Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2017: 1–17

DOI: 10.1177/2053951717743530
journals.sagepub.com/home/bds
SAGE

Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data

Michael Veale[1] and Reuben Binns[2]

Abstract

Decisions based on algorithmic, machine learning models can be unfair, reproducing biases in historical data used to train them. While computational techniques are emerging to address aspects of these concerns through communities such as discrimination-aware data mining (DADM) and fairness, accountability and transparency machine learning (FATML), their practical implementation faces real-world challenges. For legal, institutional or commercial reasons, organisations might not hold the data on sensitive attributes such as gender, ethnicity, sexuality or disability needed to diagnose and mitigate emergent indirect discrimination-by-proxy, such as redlining. Such organisations might also lack the knowledge and capacity to identify and manage fairness issues that are emergent properties of complex sociotechnical systems. This paper presents and discusses three potential approaches to deal with such knowledge and information deficits in the context of fairer machine learning. Trusted third parties could selectively store data necessary for performing discrimination discovery and incorporating fairness constraints into model-building in a privacy-preserving manner. Collaborative online platforms would allow diverse organisations to record, share and access contextual and experiential knowledge to promote fairness in machine learning systems. Finally, unsupervised learning and pedagogically interpretable algorithms might allow fairness hypotheses to be built for further selective testing and exploration. Real-world fairness challenges in machine learning are not abstract, constrained optimisation problems, but are institutionally and contextually grounded. Computational fairness tools are useful, but must be researched and developed in and with the messy contexts that will shape their deployment, rather than just for imagined situations. Not doing so risks real, near-term algorithmic harm.

Keywords

Algorithmic accountability, algorithms, discrimination, machine learning, personal data, privacy

Introduction

As machine learning techniques are taken up in an ever-wider array of sectors for decision-making and decision-support, many have pointed to harms that might result from their careless or malicious implementation. Some harms surround fairness, as it proves to be difficult to make systems that do not exhibit bias, indirectly or in subsets of data (Hajian, 2013; Kamiran et al., 2012; Barocas and Selbst, 2016). These are nested within a range of linked concerns, including algorithmic transparency and accountability (Burrell, 2016; Keats Citron and Pasquale, 2014; Kroll et al., 2016; Nissenbaum, 1996), in-the-wild reliability (Žliobaitė et al., 2016); security against adversaries (Huang et al., 2011; McDaniel et al., 2016); entrenchment of inequality (boyd and Crawford, 2012; Harcourt, 2006); risks to privacy and due process (Hildebrandt and Gutwirth, 2008); and the enablement of ambient, ubiquitous surveillance systems (Hildebrandt, 2015; Kitchin and Dodge, 2011). These have mobilised a

[1]Department of Science, Technology, Engineering and Public Policy (*STEaPP*), University College London, UK
[2]Department of Computer Science, University of Oxford, UK

Corresponding author:
Michael Veale, University College London, 36–38 Fitzroy Square, London W1T 6EY, UK.
Email: m.veale@ucl.ac.uk

wide array of researchers and practitioners to consider how these technologies can be utilised whilst minimising the pitfalls and risks that might accompany them.

This paper focuses on how fairness and discrimination in machine learning systems can be mitigated within practical institutional constraints. Machine learning systems, which identify and utilise patterns in data, are designed to discriminate. We use these systems to distinguish data points from each other based on certain predictive characteristics. Some forms of discrimination however are considered unacceptable (Hellman, 2008). Legally 'protected characteristics', usually including disability, race, sexuality, gender, pregnancy, among others, are broadly illegal to use in most decision-making. These are not set in stone. For example, while single-sex sports clubs, toilets, or specific types of job advert (e.g. modelling) are usually not illegal, acceptance of them is changing. Discrimination usually also requires cases to be otherwise comparable. In some situations, sex might not be considered discriminatory where decisions hinge on differences in statistical life expectancies (Berendt and Preibusch, 2014).

Other bars for measuring fairness are less universal. Judging based on appearance; on events that occurred some time ago; on limited data; on actions an individual has already been sanctioned for, or in conditions of high uncertainty and rapid change, is sometimes acceptable, sometimes not. Judging based on arbitrary characteristics, like favouring those who access online forms with custom web browsers (Pinsker, 2015), might also seem unfair, perhaps because of the opportunistic short-lived nature of such correlations as well as the associated ways it might discriminate against those accessing forms from schools, or from libraries.

Some sources of unfair machine learning systems

There are several interacting ways that deployment of machine learning can potentially lead to unfair or discriminatory outcomes.

