# A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection

**Bryce W. Goodman**
Oxford Internet Institute
University of Oxford
bwgoodman@gmail.com

## Abstract

Algorithms, and the data they process, play an increasingly important role in decisions with significant consequences for human welfare. This trend has given rise to calls for greater accountability in algorithm design and implementation, and concern over the emergence of algorithmic discrimination. In that spirit, this paper asks whether and to what extent the European Union's recently adopted General Data Protection Regulation (GDPR) successfully addresses algorithmic discrimination. As the first piece of legislation to explicitly address algorithmic discrimination, the GDPR sets an important precedent: its success, or failure, will have repercussions that extend well beyond Europe. We argue that while the GDPR's two primary principles for dealing with algorithmic discrimination—*data sanitization* and *algorithm transparency*—are likely inadequate, the legislation also paves the way for third party inspections of algorithms or 'algorithm audits'. If implemented properly, the algorithm audits supported by the GDPR could play a critical role in making algorithms less discriminatory and more accountable.

## 1 Introduction

The proliferation of automated decision-making in everyday life has been accompanied by a call to make algorithms accountable (Angwin, 2016). One of the chief motivations for accountable algorithms is concern over algorithmic discrimination, which occurs when an individual or group receives unfair treatment as a result of algorithmic decision-making, e.g. automated profiling. In April 2016, the European Parliament and Council officially adopted the General Data Protection Regulation (GDPR), the first set of comprehensive regulations for the collection, storage, and processing of personal data within the European Union in over two decades. While the bulk of the GDPR is specifically focused on the "right to protection of personal data", it is also the first piece of legislation to address explicitly the effect of algorithmic decision making on the "fundamental rights and freedoms of natural persons" (art. 1(2)), including algorithmic discrimination.

As the first piece of legislation to address algorithmic discrimination explicitly, the GDPR sets an important precedent: its success, or failure, will have repercussions that extend well beyond Europe. Consequently, it is crucial to analyse how effective the GDPR may be in combating algorithmic discrimination and promoting more accountable algorithms. This is the task of the following pages. Section one argues that the two principles enshrined in the GDPR—*data sanitization* and *algorithm transparency*—provide only a partial solution. Section two shows that the GDPR also lays the groundwork for implementing algorithm audits, third-party inspections of algorithms aimed at discovering and reducing discriminatory effects. If properly designed, algorithm audits could prove vital for increasing accountability and ensuring that algorithmic discrimination is kept at bay. The

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

conclusion highlights open questions for how algorithm auditing ought to be implemented under the GDPR.

## 2 Algorithmic Discrimination and the GDPR

The GDPR's most explicit mention of algorithmic discrimination occurs in *Recital 71*, which states a requirement to "implement technical and organizational measures" that "prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect." These characteristics are referred to as "special categories", and have their basis in non-discrimination legislation, such as *Article 14* of the European Convention on Human Rights (EC, 2010).

It is worth noting at the outset that the GDPR does not attempt to provide a precise definition of algorithmic discrimination. Nor does it differentiate between disparate treatment, which occurs when an individual or group receives unfavourable treatment on the basis of any special categories, and disparate impact, which occurs when ostensibly neutral practices disadvantage special categories (McCrudden and Prechal, 2009).[1] The language of *Recital 71* (e.g. "discriminatory effects") suggests that the GDPR's focus is on disparate impact. However the mechanisms introduced by the GDPR, discussed below, focus primarily on eliminating disparate treatment. This paper does not attempt to resolve this apparent inconsistency.

Moving beyond the *Recitals* and into the GDPR itself, algorithmic discrimination is addressed by two key principles. The first, *data sanitization*, is the removal of special categories from datasets used in automated decision making. This principle is introduced by *Article 9: Processing of special categories of personal data*, which establishes a *prima facie* prohibition against "the processing of data revealing racial or ethnic origin" and other "special categories". It is strengthened under *Article 22: Automated individual decision-making, including profiling*, which specifically prohibits "a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" that is "based on the special categories of personal data referred to in *Article 9*" (art. 22(2)).

