Encyclopedia of Law and Data Science

Algorithmic discrimination

Bettina Berendt, bettina.berendt@kuleuven.be, 13 October 2020

All comments welcome! - Thanks!

Terms in blue are references to other entries in the same encyclopedia in addition to those that I have referenced via "see book entry ...". I have now followed the Encyclopedia entries you sent me and referenced other entries that I have read as "see book entry ...". For the other ones, I think but cannot be sure that the cross-reference will help the reader. Please use or re-format the blue terms as you see fit!

Passages in green are changes relative to the draft version that you commented on.They include reactions to your comments and also some minor phrasing changes.

"Algorithmic discrimination" (AD) can be defined as discrimination in contexts that involve (usually digital) computers. It can be a result of bias in data, algorithms, and the socio-technical systems in which these are used. AD produced by data-mining and machine-learning algorithms, and measures for mitigating it, were first described formally in 2008, and many real-life examples have been identified. AD has become one of the most-studied ethical/legal problems of (semi-)automated decision making since the 2010s, in data science (the field of computer science / artificial intelligence dealing with big data, data mining, and machine learning) and adjoining fields.

Examples

A computer program with rules based on data gathered from prior admissions decisions, which was used in the initial screening of applicants for a medical school in London, was found to systematically discriminate against women and people with non-European sounding names (Lowry and Macpherson, 1988). In an experimental study of simulated Web users that differed only by gender, Datta, Tschantz, and Datta (2015) showed that when men and women visit Web pages associated with employment, the search engine showed high-salary job advertisements more often to the (simulated) men than to their female counterparts. The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool for predicting the risk of recidivism of pre-trial inmates was found to incorrectly judge black defendants as 'high risks' far more often than white defendants, and white defendants to be incorrectly judged as 'low risks' far more often than black defendants. (Angwin *et al.*, 2016). These studies and many others have sharpened the public perception that algorithms are far from the 'neutral', 'objective' tools that help decision-makers overcome their human prejudices and biases and help all members of society to attain equal opportunities. On the other hand, it has been contested whether certain results indeed signal 'unfairness' (e.g., Chouldechova, 2017), thus, by extension, whether the consequent treatment constitutes 'discrimination'.

Relevance (especially from a legal perspective)

AD is, by its definition, a legally relevant phenomenon: it may violate right(s) to non-discrimination and thereby presents a human-rights and/or fundamental-rights challenge. In addition, the treatment may violate other fundamental rights (such as privacy or data protection) . Obviously, the discriminated-against individuals and groups are harmed, but others and society as a whole may suffer too by the undermining of democratic principles or from economic, public-health, public-

security, and other consequences of the discrimination (e.g., Eubanks, 2018). Thus, individuals and institutions that deploy computers, whether they are public or private, have moral and legal obligations to prevent or at least mitigate AD arising from decisions and actions they are responsible for, and to provide safeguards against AD.

These obligations imply that the development, deployment and continuing monitoring/evaluation of countermeasures against AD are likely to become legally mandated and thus also commercially and politically highly relevant. The development of research, development, and legislation regarding protections against violations of the rights to privacy and data protection can be regarded as a precedent. For example, the EU General Data Protection Regulation (GDPR)[1] recognises, many threats to these rights that have become clear(er) through the widespread use of computers, and it mandates that data controllers and processors deploy a range of protections against these threats. The GDPR already recognises, albeit in rather general terms, the threat of AD, and it declares an obligation to prevent AD. The fact that discriminatory effects are mentioned explicitly 'only' in recitals (Recitals 71, 75 and 85) highlights both the importance of the phenomenon (and its recognition by lawmakers) and the need for more specific legislation.

Such legislation is expected to comprise different laws that are currently being developed or revised. One area is the (emerging or desired) regulation of Artificial Intelligence, cf. the EU White Paper on Artificial Intelligence (2020). However, algorithms that deeply affect society (and that may be associated with AD) need not be AI (Datenethikkommission, 2020); therefore, laws such as those on liability, envisaged for the future to more clearly cover computer-related artefacts and consequences (European Commission Expert Group on Liability, 2019, for details see book entry *liability*), need to become 'AD-aware' as well. Other impulses may come from data protection legislation that governs law enforcement: in contrast to the GDPR (with which it otherwise largely overlaps) the European Directive 2016/680 ("Law Enforcement Directive") forbids, in Art. 11 (3), "profiling that results in discrimination against natural persons on the basis of special categories of personal data" (see Naudts, 2019, for an analysis of the Law Enforcement Directive with respect to non-discrimination).

History

When thinking of an algorithm as a decision rule, and of the origin of such decision rules in large volumes of data, one quickly realises that the problem predates AI, big data and even computer-aided decision making. One example is the long tradition of using actuarial factors related to gender/sex in the provision of insurance and other financial services. Results included higher premiums for women in health insurance and life-annuity products, and higher premiums for men in driving and life insurance. Another example is the "redlining" practice of estimating the level of security for real-estate investments by residential neighbourhoods. Instituted in the US since the 1930s, this resulted in people living in certain areas of cities (marked by surveyors in red on city maps) being faced with higher costs and/or less availability of loans – and the "red" areas were predominantly black neighbourhoods. Both practices have received (more) notoriety by eventually being declared discriminatory and illegal, by the Fair Housing Act of 1968 and the Community Reinvestment Act of 1977 for the US, and by the *Association belge des Consommateurs ASBL v Conseil des ministres* [2011] judgement for the European Union. However, less overt examples of "redlining", via proxy variables ultimately tied to well-known grounds such as gender or ethnicity, or via new demographic categories, persist (e.g. "weblining", Andrews, 2012, see also Challenges and Future Work below).

1 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC

With increasing computational power and use for decision-making in the private and public sectors, the issue became a topic for scientific investigation also in the data science and wider information-systems fields. Three articles, selected for their contribution and influence, will be briefly described in turn. One was the first to systematically examine different ways in which bias can enter computer systems, the second provided the first formal analysis and mitigation strategies for data science, and the third provided an in-depth analysis of how big-data processing can have disparate impact in the legal sense. (While 'disparate impact' is a notion used especially in US law, for the purposes of AD analysis it is sufficiently comparable to 'indirect discrimination', which is more common in EU law. Disparate impact is probably a more frequent problem given today's heightened social resistance against disparate treatment / direct discrimination. In addition, it is certainly a more difficult problem to address, both in a social and in a computational sense.)