Unfairness in data, their collection and their processing. Many of the fairness issues in machine learning are primarily thought to arise from data. Some think, falling for what could be called the 'neutrality fallacy', that machine learning will provide a more even and objective treatment of individuals (Sandvig, 2015). As Latour indicates, we are often more than happy to declare value-laden issues as matters of fact, and let machines settle them for us (1999). This is rarely appropriate.

The high demand for labelled data in the context of supervised machine learning – the focus of this paper – can usually only be met by using data from previous decision-making. If these historical data reflect existing, unwanted discrimination in society, the model that is learned from it – essentially a similarity engine – will likely encode these same patterns, risking reproduction of past disparities. Machine learning algorithms are supposed to discriminate between data points – that is why we use them – yet some logics of discrimination, even if predictively valid, are not societally acceptable.

Furthermore, if some sub-groups are historically undersampled, or exhibit more complicated, nuanced or under-evidenced patterns compared to others, models might exhibit differential performance. It is not practically possible to have data on all individuals, quantifying or classifying all factors important to some social phenomenon. People, or aspects of their lives, are always missing. These skews fast make their way into data-driven systems.

Data are often also cleaned and transformed before use, in subjective ways. 'Feature engineering', where input variables are transformed to make them more amenable to modelling, has crucial downstream impact on the behaviour of machine learning systems. Feature engineering emphasises aspects of certain variables through augmentation, aggregation and summarisation of characteristics whilst downplaying others. For instance, aggregating those who subscribe to different branches of a religious doctrine (e.g. Catholic, Protestant; Shia, Sunni) within a single overarching doctrine (Christian, Muslim) might collapse distinctions which are highly relevant to questions of fairness and discrimination within certain contexts. Including a standard deviation of a characteristic as an input variable will make it easier for a machine learning model to emphasise divergence from a constructed average. As with many issues in machine learning, the political nature of this classifying and sorting has long been recognised (Bowker and Star, 1999). Categorisation does not just label people, it can create groups and alter future outcomes (Hacking, 1995; Harcourt, 2006), just as feature engineering can in machine learning (Rouvroy, 2011).

Unfairness from selecting and specifying a machine learning system. Humans carry their worldviews and make value-laden choices, with both foreseeable and unforeseeable consequences, during the whole modelling process. While machine learning is often portrayed as automated, a great deal of subjective human labour is involved in system design and deployment. Model choice itself can be political. Neural networks or random forests are more amenable to capturing synergy between variables than linear regression. Use of regression might omit important contextual variance, for example. Within a model family, further hyperparameters must be specified. Higher regularisation parameters penalise complexity in a model, which might help it

generalise but might trade-off for certain complicated or rare patterns not being retained. Different evaluation mechanisms for models emphasise different aspects of performance (Japkowicz and Shah, 2011). Unfortunately, 'neutral' choices in machine learning systems do not exist – candidates for these, such as software defaults, are best thought of as arbitrary.

Finally, once a model has been built, there are various ways it can be deployed in practice which may introduce additional fairness issues. The extent to which a model may have different impacts on different groups may only become evident once that model is put into a decision-making system; for instance, the setting of thresholds for positive and negative outcomes could have significant consequences for different groups which may not be evident by merely studying the model itself. The introduction of an algorithmic system may also provide spurious justification for decisions which would otherwise have been more open to challenge under a purely human decision-making process (Skitka et al., 1999).

As with any sociotechnical, value-laden problem, we cannot expect to find simple or universal panaceas. We are stuck with layered, messy techniques to define, resolve and manage these complex challenges. This paper zooms in to examine one piece of this challenge – how potentially unfair patterns in datasets that make their way into modelling and decision-making processes might be remedied in practical rather than theoretical machine learning situations. We emphasise situations where actors designing and deploying such systems wish to avoid bias themselves, for regulatory and reputation-related reasons, rather than adversarial situations where external investigators wish to discover bias against the will of the organisations undertaking analysis. Legislative discussion within a European context of the ability to investigate algorithmic systems can be found in Edwards and Veale (2017).

Can we statistically 'debias' data and algorithms?

Computational techniques to prevent machine learning methods from perpetuating these forms of bias have been proposed in recent years by research communities such as discrimination-aware data mining (DADM) and fairness, accountability and transparency in machine learning (FATML). They involve altering usual data science processes in order to correct these forms of bias. They can operate at several stages, including pre-processing, in-processing and post-processing (Hajian and Domingo-Ferrer, 2013). In each case, the aim is to induce patterns that do not lead to discriminatory decisions despite the possibility of biases in the training data.