The second principle, *algorithm transparency*, introduces the "right to explanation" (Goodman and Flaxman, 2016), whereby data subjects are entitled to "meaningful information about the logic involved, as well as the significance and the envisaged consequences" when automated decision making or profiling takes place (art. 13(2)(f); art. 14(2)(g)). In *Article 12: Transparent information, communications and modalities for the exercise of the rights of the data subject*, the GDPR further specifies that such information must be provided "in a concise, transparent, intelligible and easily accessible form, using clear and plain language."

The following sections discuss each principle in turn.

### 2.1 Special Categories and *Article 9*: The Minimal and Maximal Requirement

The GDPR's primary mechanism for combating algorithmic discrimination is to "sanitize" data used in automated decision making, i.e., to prevent the inclusion of variables related to protected categories outlined in *Article 9*. The GDPR's *Article 9* thus reflects a focus on disparate treatment: it parallels other EU non-discrimination legislation such as *Article 13* of the amended Amsterdam Treaty (2002), which prohibits "discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation."

In practice, the scope of the restrictions in *Article 9* will depend upon how the phrase "revealing" is interpreted. At least, *Article 9* clearly applies to any variables that explicitly state membership in a special category, i.e. the race, gender, etc. of data subjects. This is the *minimal requirement*. At most, *Article 9* might also apply to any variables that, individually or jointly, have a statistically significant relationship with a special category. This is the *maximal requirement*.

Unfortunately, both requirements come short of being satisfactory, for different reasons, explained in the next two sections.

[1] The tension between disparate treatment versus disparate impact in regulating big data has received and certainly deserves attention (Barocas and Selbst, 2016), but a fuller discussion is outside the scope of this paper.

### 2.1.1 The Minimal Requirement and Taste-based Discrimination

An algorithm will acquire a so-called "taste" for discrimination (Becker, 1957) if the relationship between the predictor and target variables in the dataset reflects overtly discriminatory treatment (see Custers et al., p.50). Imagine that a firm's manager is racist, and categorically gives higher performance ratings to white employees even when they perform at the same level as non-white colleagues. If a model is trained on the dataset of past performance ratings and race is explicitly coded, the resulting model will directly discriminate on the basis of race.

In this case, the minimal requirement would entail dropping the race variable from the dataset. This would cause the model to have a lower accuracy on the training data because that data reflects the manager's biased ratings, not actual performance. This loss in accuracy is desirable, however, because the algorithm no longer discriminates, and predictions are now based on actual employee performance.

Unfortunately, the minimal requirement is not effective in cases of statistical discrimination, where special category membership is genuinely predictive. In such cases, the relationship between special category membership and predicted outcomes is redundantly encoded by proxy variables in the data (see Barocas and Selbst, 2016, p.21). Removing variables for special category membership, per the minimal requirement of *Article 9*, does not affect predictions.

### 2.1.2 The Maximal Requirement

The maximal requirement seeks to remove both explicit and proxy variables for special category membership. In this case, we might begin by identifying and listing statistically significant correlations between gender and other predictor variables. However, eliminating all proxy variables will likely result in the loss of useful information with legitimate relevance to decision-making (see Calders and Verwer, 2010, p.279). Furthermore, even if one eliminated all variables that have a statistically significant correlation with gender from our dataset, there is always the possibility that remaining variables may have a significant correlation with gender on the aggregate (Dodge, 2003). As Berk notes in his study of race and crime forecasting algorithms, "ignoring race altogether can leave substantial racial effects behind" (Berk, 2009, p.239). With the further damage, one may add, that the analysis now appears officially unbiased. Thus while the minimal requirements of *Article 9* are likely ineffective, the maximal requirements are likely infeasible.