Friedman and Nissenbaum (1996) presented a framework of three categories of **bias in computer systems**, based on an analysis of case studies. They "use the term bias to refer to computer systems that *systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others" (p. 332). On the one hand, their definition overlaps with our legally-inspired definition: "A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate" (Friedman and Nissenbaum, 1996, p. 332). On the other hand, this definition allows for grounds of the differential treatment to be any criteria, ranging from ethnicity to "effecting a long/resource-intensive computing job in a multi-user computer system". Thus, whether the differentiation is unfair (in a moral sense) and whether it constitutes discrimination (in a legal sense) is a question delegated to another level of definitions.

The three proposed categories of bias are: "Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use." (p. 330)

Friedman's and Nissenbaum's conceptualisation of bias and discrimination remains central in the large majority of later computer-science literature. However, the problem has become much more apparent recently, based on two major developments in computing since the decade after their paper: big data and AI/data mining/machine learning. Two landmark papers investigated the (interlinked) effects of data mining and big data. They will be presented briefly here; the next section will present more details on how important mechanisms have changed through which pre-existing bias enters computing systems, technological constraints create technical bias, and unintended or unpredicted/unpredictable biases emerge in use.

Pedreschi, Ruggieri and Turini (2008) introduced the notion of "**discrimination-aware data mining**". They identified two essential tasks that need to be performed to mitigate AD via algorithms. They proposed several metrics to determine whether a set of patterns mined from data (in their case classification/prediction rules) showed evidence of biased data (*discrimination discovery* task) and whether the use of these patterns for prediction would lead to AD via algorithms (*discrimination prevention* task). Their metrics were inspired by heuristics that have long been used also in legal contexts (such as the 80% rule, Biddle, 2005; see Pedreschi, Ruggieri and Turini, 2012), and, like these, by measures of difference from statistical science and of the "interestingness" of patterns in data mining. The approach also includes a formal definition as well as the discovery and prevention of indirect discrimination. The methods comprise the testing of patterns with regard to whether the metrics exceed a certain threshold (discovery) and the removal of such patterns from further use (prevention).

The book edited by Custers, Calders, Schermer and Zarsky (2013) collected a multidisciplinary set of articles that surveyed important work in and around discrimination-aware data mining, which had grown substantially in various European countries. Concurrently, US scholars began to formalise and investigate 'fairness-aware data mining' (Dwork *et al.*, 2012).

Indirect discrimination as operationalised by Pedreschi *et al.* (2008) occurs, for example, when an algorithm decides based on ZIP code, but this is essentially a proxy for 'being black' (or not). This proxy may have been learned by a data-mining algorithm, and the result is one of several ways in which a decision rule or a computing system implementing such rules may seem neutral but end up adversely affecting one group of people more than another. The resulting discrimination is often unintentional, but no less real. "**Big data's disparate impact**" (Barocas and Selbst, 2016) is an apt summary expression for such effects and at the same time a dismantling of one of the core promises of big data: the promise of objectivity and neutrality. The essay analyses the concepts through the lens of US law against employment-related discrimination (Title VII). It investigates the many facets in which data-related activities can create or contribute to discrimination: from the definitions of the target variables and class labels, via the collection of training data and the labelling of examples, to feature selection. It studies the problems of proxies and masking as well as proving discrimination. The authors sketch remedial action and also the difficulties facing efforts to reforming these data- and mining-related activities. They consider external difficulties, including political choices between a focus on procedural protections against 'unfair decisions' (in US terms: anticlassification) versus a focus on enabling substantial equality (in US terms: antisubordination).

In parallel with this growing theoretical understanding, in the second half of the 2010s a growing number of case studies began to be discussed also in news media and politics. "Algorithmic discrimination" (AD) became a key spectre haunting public trust in real-life environments in which increasing numbers of functions with a clear societal impact are being performed with the help of algorithms. It is related in many ways to algorithms-related "privacy violations" as a key spectre predating it, and now continuing in parallel with it. AD can result from (properly or improperly) revealed personal data and may therefore result from violations of the right to **privacy** and/or the right to **data protection**. (Common computer-science terminology and legal terminology do not always match: while data protection and privacy are related but not the same in EU law and jurisprudence, the commonly evoked US laws talk only about 'privacy'. Computer science tends to use this term as a cover-all term.) The legal structure and, increasingly, mode of operation, of anti-discrimination and data protection law in the EU are similar and/or complementary (Gellert *et al.*, 2013). Data-science mitigation strategies against AD share formal characteristics with mitigation strategies against privacy violations, although differences and trade-offs exist (Hajian *et al.*, 2015). And both problems are examples of the "ethical" (and legal) problems that may be caused by artificial intelligence and other complex algorithms that support or make decisions with significant impacts on individuals, groups and society.

Still, protection against AD involves more than data protection. AD can also result from the processing of personal data without violations of data protection law: in *Heinz Huber v Germany* [2008], the European Court of Justice ruled that storing and accessing certain personal data only about some people, in this case nationals of selected member states of the EU, can constitute discrimination, in this case by violating the principle of non-discrimination on the basis of (EU) nationality. AD can also result from the processing of personal data without the result being 'traditional' discrimination along the lines of more or less stable group memberships: as Vedder (2018, p.2) has argued, adverse treatment of persons "on the basis of a group characteristic that is unknown to themselves or […] is dependent on their […] circumstances rather than on themselves in their own right may […] go against the individuality of persons as a fundamental value in its own right". We will return to this question in the "Challenges and Future Work" section.

In addition, the processing of non-personal data may result in systematic adverse impacts on certain groups of people, for example on those who live in locations that – via algorithmic decisions – become traffic hotspots (see case study 1 in Kulynych *et al.*, 2020, for an analysis of such effects of routing algorithms, and a mitigation strategy). 'Living in a certain area' can be, as discussed with reference to redlining above, a proxy for a legally protected ground; one could in addition ask whether it is an identity-shaping attribute in its own right, such that the systematic adverse effect may constitute discrimination.

The concept(s) of AD (especially from a computer-science and socio-technical systems perspective)

AD as a term needs further differentiation.