Anti-discrimination law has particularly motivated DADM and FATML communities, who have attempted to formalise these requirements for mathematical implementation. For instance, heuristics such as the US Equal Employment Opportunity Commission's '80% rule', which provides a suggested level of permissible disparity between protected groups and the general population, have been used to set parameters for fairness-aware models (Feldman et al., 2015). Within European contexts, non-discrimination and data protection are rights enshrined in the EU Charter of Fundamental Rights, and both potentially relate to the risks of unfairness inherent in machine learning applications (Gellert et al., 2013). Recital 71 of the EU General Data Protection Regulation (GDPR) refers in particular to fairness-aware data mining technologies and organisational measures.

There are multiple ways to define fairness formally in machine learning contexts. Most measures focus on differences in treatment between protected and non-protected groups, but there are multiple ways to measure differences in outcomes. These include: 'disparate impact' or 'statistical/demographic parity', which considers classification rates between groups;[1] 'accuracy equity', which considers the overall accuracy of a predictive model for each group (Angwin et al., 2016; Dieterich et al., 2016); 'conditional accuracy equity', which considers the accuracy of a predictive model for each group, conditional on their predicted class (Dieterich et al., 2016); 'equality of opportunity', which considers whether each group is equally likely to be predicted a desirable outcome given the actual base rates for that group (Hardt et al., 2016); and 'disparate mistreatment', a corollary which considers differences in false positive rates between groups (Zafar et al., 2016). Other measures focus not just on actual outcomes and their relation to true/false positives/negatives, but on counterfactual scenarios wherein members of the protected groups are instead members of the non-protected group (i.e. a woman classified by the system should get the same classification she would have done had she been a man) (Kusner et al., 2017).

Each of these measures of fairness are arguably reasonable ways to measure fairness. One might therefore hope that a fair system would satisfy all of these constraints. But unfortunately, recent work has formally proven that it is impossible for a model to satisfy several of these constraints at the same time, except in exceptional cases which are unlikely to hold in the real world (Berk et al., 2017; Chouldechova, 2017; Kleinberg et al., 2016). As a result, choices between the different measures will have to be made. In some cases it may be more important to focus on differences between positive classifications (e.g. loan applications), and therefore an 'equality of opportunity' measure

might be preferable; in others, the cost of a false negative might be higher (e.g. the risk a violent criminal might pose to the public). Thus the choice of a particular fairness measure therefore ought to be sensitive to the context.

Setting aside these definitional problems, fairness-aware machine learning techniques are increasingly seen as desirable, viable and even in some cases legally recommended or required. However, an important challenge remains. To be successful, these techniques depend on knowledge about the potential correlations between features in the training data and protected characteristics that are the subject of anti-discrimination and data protection law. In practice, this is a condition that is either not always met, or not always desirable to meet.

Why knowledge of protected characteristics is both necessary and problematic?

To see why knowledge of protected characteristics is necessary, it is helpful to consider why certain naïve approaches to removing bias from modelling are inadequate. One could simply delete any sensitive variables related to discrimination, e.g. age, gender, race, or religion, from the training data. Unfortunately, this does not guarantee non-discrimination in the models that are trained on this data, as non-discriminatory items might exist which in some conditions are closely correlated with the sensitive attributes. Where geography serves as a sensitive proxy, this phenomenon is termed 'redlining'. More broadly, it can be seen as an issue of redundant encoding.

In order to discover redlining in training data, one needs to be able to find out whether sensitive attributes might be encoded by other, apparently benign ones. For instance, to discover whether ZIP codes in a dataset are correlated with, e.g. race, it will be necessary to either have race as an attribute in the dataset, or to have background knowledge about the demographics of the areas in question (for instance, from census records). Proposed approaches to non-discriminatory machine learning assume that whoever is implementing the technique has access to the sensitive attributes which might be encoded (e.g. Hajian and Domingo-Ferrer, 2013; Hardt et al., 2016). Such access is necessary for assurance of computationally non-discriminatory models (Žliobaitė and Custers, 2016).

Despite this, in many cases organisations deploying machine learning will lack this necessary access, often for legitimate reasons.

