. The first level involves laws to govern AI-related activities, including public standards backed up by public institutions and enforcement mechanisms, which claim to be morally binding on all citizens in virtue of their formal enactment. Some efforts discussed in the "Legal perspective" section can be seen as examples of this. However, this evades the problem that not all of the socially entrenched standards that govern our lives are legal standards. We rely not only on the law to discourage people from wrongful behavior, but also on moral standards that are instilled in us from childhood and reinforced by society.

The second level is the social morality around AI. The definition of such a morality is problematic as it involves a potential infinity of reference points, as well as the cultivation of emotional responses such as guilt, indignation, and empathy—both of which are effects of human consciousness and cognition [29]. The third and final level includes individuals and their engagement with AI. Individuals and associations will still need to exercise their own moral judgment by, for instance, devising their own codes of practice. However, how these levels can be operationalized (or to what extent) from a technical AI point of view is not yet clear.

Open challenges

Addressing and attesting digital discrimination and remedying its corresponding deficiencies will remain a problem for technical, legal, social, and ethical reasons. Technically, there are a number of practical limits to what can be accomplished, particularly regarding the ability to automatically determine the relationship between biases and discrimination and how to translate the social realities into machine code. Current legislation is poorly equipped to address the classificatory complexities arising from algorithmic discrimination. Social inequalities and differing attitudes toward computation further obfuscate the distinction between bias and discrimination. From an ethical perspective, existing moral standards need to be reassessed and frequently updated in light of the risks and benefits AI might pose.

In sum, the design and evaluation of AI systems are rooted in different perspectives, concerns, and goals (see Table 1). To posit the existence of a predefined path through these perspectives would be misleading. What is needed, instead, is a sensitivity to the distinctions concerning what is desirable AI implementation, and to a dialogical orientation toward design processes. Finding solutions to discrimination in AI requires robust cross-disciplinary collaborations. We conclude here by summarizing what we believe to be some of the most important cross-disciplinary challenges to advance research and solutions for attesting and avoiding discrimination in AI.

Table 1. Summary of challenges for the different perspectives.

Perspective	Challenges/Limitations
Technical	Relationship between bias and discrimination is difficult to determine and generalise solely from a technical perspective.
Legal	Legislation is poorly equipped to address the classification complexities arising from algorithmic discrimination.
Social	Differing attitudes towards computation and literacy obfuscate the distinction between bias and discrimination.
Ethical	Existing moral standards need to be reassessed in light of the risks and benefits AI might pose.

How much bias is too much?

Whether a biased decision can be considered discriminatory or not depends on many factors, such as the context in which AI is going to be deployed, the groups compared in the decision, and other factors like a tradeoff between individualist- meritocratic and outcome-egalitarian values. To simplify these problems, technical implementations tend to borrow definitions from the legal literature, such as the thresholds that constitute prima facie evidence of discrimination, and use it as a general rule to attest algorithmic discrimination. Yet this cannot be addressed by simply encoding the legal, social and ethical context, which in and of itself is nontrivial. Bias and discrimination have a different ontological status: while the former may seem easy to define in terms of programmatic solutions, the latter involves a host of social and ethical issues that are challenging to resolve from a positivist framework.

Critical AI literacy

Another challenge is the need for an improvement in critical AI literacy. We have noted the need to take into account the end-user of AI decision making systems, and the extent to which their literacy of these systems can be targeted and improved. In part, this entails end-user knowledge

of particularities such as the attributes being used in a data set, as well as the ability to compare explanation decisions and moral rules underlying those choices. This is, however, not solely a technical exercise, as decision-making systems render end-users into algorithmically constructed data subjects. This challenge could be addressed through a socio-technical approach that can consider both the technical dimensions and the complex social contexts in which these systems are deployed. Building public confidence and greater democratic participation in AI systems requires ongoing development of not just explainable AI but of better Human–AI interaction methods and sociotechnical platforms, tools, and public engagement to increase critical public understanding and agency.