Discrimination (in an epistemic sense) is an essential feature of many algorithms. This is most obvious in typical classification, regression and ranking algorithms such as those mentioned in the examples above. A search engine or other recommender system distinguishes, via the algorithm(s) it uses, **between informational items** deemed relevant or irrelevant for the current information need or user. It sorts – and thus classifies – items such as Web pages by some measure of fit, importance or other quality criterion. A risk assessment system classifies 'people-representing items' into risk groups.

An algorithm may behave differently for specific groups of items. If more undesirable outcomes systematically result for a certain population group G, the algorithm can be said to exhibit **algorithmic bias** (AB) against G. Typical biases are predictions of more undesirable outcomes (as in educational test scores or recommendations of bank loans with worse conditions for G, fewer recommendations to hire members of G, see examples in Barocas and Selbst, 2016; Hutchinson and Mitchell, 2019) or higher error rates in predictions for G (e.g. in facial-recognition software, see Buolamwini and Gebru, 2018). The term AB is not only used with regards to people, but also other systematic 'preferences' of algorithms, e.g. for or against news with a certain political perspective (Hamborg, Donnay and Gipp, 2019). For further meanings of "bias", see also book entry *bias*.

The concept of AB is problematic for four reasons. The first reason arises from the dual nature of big-data algorithms (and actuarial and other statistical methods before them) as **descriptive and predictive/prescriptive**. Decision algorithms learned from data by learning algorithms are born *descriptive*. An algorithm that classifies women to be less tall then men, is not biased but reveals a biological regularity; an algorithm that classifies women as earning less money than men, may reflect a biased world but is not in itself biased. Such an algorithm can however easily become biased when – as is common for data analytics – it is re-purposed in a *predictive* and/or *prescriptive* way. An algorithm that predicts that female employees are likely to be 'worth' less than males, that proposes to pay female employees less than male employees, would be regarded as biased. AB is therefore not a property of the algorithm itself, but of the way it is used.

A second complication arises from **biased data**, a term that is used to denote a range of phenomena, which are not mutually exclusive. (1) In a data analyst's ideal world, the data underlying the algorithm would be objective and accurate representations of a world. Even then, the world may be one in which there are human and other biases against females in the workplace – pre-existing biases that manifest themselves in the data. (2) The data may be "biased" in one of several statistical senses, e.g. by being a convenience sample that is not representative of the population. (3) All data – input descriptors as well as outcome labels – are influenced by the concepts available to the data modellers, collectors and processors, i.e. dependent on the framings of their socio-historical context (Kitchin, 2014; Barocas and Selbst, 2016). This can be regarded as an instance of technical bias.

The term “biased data” is used for all three phenomena. In sum, AB is strongly determined by the data from which an algorithm has been learned.

The third reason arises from the moral and legal evaluation of “bias” as **“unfairness”**. Women earning systematically less than men seems unfair, but if the female employees are, on average, less highly skilled than the male ones, is paying them less (on average) really unfair? Some decades ago, the answer to this question would probably have been a clear “no”; today, many people would probably add that the education and training system that produces such skill distributions, as well as the relative valuation of different types of work, are intrinsically unfair. Different philosophical and political notions of equality, justice and fairness (such as equality of outcome vs. equality of opportunity) give rise to different metrics of when a classification or resulting allocation would be regarded as fair (or not). In “discrimination-aware” or “fairness-aware” data-mining and machine learning, a wide number of such metrics are being discussed, see Section “Measures to mitigate AD”. Even some laws refer to metrics and thresholds in order to decide when a differential treatment is sufficiently unfair to become discrimination (such as the so-called 80% rule, Biddle, 2005), even though legal terms and concepts cannot be fully formalised in general and often deviate from the use of the terms in AD specifically (e.g. Berendt and Preibusch, 2014; Xiang and Raji, 2019). The debate mirrors an earlier one on the ‘right’ metric of fairness in educational testing, and while it is important for algorithm designers and analysts to operationalise their concepts in a clear mathematical way, the historical development of that earlier debate (Hutchinson, 2019) suggests that trying to find the ‘right metric’ is futile and less important than furthering the understanding of the larger context and how to change it. Also, other (including legal) related notions, such as unfair commercial practices (see book entry *unfairness*), could be drawn on to enrich the field of AD.

A fourth reason is that AB is typically a property of a **machine-learned** “decision algorithm”. These differ from algorithms with human-coded decision rules that may have pre-existing bias ‘directly engineered in’ (such that it can be ‘engineered out’). In machine-learned algorithms, bias tends to work indirectly and therefore requires **indirect** bias-reduction strategies. Do these algorithms contain, in the sense of Friedman and Nissenbaum (1996), *pre-existing* bias, and can this bias be ‘engineered out of them’ easily? Unfortunately, this is often not the case. Decision algorithms are, in a big-data era, generally machine-learned by applying comparatively general-purpose learning procedures (themselves algorithms) to specific data. For example, a decision algorithm could state that “if the user is a woman, then income is/should be in the middle or low range”, learned by a general-purpose algorithm that learns if-then rules from statistical regularities in a dataset. This can produce *emergent* bias. It is generally impracticable, or even impossible, to check and post-correct each decision algorithm learned, in a given configuration and situation and often dynamically changing. Instead, algorithm designers mostly concentrate on indirect mitigation strategies: to modify the general-purpose learning procedures or the input data.

Discrimination between items, biased data, and AB can result in **differential treatment of people** in various ways. The ‘item’ may in itself be the representation of a person, as in the example of risk assessment systems. The discrimination that results from the application of the algorithm to the data describing different individuals thus results in differential treatment such as some individuals being granted and others not being granted parole. If the item is a piece of information (such as a search result) or another resource (such as a loan) that is being made available or not, or being made available in forms that are more or less salient, differently framed, etc. for different persons, the effects are more indirect, but they can still be significant. People are influenced by item discrimination cognitively and emotionally, as well as in their actions. For example, not receiving the information about certain job advertisements may prevent individuals from applying and thus from obtaining such jobs, and/or it may convince them they are not good enough for certain jobs. As a result, individuals and groups will remain in or move to different (often: worse) employment histories and different (often: lower) socio-economic status. In general, effects of the differential

treatment may range from slight annoyance to grave violations of fundamental rights, and perceptions may range from ‘none’ to ‘extreme and inacceptable’.