First, the collection of personal data inevitably creates privacy risks. Many organisations have internalised the dictum of regulators and privacy advocates only to collect data that is necessary for their purposes. The concepts of data minimisation and purpose limitation within the GDPR are intended to prevent collection and processing of data for unspecified or disproportionate ends. Furthermore, the kinds of protected characteristics involved in cases of discrimination raise higher privacy and data protection risks than other kinds of data, and are given special protection under both the GDPR and other laws (Edwards and Veale, 2017). The proposition that organisations ought to collect a wide range of sensitive data that isn't directly necessary for their primary purposes contradicts this general dictum. Yet fairness-aware machine learning seems to require organisations to do exactly that to adequately inspect and modify their models.[2]

It is not our aim here to analyse the extent to which privacy and data protection law and best practice is substantively in conflict with the collection and processing of sensitive attributes for the purposes of fairness-aware machine learning.[3] It may be that collection and processing for such purposes is legitimate; however, it may still not be desirable. It would require data subjects to share sensitive attributes along with non-sensitive ones every time their data was to be used to train a model. The general result would be much more sensitive data in the hands of data controllers – a security risk even if it is intended to be used for the legitimate purposes of avoiding discriminatory outcomes. Even if organisations are permitted to collect and process such data, requiring consumers to provide it might make their service less competitive, or less trusted. For purposes of building a model that serves some narrowly prescribed goal, they may not see the need to collect sensitive data. In the context of data minimisation, the data controller must argue that it is proportionate to collect and process sensitive categories of data, and they may not be sufficiently incentivised to do so. Where individuals fear they are being treated unfairly, the collection of sensitive data by the organisation in question, even to explicitly remedy fairness issues, might not alleviate that perception-based fear. It could even make it worse.

Some approaches have been proposed to transform training data with anonymisation procedures to protect the sensitive attributes. This can be performed in tandem with pre-processing techniques to prevent discrimination (Hajian et al., 2014; Hajian and Domingo-Ferrer, 2012). While promising, this still mandates the comprehensive collection of sensitive attributes from individuals in training data for each form of discrimination for which mitigation is desired. Despite meaningful privacy protections, the concerns raised above are still likely to apply. Individuals are unlikely to be happy providing a comprehensive range of sensitive personal data to the very organisations who are in position to

discriminate, no matter how technically robust their anonymisation process is.

Three approaches for appraising and improving fairness with limited data

Organisations developing learning systems need strategies to mitigate discrimination concerns in the absence of sensitive data. The challenge is to implement the techniques, such as those outlined above, without having to take on the additional burden and risk of collecting detailed sensitive data on the training sample.

We present three alternative approaches to overcome this challenge. The first is based on a multi-party data governance model, suited to contexts where little background knowledge about discrimination exists and a comprehensive assessment of potential forms of discrimination is needed. The second involves a collaborative knowledge sharing approach in which organisations can learn from each other's experiences in similar contexts as well as relevant sociological and demographic correlations. The third involves exploratory analysis to build hypotheses of potential unfair characteristics of the data or system, which can be more formally tested as part of a due diligence process. Figure 1 pictographically illustrates these three distinct approaches.

We do not argue that these three methods are perfect, nor that they provide complete solutions or assurances to the multitude of challenges surrounding machine learning systems. We argue instead that these are avenues that are important to explore to make fairer machine learning a practical reality in the multitude of settings that automated and semi-automated decisions will be occurring in our society in the coming years and decades.

Trusted third parties holding protected characteristics

Various proposals have been made for the involvement of external parties in the evaluation and auditing of algorithmic systems (Mantelero, 2016; Pasquale, 2010; Sandvig et al., 2014; Tutt, 2016). Some of these are reflected in law. Article 35 of the GDPR obliges organisations to undertake 'data protection impact assessments' wherever 'profiling' is used to automatically make decisions which have legal or significant effects on data subjects. In some cases these assessments may be audited by a data protection authority (Recital 84). In most governance approaches, external auditors are given access to an organisation's policies, personnel, data collection procedures, training data, models, proprietary code, and other relevant aspects, in order to assess the ethical dimensions and legal compliance of a particular algorithmic system (see Binns, 2017).

This model assumes that the relevant information required to perform an audit will lie in the hands of the organisation being audited. As argued above, this might not be the case, rendering external audit process incapable of ensuring the kinds of algorithmic fairness that DADM and FATML techniques aim for.

This might be different, were trusted third parties enlisted to work alongside organisations from when data collection begins. This proposal could be achieved with a variety of different institutional and technical arrangements. Below, we illustrate several possible implementations.