Before concluding our discussion of *Article 9*, it is important to note an important risk for both the minimal and maximal interpretations. In short, *prohibiting the collection or processing of data revealing special category membership may worsen the problem it is intended to solve*. The methods proposed for identifying and reducing discrimination in algorithms are only effective if special category membership is indicated in the dataset (Feldman et al., 2015). Furthermore, variables with no theoretical interpretation whatsoever may be highly indicative of special category membership. However, there is no way to establish whether this is or is not the case without information about special category membership. Eliminating the collection of data revealing sensitive categories may, perversely, allow discrimination to continue and deepen by making it impossible to be detected in the first place.

## 2.2 The Inadequacy of Transparency

One of the advantages of learning algorithms versus traditional approaches is that they are able to develop highly sophisticated models capable of representing complex, non-linear decision boundaries. For example, neural networks are composed of multiple nodes each representing a particular set of functional transformations and arranged into "visible" and "hidden" layers (Bishop, 2006). This multi-level structure gives great flexibility from a representational point of view, but makes the "internal logic" of neural networks elusive. The trade-off between model flexibility and interpretability is not new *per se*: over two decades ago, Wildberger (1994) noted that "the opacity of neural networks [has] meant that it has not been possible to derive any clear logical relationship between their interior configuration and their external behaviour." However, the enthusiastic adoption of neural network approaches, and their deployment in novel prediction contexts, adds a new sense of urgency to the challenge Wildberger identified (cf. Lisboa, 2013; Kim et al., 2015; Burrell, 2016; Kroll et al., 2016).

Another barrier to transparency in machine learning is the nature of the data. Traditional analyses (e.g. actuarial predictions) typically employ a small number of theoretically important variables with known distributions (Trowbridge, 1989). By contrast, machine learning techniques are agnostic to the theoretical importance of variables, making them suitable for very large, high-dimensional datasets (cf. Berk and Hyatt, 2015, p.223). In these applications, the challenge for providing a meaningful account may be twofold: first, the sheer complexity of the underlying model and, second, the fact that many predictor variables have no theoretical interpretation whatsoever.

Finally, many machine learning algorithms cannot be sensibly scrutinized independently of the training data that was used to generate models. This is especially true with non-parametric methods, such as the K-nearest-neighbours approach (Bishop, 2006, pp.125-126). As the name suggests, this model classifies new observations according to the class-membership of its K-nearest-neighbours, which is based on the training data. When separated from the training data, however, the decision rule is meaningless. In addition to concerns over technical feasibility, the need to store all data used to develop decision rules highlights a potential conflict between effectively monitoring for algorithmic discrimination and the GDPR's principle of data minimization (cf. art 5(1)(c)).

## 3 Algorithm Auditing and the GDPR

The two principles proposed by the GDPR, namely data sanitization and algorithmic transparency, are unlikely to address algorithmic discrimination successfully when this is:

- *unintentional*, in the sense that whoever employs the algorithm may genuinely harbour no ill will towards the discriminated individual or group;
- *opaque*, in the sense that discrimination may be difficult or impossible to detect *a priori*.

The upshot is that identifying and rectifying algorithmic discrimination is not trivial, and it is unlikely that all of those implementing automated profiling (especially business organisations) will also be best positioned to evaluate and ensure an adequate level of compliance.

One potential solution is to implement *algorithm audits*, third party inspections of algorithmic decision-making modelled on audit studies from social science. As Sandvig notes, audit studies are "the most prevalent social scientific method for the detection of discrimination... [and] are typically field experiments in which researchers or their confederates participate in a social process that they suspect to be corrupt in order to diagnose harmful discrimination" (Sandvig et al., 2014, p.5). More recently, algorithm audits have been used by researchers and journalists to reveal discrimination in online advertising (Sweeney, 2013; Datta and Tschantz, 2015), interest rates (Miller, 2015) and pricing (Valentino-DeVries et al., 2012).

To date, however, efforts to implement algorithm auditing have encountered legal resistance (Farivar, 2016). In the United States, a recent lawsuit brought about by researchers challenges the barriers that the Computer Fraud and Abuse Act (CFAA) has placed against third party "testing for discrimination on the Internet" (ACLU, 2016). As the plaintiffs note, "many common website terms of service prohibit... [activities] necessary for robust audit testing to uncover discrimination on the Internet", and the CFAA effectively makes such terms of service enforceable as criminal law. Thus one area where the GDPR could have a significant impact on algorithmic discrimination is in providing a legal mechanism for allowing, or even requiring, algorithm auditing.