The differential treatment of people can result in **unlawful discrimination**, **or** in **discrimination that is** not (or not yet) unlawful but considered **socially undesirable**. There are legal, social, and ethical questions regarding when differential treatment becomes discrimination, and we refer the reader to the entry on discrimination. For the present purposes, we will simplify this question by highlighting that most analyses of AD focus on the criterion of differentiation being a legally protected ground, often represented by an attribute of the individual or group (see also the “Challenges and Future Work” section below). The discrimination can be direct (~ disparate treatment) or indirect discrimination (~ disparate impact).

Discrimination via algorithms in algorithmic systems vs. discrimination via algorithmic systems. An algorithm by itself does not discriminate (in the social or legal sense that we are interested in here). An algorithm is a procedure for effecting certain computations, including the input and output of data. Thus, AD cannot be “discrimination *by* algorithms”. Rather, the discrimination happens in an algorithmic system, which we define as a socio-technical system that includes algorithms deployed on computers and generally working on data, as well as people and social rules and institutions. It is therefore meaningful to speak of, analyse, and improve on, AD in the sense of “discrimination *via* algorithms in algorithmic systems”. For example, women tend to be discriminated against not by a performance assessment algorithm as a computational procedure, but by this performance assessment algorithm being deployed in workplaces and by the ways they are used to make or support decisions with regard to promotions, pay rises, etc.

However, the term neglects the potentially discriminatory effects of the many other choices made to shape an algorithmic system. We call such effects “discrimination via algorithmic systems”. One set of choices concern the computational system, which includes not only algorithms but also, e.g., user interfaces (Berendt and Preibusch, 2014, 2017). Bias and discriminatory effects can result from the *technology* per se, or *emerge* in contexts of use, especially those that designers probably did not expect. For example, the need to file a social-welfare application online discriminates against poor people, since these may not have sufficient access to the internet, and against people with lower degrees of digital literacy (Eubanks, 2018).

Another set of choices concern the wider socio-technical system. For example, political/administrative choices to deploy certain computational systems on selected demographics can also amount to discrimination. Examples include (a) the comparison of the large amount of personal data that are routinely collected from applicants for housing assistances, i.e. the homeless, vs. the much fewer data that are routinely collected from applicants for other state assistance, such as student loans, by Eubanks, 2018; and (b) the deployment of the Dutch system for detecting welfare fraud, SyRI, exclusively in poor neighbourhoods (Van Veen, 2019). SyRI was found to be in breach of European Convention on Human Rights by the The Hague District Court for data-protection reasons, but the court also recognised the potential for discriminatory effects, which had been a major argument against the system by civil society (*Nederlands Juristen Comité voor de mensenrechten et al. v The Netherlands* [2020]). It can be argued that an overly strong focus on discriminatory effects of algorithms and therefore on remedial effects of algorithm designers, corresponds to the politically often naive attribution of bias and discrimination to individuals (see D’Ignazio and Klein, 2020). A focus on algorithmic systems, instead, helps to focus also on structural discrimination and on the need for measures to alleviate it.

In sum, AD is not an effect of algorithms in themselves, but of the way in which they are being embedded in computational and socio-technical systems.

Measures to mitigate AD

AD has become a sizeable and intensely growing area of research since the first formulations of countermeasures against it from the data-science community itself. In this section, we briefly describe major research themes. Throughout, we will talk about "mitigating" AD, even if we agree with Pedreschi *et al.* (2008) that the goal should be to "prevent" it. This word choice appears more realistic, especially given that (a) algorithmically, often a reduction in bias and its consequences is the best possible result given other constraints, and (b) we want to avoid giving the impression that there are changes to algorithms and data that can truly and completely eradicate the possible discriminatory effects of complex algorithmic systems.

A first major theme is the development of **techniques to detect and mitigate AD with modelling and algorithmic means**. The "post-processing" of patterns by Pedreschi *et al.* (2008) as an approach to AD mitigation was complemented by approaches that modify ("de-bias") the learning algorithm or the input data (Hajian and Domingo-Ferrer, 2013).

The years following these proposals have been characterised by the development of a large number of approaches for detecting and mitigating AD , the increasing use of the term "unfairness" and "fairness" to describe the spectrum between AD and the ideal of its absence, and the discovery of a multitude of real-world examples of AD. A core research theme has been the question of which metrics best capture AD and fairness (e.g. Zliobaite, 2017), and the demonstration that certain combinations of fairness cannot be achieved simultaneously (Chouldechova, 2017; Kleinberg, Mulleinathan, and Raghavan, 2017; see also Binns, 2018, for a contextualisation of these questions with respect to political philosophy). Many metrics are based on statistical disparities between groups as indicators of deficiencies in *group fairness*, while others measure potential differences between treatment of individuals that should be treated alike *(individual fairness).*

The book entries *discrimination data analysis* and *algorithmic fairness* provide excellent surveys of these developments. More extensive overviews of the field include the book by Barocas, Hardt and Narayanan (2019) with a wide scope including causality, legal and political aspects. Romei and Ruggieri (2014) provided a broad and deep multidisciplinary survey drawing on perspectives from law, statistics, economics and computer science. Friedler *et al.* (2019) conducted an in-depth comparison of performance over four algorithms, five real-world datasets, default accuracy measures, and eight notions of fairness. Dunkelau and Leuschel (2019) presented an extensive overview of formalisations and algorithms.

Large vendors and platforms now offer toolkits with which designers can inspect and improve their algorithms. An overview is given by Dunkelau and Leuschel (2019). Data and model modifications also give rise to new questions regarding their lawfulness (Ntoutsi *et al.*, 2020).