Discrimination-aware AI

Third, AI should not just be seen as a potential problem causing discrimination, but also as a great opportunity to mitigate existing issues. The fact that AI can pick up on discrimination suggests it can be made *aware* of it. For instance, AI could help spot digital forms of discrimination, and assist in acting upon it. For this aim to become a reality we would need, as explored in this work, a better understanding of social, ethical, and legal principles, as well as dialogically constructed solutions in which this knowledge is incorporated into AI systems. Two ways to achieve this goal are: 1) using data-driven approaches like machine learning to actually look at previous cases of discrimination and try to spot them in the future and 2) using model-based and knowledge-based AI that operationalizes the socio-ethical and legal principles mentioned above (e.g., normative approaches that include nondiscrimination norms as part of the knowledge of an AI system to influence its decision making). This would, for instance, facilitate an AI system realizing that the knowledge it gathered or learned is resulting in discriminatory decisions when deployed in specific contexts. Hence, the AI system could alert an expert human about this, and/or proactively address the issue spotted. ■

Acknowledgments

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/R033188/1. It is part of the Discovering and Attesting Digital Discrimination (DADD) project—see https://dadd-project.org.

References

[1] C. O'Neil, *Weapons of Math Destruction: How Big Data increases Inequality and Threatens Democracy*. Portland, OR, USA: Broadway Books, 2017.

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD),* 2008, pp. 560–568.

[3] N. Criado and J. M. Such, "Digital discrimination," in *Algorithmic Regulation*. Oxford, U.K.: OUP, 2019.

[4] D. Danks and A. London, "Algorithmic bias in autonomous systems," in *Proc. IJCAI*, 2017, pp. 4691–4697.

[5] X. Ferrer et al., "Discovering and categorising language biases in reddit," in *Proc. Int. AAAI Conf. Web Social Media (ICWSM*), 2020.

[6] S. T. Mueller et al., "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI," Feb. 2019, *arXiv:1902.01876*. [Online]. Available: http://arxiv.org/abs/1902.01876

[7] R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Comput. Surv*., vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018.

[8] R. Guidotti et al., "Factual and counterfactual explanations for black box decision making," *IEEE Intell*. Syst., vol. 34, no. 6, pp. 14–23, Dec. 2019.

[9] S. Ruggieri, D. Pedreschi, and F. Turini, "Integrating induction and deduction for finding evidence of discrimination," *AI Law*, vol. 18, no. 1, pp. 1–43, 2010.

[10] N. Kilbertus et al., "Avoiding discrimination through causal reasoning," in *Proc*. NIPS, 2017, pp. 656–666.

[11] R. M. Byrne, "Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning," in *Proc. IJCAI*, 2019, pp. 6276–6282.

[12] T. Calders and I. Zliobaite, "Why unbiased computational processes can lead to discriminative decision procedures," in *Discrimination and Privacy in the Information Society*. Berlin, Germany: Springer, 2013, pp. 43–57.

[13] N. Criado, X. Ferrer, and J. M. Such, "A normative approach to attest digital discrimination," in *Proc. Advancing Towards SDGS Artif. Intell. Fair, Just Equitable World Workshop 24th Eur. Conf. Artif. Intell. (ECAI),* 2020.

[14] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell*., vol. 267, pp. 1–38, Feb. 2019.

[15] N. Grgic-Hlac et al., "Beyond distributive fairness in algorithmic decision making," in *Proc. AAAI*, 2018, pp. 51–60. [Online]. Available: https://people.mpi-sws.org/nghlaca/papers/fair_feature_selection.pdf

[16] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. IEEE/ACM FairWare*, May 2018, pp. 1–7.

[17] C. Dwork et al., "Fairness through awareness," in *Proc. ITCS*. New York, NY, USA: ACM, 2012, pp. 214–226.

[18] M. Feldman et al., "Certifying and removing disparate impact," in *Proc. ACM-SIGKDD*. New York, NY, USA: ACM, 2015, pp. 259–268.

[19] T. Bolukbasi et al., "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4349–4357.