Although the GDPR does not explicitly propose auditing algorithms, it is anticipated by three distinct instruments:

- *Data impact assessments* (art. 24) requires data controllers to evaluate "the risks of varying likelihood and severity for the right and freedoms of natural persons" posed by data processing and to "implement appropriate technical and organizational measures to ensure and *to be able to demonstrate* that processing is performed in accordance with this Regulation" (emphasis added);
- *Codes of conduct* (art. 40) encourages designated bodies to "prepare codes of conduct...such as with regard to fair and transparent processing" and "to carry out the mandatory monitoring of compliance"; and

- *Certification* (art. 42) authorizes "the establishment of data protection certification mechanisms and data protection seals and marks... available via a process that is transparent" and subject to periodic review.

### 3.1 Auditing as Safety Engineering

Each of the three instruments listed at the end of the previous section offers a potential route for introducing algorithm auditing. However, before anything can be said about how audits should be conducted, it is critical to specify what, exactly, algorithm audits are intended to achieve. In safety engineering, auditing identifies key process risks, evaluates whether adequate safeguards are in place and, where gaps are found, provides guidance on risk prevention going forward (Hansson, 2014).

Within safety engineering it is common to distinguish between "primary" or "inherent safety" and "secondary" risk prevention (Kletz, 1978). Primary prevention prevents risks during initial design. By contrast, secondary prevention is the mitigation of risk through additional measures. Equipping a building with fire extinguishers would be an example of secondary prevention, while using non-flammable building materials is an example of primary prevention (Hansson, 2014). These two concepts roughly map on to pre, in and post processing approaches to discrimination prevention (Singh and Sane, 2014).

Pre-processing techniques seek to eliminate the possibility of constructing a discriminatory classifier by 'correcting' the training data for discrimination, i.e. by modifying the dataset in some way (Kamiran and Calders, 2012; Feldman et al., 2015). This may include identifying and removing attributes that are most highly correlated with protected categories (e.g. race, gender, etc.) or modifying labels for the training data to reduce traces of discrimination (Calders and Verwer, 2010). By contrast, in- and post-processing approaches are based upon reframing a general supervised learning classification problem to become "discrimination aware", i.e. to learn a classifier such that accuracy is high and discrimination with respect to the protected category is low (Pedreshi et al., 2008; Hajian et al., 2012). The discrimination aware classification problem is thus a case of multiple optimization, and is typically approached by modifying a classic learning algorithm to include a measure of discrimination (a 'discrimination constraint') during model training or, in the case of post-processing, model testing (Kamiran and Calders, 2009; Singh and Sane, 2014).

Although these approaches offer promising tools for identifying and correcting algorithmic discrimination, they are not without limitations. It is widely acknowledged that there will often be a trade-off between discrimination removal and classifier performance (Calders and Verwer, 2010; Kamiran and Calders, 2012; Singh and Sane, 2014). Traits that are highly correlated with a protected category may also be genuinely and legitimately informative. The proposed approaches to discrimination detection and removal will also fail when discrimination arises from a mismatch between the sample used for training and the actual population. Lastly, none of the research to date has addressed discrimination in an active learning environment, where the output of algorithmic decision-making becomes part of the basis of future decisions (Goodman and Flaxman, 2016).

More generally, the success of in-, pre- and post-processing approaches hinge entirely on how well discrimination is formally defined. As discussed in Section 1, non-discrimination is a multivalent concept. In some cases, it concerns the outcome of decisions, whereas in others it may concern the procedure by which decisions are made. One can also distinguish between non-discrimination as *consistency* or *individual fairness*, in which similar people experience similar outcomes, versus *statistical parity*, where protected and unprotected groups have the same probability of outcomes (Bonchi, 2016). This does not mean that non-discrimination requirements are essentially ambiguous, but that that they are not translatable into a formal set of universal necessary and sufficient conditions. As Wouter Vandenhole (2005) notes, "[t]here is no universally accepted definition of discrimination." The literature on algorithms is no exception (Pedreschi et al., 2012). One consequence is that a wholly automated approach to algorithmic auditing is unlikely to succeed. Rather, human judgement will play an essential role in determining whether bonafide discrimination takes place.