A second major theme is the **relationship between AD mitigation and transparency**. The challenge results from the concomitant rise, during this same period, of (a) highly complex neural-network architectures in data science and processing ("deep learning") and (b) increasingly complex pipelines of data processing that combine data and algorithms from different sources. These general developments in Artificial Intelligence, coupled with (c) an increasing deployment of AI and other complex algorithmic systems in real-life contexts, make it increasingly difficult to understand – and hence, to challenge – algorithms in general. This also implies that it becomes increasingly difficult to understand and challenge possible AD. One proposed countermeasure is to develop and deploy algorithmic and other technical as well as organisational measures to enhance transparency – a requirement that pervades politics as well as the law anyway, but which needs conceptual enquires into the relationship between transparency, explanation, understandability, intervenability,

accountability, and similar concepts, as well as research, development, and clear legal specifications and enforcement. These topics are investigated in work on 'explainable AI' (e.g. Guidotti *et al.*, 2019), on whether/how there exists a 'right to explanation' in the GDPR (e.g. Selbst and Powles, 2018, Malgieri and Comandé, 2017), and on whether transparency and explanations are in fact always the remedy wanted (Edwards and Veale, 2017).

A much-studied example of such effects of processing-chain dependencies is that of **bias in natural language**, and a much-studied example domain is that of recruiting and other labour-market tasks. In some labour markets, the majority of applications is never seen by a human, and the training data of algorithms encode the widespread and persistent disadvantages for female job-seekers and employees. Occasionally, such bias is detected and the algorithmic procedure modified or discontinued (for an overview and references, see Barocas *et al.*, 2019). Decision algorithms that involve natural-language input data (such as the derivation of job-posting recommendations from search queries or résumés) increasingly rely on *language models* that encode regularities such as co-occurrences of words and word sequences. These models are learned, independently of the later task, from large text corpora, and they therefore often contain AB (such as associating male pronouns with high-prestige job titles and female pronouns with low-prestige job titles, see the study by Datta *et al.*, 2015, described in the "Examples" section above). Unless *these* models are analysed and, ideally, de-biased, the subsequent decision algorithms will carry the bias further. State-of-the-art language models are themselves complex neural networks, and detecting, mitigating, and even defining in what sense they are biased are non-trivial tasks (Blodgett *et al.*, 2020). One reason is that the question of what constitutes bias in language remains an evolving and controversial (socio-)linguistic question.

The need for **interdisciplinary approaches** to understanding and mitigating AD is not limited to data that involve natural language. The increasing recognition of this need is a third major theme in current research and practice. Discrimination has been studied for many years in many fields (Romei and Ruggieri, 2014). Many AD researchers collaborate in interdisciplinary teams, and one of the key conferences in the area in 2018 strategically shed the reference to "machine learning" in its name to explicitly integrate *all* research on "fairness, accountability and transparency in socio-technical systems" (see www.facct.org) while at the same time becoming a conference under the auspices of the ACM, one of the most relevant international associations of *computing* scholars and professionals. Interdisciplinary collaborations can profit from methods that go beyond addition of methodology and expertise and also encourage and strengthen the reflection of implicit assumptions and terminology (Allhutter and Berendt, 2020). The area has strong overlaps with critical and feminist data science, fields that analyse how many of the common framings of mainstream data science help co-produce the social-justice problems that AD research tries to address (D'Ignazio and Klein, 2020).

It is difficult to assess which mitigation strategies were **deployed** and have had what success (or not) **in real-life settings**. Several instances of AD have received much attention in science and the media. Follow-up media reports suggest that companies took a range of mitigation measures, including: mitigating data bias by enhancing previously unrepresentative training data (Raji and Buolamwini, 2019), and terminating or pausing the use of algorithms found to exhibit AB (Dastin, 2018; Heilweil, 2020). Several public-sector algorithmic systems have been discontinued after political pressure and court judgements (Niklas, 2019; *Nederlands Juristen Comité voor de mensenrechten et al. v The Netherlands* [2020]). In other cases, steps have been taken to de-bias algorithms, possibly by the manual addition of edge-case treatment. An example of the latter is the change in Google's query completion after the UN Women (2013) campaign publicised that the query "Women should" was completed by proposals such as "be slaves", and "Women shouldn't" by "have rights" - completions learned from past searches. Like the biases in language models mentioned above, query recommendations like these are instances of AD being reprehensible by

perpetuating stereotypes and cultural denigration (*representational harm*) rather than by withholding opportunities or resources (*allocative harm*, Barocas *et al.*, 2019, also referred to as *distributive harm*, Binns, 2018).

At the same time, due to standard business secrecy and the widespread use of proprietary software, it is impossible to have an overview of the use of algorithms and the possible AD caused by them in the private sector, and calls for greater transparency and algorithm audits may actually lead companies away from transparency (Raji *et al.*, 2020). Regarding the public sector, even the attempt to compile an overview of where algorithmic decision-making is employed in the public sector in the EU proved to be challenging (AlgorithmWatch and Bertelsmann Stiftung, 2019).

Challenges and future work

Efforts to understand and mitigate AD face many challenges.

A first challenge is the question of **who** the "patients" and the "agents" of AD are, in the moral and legal senses.

As the examples have demonstrated, the *patient* of AD can be an individual represented in a computational system, or a user of a computational system. It can also be someone with the features of a group that was not, or not sufficiently, or only in a biased way, represented in data used to train a machine-learning system. These categories are not mutually exclusive and often, many stakeholders are affected (for example, the presentation of, and exposure to, repeated and stereotyped depictions of population groups affect "the user" as well as different population groups). Especially when faced with non-transparent algorithmic systems, proprietary data and algorithms, and disputed metrics, it can be very difficult to prove that AD has occurred and who has been discriminated against how.

It is also difficult to determine the *agents* of AD. Understanding where AD may occur, let alone mitigating it, *after* algorithmic decision-making has been deployed in largely unregulated ways in complex socio-technical systems, is often impracticable. A more promising – technical as well as regulatory – approach is a by-design methodology, as manifested for example in the GDPR's combination of (a) the requirement to perform a technology impact assessment (specifically: a data protection impact assessment) before processing data (Art. 35), (b) the requirement to build systems with the desired value by default and by design (Art. 25), and (c) the requirement to monitor and be accountable for compliance (Art. 5 (2)). The GDPR itself as well as relevant interpretations of it allow such assessments to also consider possible discriminatory impacts (Naudts, 2018). However, this approach may only work well for monolithic systems with clear control and responsibility. Modern software engineering practices tend to be decentralised and agile, and they exhibit a high degree of interconnectedness and dynamicity (as well as the uncertainties about what a self-learning system will learn when deployed). In addition, different roles and different people may qualify as responsible for the effects of algorithmic systems, including AD.