[20] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *Knowl. Eng. Rev.*, vol. 29, no. 5, 2014, Art. no. 582638.

[21] S. Barocas and A. Selbst, "Big data's disparate impact," *California Law Rev.*, vol. 104, no. 3, pp. 671–729, Jun. 2016. [Online]. Available: https://ssrn.com/abstract=2477899

[22] M. Hildebrandt and B.-J. Koops, "The challenges of ambient law and legal protection in the profiling era," *Mod. Law Rev.*, vol. 73, no. 3, pp. 428–460, 2010.

[23] M. Leese, "The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European union," *Secur. Dialogue*, vol. 45, no. 5, pp. 494–511, Oct. 2014.

[24] S. Walby, J. Armstrong, and S. Strid, "Intersectionality: Multiple inequalities in social theory," *Sociology*, vol. 46, no. 2, pp. 224–240, Apr. 2012.

[25] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media Soc.*, vol. 20, no. 3, pp. 973–989, Mar. 2018.

[26] T. van Nuenen et al., "Transparency for whom? Assessing discriminatory artificial intelligence," *Computer*, vol. 53, no. 11, pp. 36–44, Nov. 2020.

[27] R. Binns et al., "'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions," in *Proc.* CHI. New York, NY, USA: ACM, 2018, p. 377.

[28] O. S. Curry, D. A. Mullins, and H. Whitehouse, "Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies," *Current Anthropol.*, vol. 60, no. 1, pp. 47–69, Feb. 2019.

[29] J. Tasioulas, "First steps towards an ethics of robots and artificial intelligence," *J. Practical Ethics*, vol. 7, no. 1, 2019.

Xavier Ferrer received the Ph.D. degree in informatics from the Artificial Intelligence Institute, Spanish National Research Council (IIIA-CSIC), Barcelona, Spain, and the Universitat Autonoma de Barcelona (UAB), Barcelona, in 2017.

He is currently a Research Associate in digital discrimination with the Department of Informatics, King's College London, London, U.K. His research interests are related to natural language processing, machine learning, and fairness.

Tom van Nuenen received the Ph.D. degree in cultural studies from Tilburg University, Tilburg, Netherlands, in 2016.

He is currently a Research Associate with the Department of Informatics, King's College London, London, U.K., and a Visiting Scholar in digital humanities with the University of California at Berkeley, Berkeley, CA, USA. He focuses on the use of mixed methods to identify and solve social, ethical, and development questions related to big data, artificial intelligence (AI), and machine learning.

Jose M. Such is currently a Reader (Associate Professor) with the Department of Informatics and the Director of the KCL Cybersecurity Centre, King's College London, London, U.K. He has been a Principal Investigator for a number of large projects funded by EPSRC, including the Discovering and Attesting Digital Discrimination (DADD) Project and the Secure AI Assistants (SAIS) Project. His research interests are at the intersection of artificial intelligence (AI), human–computer interaction, and cybersecurity, with a strong focus on human-centered AI security, ethics, and privacy.

Mark Coté is currently a Senior Lecturer in data culture and society with the Department of Digital Humanities, King's College London, London, U.K. He is a Principal Investigator (PI) and a Co-Principal Investigator (CI) on a range of H2020 and UKRI grants, namely, through the Engineering and Physical Science Research Council and the Arts and Humanities Research Council. He is involved in research on critical interdisciplinary methods focusing on the social, cultural, and political–economic dimensions of big data, algorithms, and machine learning.

Natalia Criado is currently a Senior Lecturer in computer science with the Department of Informatics and the Co-Director of the UKRI Centre for Doctoral Training in Safe and Trusted AI, King's College London, London, U.K. Her research interests are computational norms and normative multiagent systems, and the application of multiagent systems, data science, and artificial intelligence to enhance cybersecurity and privacy.

■ Direct questions and comments about this article to Xavier Ferrer, King's College London, London WC2R 2LS, U.K.; xavier.ferrer_aran@kcl.ac.uk.