Auditing is not a panacea, but sets up minimal conditions for workability: a process that passes a safety audit may fail for other reasons (e.g. inefficiency). Passing a safety audit does not mean that all risk is eliminated but, rather, that risk is reduced to an acceptable level. Choosing an acceptable

level of risk depends in turn on the process evaluated and, in particular, both the likelihood and severity of a failure.

To this end, one may develop various tiers of auditing requirements that correspond to the probability and magnitude of harm arising from algorithmic discrimination. In practice, this may mean restrictions on the types of algorithms that can be employed to accomplish certain prediction tasks. In "safety critical" environments (e.g. where there is a high risk of discrimination and associated harms are severe), one may exclude some specific input variables either because they are too strongly correlated with special category membership, or because they do not represent legitimate desiderata. For example, in order to avoid inadvertent red-lining, one might prohibit some algorithms from including variables that reveal location, e.g. IP address (Peck, 2013). In other cases, one may insist on human-interpretable models that only employ input features that are believed to be causally related to the output (Varshney, 2016).

The upshot is that, both in designing accountable algorithms and developing a framework for auditing, it is important to consider the relative consequences of algorithmic decision-making. Algorithms for optimizing advertising and evaluating creditworthiness may both exhibit discrimination, but the effects of the former, while potentially significant, are different in kind from the latter. The goal should be to set both safeguards and auditing standards that are appropriate to the setting in which the algorithm is deployed. While the standards themselves may be formalized, the choice of which standards to implement will always depend on an interpretation of relevant risks. This suggests that auditing should be a dialectical process, wherein the auditor exists less to provide answers than to ensure the right questions have been asked.

## 4 Conclusion

This paper began with a question—whether and to what extent the GDPR adequately addresses algorithmic discrimination. Our inquiry reveals that the answer crucially depends upon how the discrimination manifests, both in terms of its underlying technical causes and outgoing social effects. In light of this, we argue that the most promising approach is to employ audits that assess algorithms on a case by case basis. Such audits will, if properly executed, provide both quantitative and qualitative indicators that can inform the design and operation of more fair and transparent algorithms.

Before concluding, three open questions should be highlighted.

First: who will be responsible for performing audits? The GDPR does not make clear whether monitoring codes of conduct or certification should be conducted by a public, non-governmental or for-profit entity. From a practical point of view, the track record of public monitoring in other industries, such as the financial sector (Weil and Wilke, 2002), raises doubts about whether public bodies will have the resources and technical expertise to be effective in practice. At the same time, collusion between private auditors and auditees to circumvent requirements is an obvious concern (Baiman et al., 1991; Acemoglu and Gietzmann, 1997; Lee and Gu, 1998). Private auditing may also create novel power imbalances. For example, auditing in the financial sector is currently dominated by the "big four" auditing agencies. These quasi-monopoly conditions inflate the cost of auditing services, and may also create an incentive for auditors to endorse auditing practices that are needlessly complex and obscure (Power, 1997, 2003).

Second: who will bear what costs? Even if governments completely subsidize the cost of audits, companies will face additional expense for implementing internal compliance. Larger companies may face higher compliance costs, but it is unlikely that these costs will scale. This means that the cost of auditing will likely have a disproportionate impact on smaller companies, and may pose a barrier to entry. This cost should be kept in check to ensure that it does not hinder competition.

Third: how much should companies be expected to assist with audits? If auditors have no information about special category membership, they cannot determine whether and to what extent disparate impact exists. This raises an important problem, namely whether data controllers should be encouraged or required to collect information about special category membership for the sake of ensuring that auditors can check for algorithmic discrimination.