Together, these conditions imply that the responsibility for components of algorithmic decision-making tends to be widely distributed and that previewing the impact of design decisions may be extremely difficult (Gürses and Van Hoboken, 2018). This renders the need to review and revise existing liability schemes more challenging and at the same time essential for protecting both those who produce and those who are affected by algorithmic systems (Datenethikkommission, 2020; European Commission Expert Group on Liability and New Technologies, 2019, see book entry *liability*).

A second challenge is to determine **what** types of differential treatment and of AD should be in the focus of scientific and practical/political investigations and in the focus of legal reform. For example, should we focus on traditional *legally protected grounds* such as "racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation" (GDPR Recital 71)? Should a list be extended by a "such as" as in the Universal Declaration of Human Right's Article 2 ("distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status [; or] on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs") or in the EU's European Charter of Fundamental Rights, Article 21 ("any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation")?

What about usually *not (or not explicitly) legally protected grounds* such as socio-economic status (i.e., discrimination against the poor, Eubanks 2018) or grounds that change with policies (such as 'migration status', Taylor 2016)? And what about *new* differentiation criteria that are machine-learned, such as 'having a single parent' or 'having shopped for guitars', both of which have been found to lead to unfavourable outcomes from prediction models (EPIC, n.d.; Andrews, 2012)? While the first may turn out to be a proxy for ethnicity and thus a reason to suspect indirect discrimination, the second may not clearly correlate with a known demographic feature but still be perceived as discriminatory.

Are these 'new' grounds justified bases for differential treatment? Should such 'new' grounds be considered discriminatory, and why or why not? In an analysis of European Court of Human Right's case-law given the conceptual freedom of "such as" and "other status" in Article 14 of the European Convention on Human Rights, Naudts (2019) traces a complex history of which differentiations have been regarded as constituting (or not) discrimination. This history suggests an underlying search for a social-psychology construct: for example, do grounds have to be "innate or inherently linked to the identity of the individual" in order for the differentiation to be discriminatory? If so, who gets to decide what counts as "inherently linked to identity"? To what extent is the recognition (or not) of a factor as identity-linked an external attribution and thus an expression of the same political power that is the root cause of the discrimination? What role does history play: are grounds describing historically disadvantaged groups clearer indicators of discrimination? How much and how long does disadvantage have to persist to qualify as a protected ground alongside the known historically disadvantaged groups? Is recurrent AD that is based on changing, ephemeral grounds such as 'having shopped for guitars' less, equally, or more damaging than persistent AD based on a stable ground? Is it even possible to respect people's individuality in such decision contexts, rather than resorting – more or less strongly – to treating them as instances of categories? Thus, many questions and arguments from (political) philosophy on what exactly egalitarianism entails return with a vengeance when algorithms are involved (Binns, 2018). Among them is the issue of the "spheres of justice": that in different realms of life (such as employment, voting, and dating), different types of inequalities appear morally wrong to a given society – or acceptable, or even deserved and morally right.

From a computational perspective, such grounds – once they are known – are all alike in the sense of being one feature, one variable that for every person represented in a dataset could take on some value. Thus, they can all be addressed by the detection and mitigation strategies described above. However, these grounds need to be known before the analysis to allow for metrics as well as mitigation strategies to be applied.

Today's social debates revolve, increasingly, around intersectionality and multidimensional discrimination. Some forms are well-known (Taylor, 2017) and continue to describe specifically

marginalised groups (e.g. black women); a simple approach would be to just declare such a *known* combination a new ground and then apply the same methods. However, this approach faces (also) statistical problems. In recent years, formal methods have been developed that can *discover*, from data, that a certain group (maybe not known in this combination before) is segregated (Baroni and Ruggieri, 2019) and/or discriminated against (Kearns *et al.*, 2018). However, how can and should they be deployed in real-life discrimination mitigation strategies? They form 'new grounds' similarly to the one-criterion new grounds sketched above, but given that the essence of decision models is to differentiate and given that they do this on combinations of grounds, there will always be *some* disadvantaged group. Which of these, and when, should be considered for protection and AD mitigation?

Types of AD can be distinguished by whether they arise from fully automated or only partially automated decision making. Many authors, citizens, and lawmakers agree that special safeguards should be in place to allow people to contest individual decision making solely based on automated processing, including its errors and AD that it may cause. In the GDPR (Article 22), and the Data Protection Directive before it, data subjects are offered specific rights when faced with such decision making, but this clause has not had much effect in the past (Berendt and Preibusch, 2017). The implicit, and already in the past debated, assumption is that involving a human – in any way – in the decision making will help resolve problems with fully automated decisions. But humans are also influenced by machine output in computer-assisted decision making, and thus the question becomes whether and how this should be regulated (e.g., Citron, 2007; Ferguson, 2015).

A further unresolved question is **how** data scientists can or should build algorithms and tools. Data science, which is strongly shaped by artificial intelligence research, tends to *model* persons in terms of a set of characteristics rather than as individuals with agency (Berendt and Preibusch, 2017; D'Ignazio and Klein, 2020). Much of AD research implicitly assumes that discrimination consists in treating people who should be treated alike differently. Thus, the goal of non-discrimination consists in treating people who should be treated alike (as) equally (as possible). This ignores the fact that discrimination may also consist in the failure to treat different people differently, as when disabled employees have the right to be provided with special help. It also models people as *being* something, rather than as persons with agency in social contexts.

Consider the 2016 *Taddeucci and McCall v Italy* case in which the European Court of Human Rights ruled that it had been discriminatory to treat an unmarried homosexual couple equally to unmarried heterosexual couples (the non-European partner had been denied a residence permit on family grounds). The court argued that a heterosexual couple could get married (and thus change its feature value relevant to this domain), whereas at the time and place under consideration a homosexual couple was not able to do so. The judgement illustrates, first, how reasoning about discrimination must go beyond what people "are" and take into account what they "could be" – and that the context as well as their agency and restrictions on it will determine which of the possible alternatives can materialise. It illustrates, second, that not only the individual or the discriminated-against and their actions count, but also those of others. It emphasises, third, how non-discrimination interacts with other fundamental rights such as autonomy and agency. It is difficult, if not impossible, to model all these aspects of discrimination formally and "solve AD with artificial intelligence". Data scientists and engineers should also focus on building interactive tools that help people detect and reason about AD with human intelligence.

To respect and support the autonomy of people potentially affected by discrimination, another consideration should be remembered: the ultimate goal of AD mitigation should not be less bias in the distribution of the same (often bad) outcomes over different groups, but more justice for all.

References

AlgorithmWatch and Bertelsmann Stiftung, *Automating Society Taking Stock of Automated Decision-Making in the EU* (2019) <https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf> accessed 11 August 2020

L Andrews, *I Know Who You Are and I Saw What You Did: Social Networks and the Death of Privacy* (Free Press 2012)

J Angwin, J Larson, S Mattu and L Kirchner, 'Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 11 August 2020

D Allhutter and B Berendt, 'Deconstructing FAT: using memories to collectively explore implicit assumptions, values and context in practices of debiasing and discrimination-awareness' in *FAT* 2020: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) 687

S Barocas, M Hardt and A Narayanan, *Fairness and Machine Learning* (2019) <http://www.fairmlbook.org> accessed 11 August 2020

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

A Baroni and S Ruggieri, 'SCube: A Tool for Segregation Discovery' in *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019*, (OpenProceedings.org 2019) 542

B Berendt and S Preibusch, 'Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence' (2014) 22 (2) Artificial Intelligence and Law 175

B Berendt and S Preibusch, 'Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass' (2017) 5 (2) Big Data 135

D Biddle, *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing* (Gower 2005)

R Binns, 'Fairness in Machine Learning: Lessons from Political Philosophy' in *Conference on Fairness, Accountability and Transparency, FAT 2018* (Proceedings of Machine Learning Research 81, 2018) 149

S L Blodgett, S Barocas, H Daumé III and H Wallach, 'Language (Technology) is Power: A Critical Survey of "Bias" in NLP' in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL 2020) 5454

J Buolamwini and T Gebru, 'Gender Shades: Intersectional accuracy disparities in commercial gender classification' in *Conference on Fairness, Accountability and Transparency, FAT 2018* (Proceedings of Machine Learning Research 81, 2018) 77

A Chouldechova. Fair predictions with disparate impact: A study of bias in recidivism prediction instruments. 5 (2) Big Data.153

D K Citron, ‘Technological Due Process’ (2007) 85 Washington University Law Review 1249 <http://ssrn.com/abstract=1012360> accessed 11 August 2020

B Custers, T Calders, B Schermer and T Zarsky, (eds.), *Discrimination and Privacy in the Information Society. Data mining and Profiling in Large Databases* (Springer: Studies in Applied Philsophy, Epistemology and Rational Ethics (3) 2013)

J Dastin, ‘Amazon scraps secret AI recruiting tool that showed bias against women’ (*Reuters*, 10 October 2018) <<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> accessed 11 August 2020

Datenethikkommission, *Opinion of the Data Ethics Commission* (2020) <https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3> accessed 11 August 2020
A Datta, M C Tschantz and A Datta, ‘Automated experiments on ad privacy settings’ in *Proceedings on Privacy Enhancing Technologies* (2015) (1) 92
C D’Ignazio and L.F. Klein, *Data Feminism* (MIT Press 2020)

J Dunkelau and M Leuschel, *Fairness-Aware Machine LearningAn Extensive Overview* (Heinrich-Heine Universität Düsseldorf, Working Paper 2019) <https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft/KMW_I/Working_Paper/Dunkelau___Leuschel__2019__Fairness-Aware_Machine_Learning.pdf> accessed 11 August 2020

C Dwork, M. Hardt, T. Pitassi, O Reingold and R S Zemel, ‘Fairness through awareness’ in *Proceedings of Innovations in Theoretical Computer Science, ITCS 2012* (ACM 2012) 21

L Edwards and M Veale, ‘Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for’ (2017) 16 Duke Law & Technology Review 18 <https://ssrn.com/abstract=2972855> accessed 11 August 2020

EPIC, ‘Algorithms in the Criminal Justice System: Risk Assessment Tools’ (n.d.) <https://epic.org/algorithmic-transparency/crim-justice/> accessed 11 August 2020

V Eubanks, *Automating Inequality. How High-Tech Tools Profile, Police and Punish the Poor* (St. Martin’s Press 2018)

European Commission, *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*. (COM(2020) 65 final, 2020) <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en> accessed 11 August 2020

European Commission Expert Group on Liability and New Technologies – New Technologies Formation, *Liability for Artificial Intelligence and Other Emerging Technologies* (2019) <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608> accessed 11 August 2020

A G Ferguson, ‘Big Data and predictive reasonable suspicion’ (2015) 1632 University of Pennsylvania Law Review 327

S A Friedler, C Scheidegger, S Venkatasubramanian, S Choudhary, E P Hamilton and D Roth, ‘A comparative study of fairness-enhancing interventions in machine learning’ in *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* 2019* (ACM 2019) 329

B Friedman and H Nissenbaum, ‘Bias in computer systems’ (1996) 14 (3) ACM Transactions on Information Systems 330

R Gellert, E De Vries, P De Hert, and S Gutwirth, ‘A comparative analysis of anti-discrimination and data protection legislations’ in Custers *et al.* (eds.) (2013) 61

R Guidotti, A Monreale, S Ruggieri, F Turini, F Giannotti and D Pedreschi, ‘A survey of methods for explaining black box models’ (2019) 51 (5) ACM Computing Surveys 93:1

S Gürses and J Van Hoboken, ‘Privacy after the agile turn’ in J Polonetsky, O Tene and E Selinger (eds.) *Cambridge Handbook of Consumer Privacy* (Cambridge University Press 2018) 579

S Hajian and J Domingo-Ferrer, ‘A methodology for direct and indirect discrimination prevention in data mining’ (2013) 25 (7) IEEE Transactions on Knowledge and Data Engineering 1445

S Hajian, J Domingo-Ferrer, A Monreale, D Pedreschi and F Giannotti, ‘Discrimination- and privacy-aware patterns’ (2015) 29 (6) Data Mining and Knowledge Discovery 1733

F Hamborg, K Donnay and B Gipp, ‘Automated identification of media bias in news articles: an interdisciplinary literature review’ (2019) 20 International Journal on Digital Libraries 391