There are, more generally, open questions regarding how discrimination ought to be both legally and formally defined. In particular, there is debate around whether the focus ought to be on ensuring formal equality in how decisions are made versus their distributive effects, and protecting special

categories versus equal treatment of individuals (Gosepath, 2011; Barocas, 2014; Bonchi, 2016). These questions are both foundational and inherently multidisciplinary. Meaningful engagement thus requires expertise drawn from computer science and engineering as well as social science, law and philosophy (see Kroll et al., 2016).

It is outside the scope of this paper to resolve these questions. However, one of the most important contributions of the GDPR is to give these questions a greater sense of urgency. In the past, companies have devoted immense resources to improving algorithmic performance. Going forward, one hopes to see similar investments in promoting fair and accountable algorithms.

## 5 Acknowledgements

Thanks to Luciano Floridi and Seth Flaxman for conversations and feedback.

## References

D. Acemoglu and M. B. Gietzmann. Auditor independence, incomplete contracts and the role of legal liability. *European Accounting Review*, 6(3):355–375, 1997.

ACLU. Sandvig v. lynch — challenge to cfaa prohibition on uncovering racial discrimination online, Jun 2016. URL `https://www.aclu.org/cases/sandvig-v-lynch`.

J. Angwin. Make algorithms accountable. *The New York Times*, Aug 2016. ISSN 0362-4331. URL `http://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html`.

S. Baiman, J. H. Evans, and N. J. Nagarajan. Collusion in auditing. *Journal of Accounting Research*, page 1–18, 1991.

S. Barocas. *Data Mining and the Discourse on Discrimination*. 2014. URL `https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf`.

S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2512208`.

G. Becker. *The economics of discrimination*. University of Chicago Press, Chicago, 1957.

R. Berk. The role of race in forecasts of violent crime. *Race and social problems*, 1(4):231–242, 2009.

R. Berk and J. Hyatt. Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter*, 27(4):222–228, Apr 2015. ISSN 10539867, 15338363. doi: 10.1525/fsr.2015.27.4.222.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. URL `http://www.academia.edu/download/30428242/bg0137.pdf`.

F. Bonchi. Tutorial on algorithmic bias, Aug 2016. URL `http://francescobonchi.com/algorithmic_bias_tutorial.html`.

J. Burrell. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data Society*, 3(1), 2016. URL `http://bds.sagepub.com/content/3/1/2053951715622512.abstract`.

T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, Jul 2010. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-010-0190-x.

B. Custers, T. Calders, B. Schermer, and T. Zarsky. *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer Berlin Heidelberg. ISBN 978-3-642-30486-6. URL `http://link.springer.com/10.1007/978-3-642-30487-3`.

A. Datta and Tschantz. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

Y. Dodge. Interaction effect. *Oxford Dictionary of Statistical Terms*, 2003.

EC. *European Convention on Human Rights*. European Court of Human Rights, 2010. URL http://link.springer.com/chapter/10.1007/978-94-017-7173-3_1.

C. Farivar. To study possibly racist algorithms, professors have to sue the us, Jun 2016. URL http://arstechnica.co.uk/tech-policy/2016/06/do-housing-jobs-sites-have-racist-algorithms-academics-sue-to-find-out/.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *Certifying and removing disparate impact*, page 259–268. ACM, 2015. URL http://dl.acm.org/citation.cfm?id=2783311.

B. Goodman and S. Flaxman. *European Union regulations on algorithmic decision making and a "right to explanation"*. Jun 2016. URL https://arxiv.org/abs/1606.08813.

S. Gosepath. *Equality*. Spring 2011 edition, 2011. URL http://plato.stanford.edu/archives/spr2011/entries/equality/.

S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer, and F. Giannotti. *Injecting discrimination and privacy awareness into pattern discovery*, page 360–369. IEEE, 2012. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6406463.