R Heilweil, 'Big tech companies back away from selling facial recognition to police. That’s progress' (11 June 2020) Vox <https://www.vox.com/recode/2020/6/10/21287194/amazon-microsoft-ibm-facial-recognition-moratorium-police> accessed 11 August 2020

B Hutchinson and M Mitchell, ‘50 years of test (un)fairness: Lessons for machine learning’ in *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* 2019* (ACM 2019) 49

M Kearns, S Neel, A Roth and Z S Wu, ‘Preventing fairness gerrymandering: Auditing and learning for subgroup fairness’ in *Proceedings of the 35th International Conference on Machine Learning* (Proceedings of Machine Learning Research (80) 2018) 2564

J Kleinberg, S Mullainathan and M Raghavan, ‘Inherent trade-offs in the fair determination of risk scores’ in Proceedings of the 8th Conf. on Innovations in Theoretical Computer Science, ITCS (2017) 43:1

R Kitchin, *The Data Revolution. Big Data, Open Data, Data Infrastructures & Their Consequences* (Sage 2014)

B Kulynych, R Overdorf, C Troncoso and S F Gürses, ‘POTs: protective optimization technologies’ in *FAT* 2020: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) 177

S Lowry and G Macpherson, ‘A blot on the profession’ (1988) 296 British Medical Journal 657

G Malgieri and G Comandé, ‘Why a right to legibility of automated decision-making exists in the General Data Protection Regulation’ (2017) 7 (3) International Data Privacy Law 243

L Naudts, ‘Criminal profiling and non-discrimination: On firm grounds for the digital era?’ in A Vedder, J Schroers, C Ducuing and P Valcke (eds) *Security and Law. Legal and Ethical Aspects of Public Security, Cyber Security and Critical Infrastructure Security* (Intersentia 2019) 63. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3508020> accessed 11 August 2020

L Naudts, ‘How machine learning generates unfair inequalities and how data protection instruments may help in mitigating them’ in R Leenes, R van Brakel, S Gutwirth and P De Hert (eds.) *Data Protection and Privacy: The Internet of Bodies (Computers, Privacy and Data Protection Vol. 11) 2018* (Hart Publishing 2018) <https://ssrn.com/abstract=3468121> accessed 11 August 2020

J Niklas, ‘Poland: Government to scrap controversial unemployment scoring system’ (16 April 2019) <https://algorithmwatch.org/en/story/poland-government-to-scrap-controversial-unemployment-scoring-system/> accessed 11 August 2020

E Ntoutsi, P Fafalios, U Gadiraju, V Iosifidis, W Nejdl, M-E Vidal, S Ruggieri, F Turini, S Papadopoulos, E Krasanakis, J Kompatsiaris, K Kinder-Kurlanda, C Wagner, F Karimi, M Fernández, H Alani, B Berendt, T Kruegel, C Heinze, K Broelemann, G Kasneci, T Tiropanis, S Staab, ‘Bias in data-driven artificial intelligence systems – An introductory survey’ (2020) 10 (3) Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery <https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1356> accessed 11 August 2020

D Pedreschi, S Ruggieri and F Turini, ‘Discrimination-aware data mining’ in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM 2008) 560

D Pedreschi, S Ruggieri and F Turini, ‘A study of top-k measures for discrimination discovery’ (2012) in *Proceedings of the Symposium on Applied Computing, SAC’12* (ACM 2012) 126

A D Selbst and J Powles, ‘ “Meaningful Information” and the Right to Explanation’ in *Conference on Fairness, Accountability and Transparency, FAT 2018* (Proceedings of Machine Learning Research 81, 2018) 48

I D Raji and J Buolamwini, ‘Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products’ in *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2019) 429

I D Raji, T Gebru, M Mitchell, J Buolamwini, J Lee and E Denton, ‘Saving face: Investigating the ethical concerns of facial recognition auditing’ in *AIES '20: Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2020) 145

A Romei and S Ruggieri, ‘A multidisciplinary survey on discrimination analysis’ (2014) 29 (5) Knowledge Engineering Review 582

L Taylor, ‘Refugees, migrants and data science: it’s not just research’ (24 November 2016) <https://linnettaylor.wordpress.com/2016/11/24/refugees-migrants-and-data-science-its-not-just-research/> accessed 11 August 2020

L Taylor, ‘What is data justice? The case for connecting digital rights and freedoms globally’ (2017) 4 (2) Big Data & Society <https://doi.org/10.1177/2053951717736335> accessed 11 August 2020

UN Women, ‘UN Women ad series reveals widespread sexism’ (21 October 2013) <http://www.unwomen.org/en/news/stories/2013/10/women-should-ads> accessed 11 August 2020

C van Veen, ‘Profiling the Poor in the Dutch Welfare State’ (1 November 2019) <https://chrgj.org/2019/11/01/profiling-the-poor-in-the-dutch-welfare-state/> accessed 11 August 2020

A Vedder, ‘Why data protection and transparency are not enough when facing social problems of machine learning in a big data context’ in E Bayamlioglu *et al.* (eds) *Being profiled: Cogitas, ergo sum. 10 Years of Profiling the European Citizen.* (Amsterdam University Press 2018). <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3407853> accessed 11 August 2020

A Xiang and I D Raji, ‘On the legal compatibility of fairness definitions’, CoRR abs/1912.00761 (2019) <http://arxiv.org/abs/1912.00761v1> accessed 11 August 2020

I Zliobaite, ‘Measuring discrimination in algorithmic decision making’ (2017) 31 (4) Data Mining and Knowledge Discovery 1060

CASES

European Court of Human Rights

Taddeucci and McCall v Italy App no 51362/09 (EctHR, 30 June 2016)

European Court of Justice

Case C-236/09 Association belge des Consommateurs Test-Achats ASBL v Conseil des ministres [2011] ECR *2011 I-00773*

Case C-524/06 Heinz Huber v Bundesrepublik Deutschland [2008] *ECR 2008 I-09705*

Other courts

Nederlands Juristen Comité voor de mensenrechten et al. v The Netherlands [2020] Rechtbank Den Haag [2020] C-09-550982-HA ZA 18-388. English version of the judgment at <https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878> accessed 11 August 2020