S. O. Hansson. *Risk*. Spring 2014 edition, 2014. URL http://plato.stanford.edu/archives/spr2014/entries/risk/.

F. Kamiran and T. Calders. *Classifying without discriminating*, page 1–6. IEEE, 2009. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4909197.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

B. Kim, K. Patel, A. Rostamizadeh, and J. A. Shah. *Scalable and Interpretable Data Representation for High-Dimensional, Complex Data.*, page 1763–1769. 2015. URL http://people.csail.mit.edu/beenkim/papers/BeenPRSAAAI15.pdf.

T. A. Kletz. What you don't have, can't leak. *Chemistry and Industry*, 6:287–292, 1978.

J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. *Accountable Algorithms*. Number ID 2765268. Mar 2016. URL https://papers.ssrn.com/abstract=2765268.

C.-W. J. Lee and Z. Gu. Low balling, legal liability and auditor independence. *Accounting Review*, page 533–555, 1998.

P. J. Lisboa. *Interpretability in Machine Learning–Principles and Practice*, page 15–21. Springer, 2013. URL http://link.springer.com/chapter/10.1007/978-3-319-03200-9_2.

C. McCrudden and S. Prechal. *The Concepts of Equality and Non-Discrimination in Europe: A practical approach*. Nov 2009.

C. C. Miller. When algorithms discriminate. *The New York Times*, Jul 2015. ISSN 0362-4331. URL http://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html.

D. Peck. They're watching you at work. *The Atlantic*, Dec 2013. ISSN 1072-7825. URL http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

D. Pedreschi, S. Ruggieri, and F. Turini. *A study of top-k measures for discrimination discovery*, page 126–131. ACM, 2012. URL http://dl.acm.org/citation.cfm?id=2245303.

D. Pedreshi, S. Ruggieri, and F. Turini. *Discrimination-aware data mining*, page 560–568. ACM, 2008. URL http://dl.acm.org/citation.cfm?id=1401959.

M. Power. Expertise and the construction of relevance: Accountants and environmental audit. *Accounting, Organizations and Society*, 22(2):123–146, Feb 1997. ISSN 0361-3682. doi: 10.1016/S0361-3682(96)00037-2.

M. Power. Auditing and the production of legitimacy. *Accounting, Organizations and Society*, 28 (4):379–394, May 2003. ISSN 0361-3682. doi: 10.1016/S0361-3682(01)00047-2.

C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 2014. URL http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf.

J. Singh and S. S. Sane. Discrimination discovery and prevention in data mining: A survey. *International Journal of Engineering Research and Applications*, 4(6):54–57, 2014.

L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

C. L. Trowbridge. *Fundamental Concepts of Actuarial Science*. Actuarial Education and Research Fund, 1989.

J. Valentino-DeVries, J. Singer-Vine, and A. Soltani. Websites vary prices, deals based on users' information. *Wall Street Journal*, Dec 2012. ISSN 0099-9660. URL http://www.wsj.com/articles/SB10001424127887323777204578189391813881534.

W. Vandenhole. *Non-discrimination and equality in the view of the UN human rights treaty bodies*. Intersentia nv, 2005. URL https://books.google.co.uk/books?hl=en&lr=&id=FjrMK-1BW38C&oi=fnd&pg=PR5&dq=Vandenhole,+Wouter.+2005.+Non-Discrimination+and+Equality+in+the+View+of+the+UN+Human+Rights+Treaty+Bodies,+Oxford,+UK:+Intersentia.&ots=xDobVtqIlc&sig=qb0rCleKuVAoGAlMWHaqAG9CMTY.

K. R. Varshney. Engineering safety in machine learning. *arXiv preprint arXiv:1601.04126*, 2016. URL http://arxiv.org/abs/1601.04126.

J. Weil and J. Wilke. Systemic failure by sec is seen in enron debacle. *Wall Street Journal*, Oct 2002. ISSN 0099-9660. URL http://www.wsj.com/articles/SB1033944629262271233.

A. Wildberger. *Alleviating the opacity of neural networks*, volume 4, page 2373–2376 vol.4. Jun 1994. doi: 10.1109/ICNN.1994.374590.