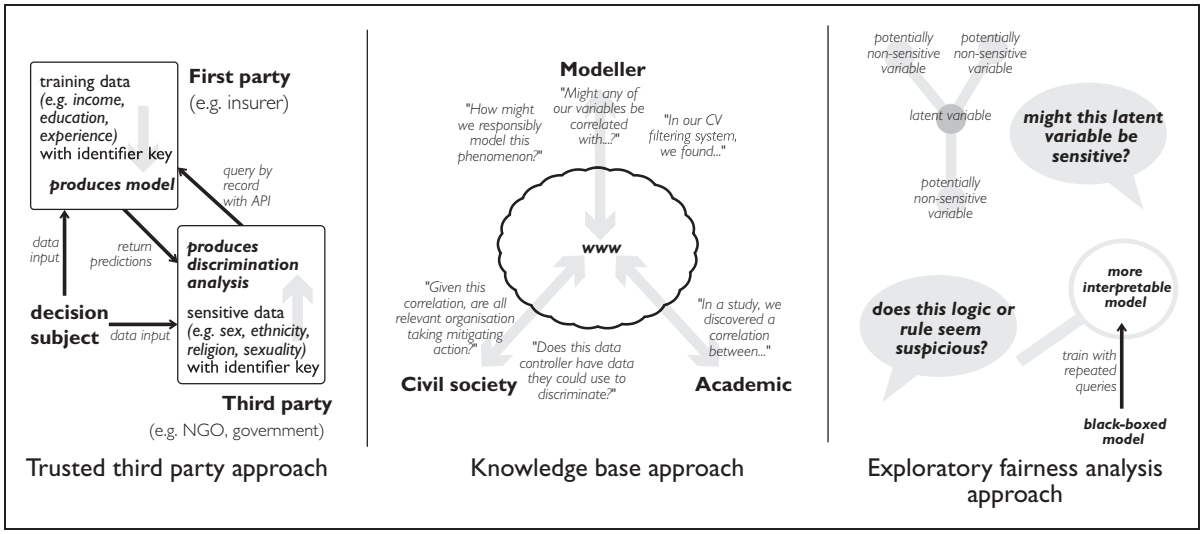

Figure 1. Three approaches to fairness-aware machine learning without holding sensitive characteristics.

The first party (the organisation implementing the algorithmic decision-making system) has access to historical data relevant to the classification or prediction task for which they are building a model. However, the first party does not and should not have access to any of the protected characteristics associated with the population used to train the model.

As discussed above, in order to statistically test the model for potential discrimination, the protected characteristics need to be linked somehow to the records used in the training data. To achieve this, a trusted third party is enlisted to collect data on the protected characteristics of those individuals whose data is used to train the model. For each individual, protected characteristics like race, gender, religious beliefs or health status are collected by the third party in parallel to the collection of the non-protected characteristics by the third party. The channel for communicating this information from the individual to the third party may depend on the platform (e.g. online, telephone, or in-person). It could be as part of a separate collection process, although this prove unwieldy, or be encrypted simultaneously and seamlessly at the point of collection (e.g. locally through JavaScript in a web browser[4]) with the public key of a third party, and transmitted to the organisation in question.[5]

Consider the following illustrative example:

> An insurer wishes to use a machine learning model to help determine customers' premiums. They have access to historical customer, and use it to train a model to predict the amount of compensation a customer will claim over the term of their cover given certain attributes (e.g. postcode, occupation, qualifications). The estimated size of a potential claim – the output of the model – is used to automatically set premiums.

The insurer enlists a third party organisation (for instance, a consumer rights group) to simultaneously collect protected characteristics about each customer as they purchase their insurance policy. For online purchases, the customer is directed to the consumer rights group's domain, and asked to provide protected characteristics for the purposes of discrimination prevention.

Based on this multi-party data governance model, there are multiple ways to proceed, depending on whether the goal is merely to detect bias or to both detect and prevent it, and what prevention techniques will be used (e.g. pre-processing, in-processing, or post-processing). We outline a set of possible variations here, and discuss their relative advantages and drawbacks.

Variation 1: Third party as ex post disparate impact detector. In cases where the third party's only role is to detect discrimination (but not prevent it), the third party need only collect protected characteristics from each individual featured in the dataset used to train (and test) the model, along with an identifier. The records held by the first party for the purposes of model training could be linked by this identifier to the records held by the third party which contain the protected characteristics. The third party would be given access to the model developed by the first party (either directly or via an application programming interface (API)). By testing the outputs of the model on each of the individuals in their sensitive attribute dataset (using the individual's identifier), the third party could detect disparate impacts.

An advantage of this variation is that the third party can only access the sensitive attributes, not the potentially non-sensitive ones. Since each record only contains sensitive attributes and an identifier this represents a lesser privacy risk; while the data itself is sensitive, it would be harder to re-identify an individual without other data types. This may also be beneficial from the perspective of a first party concerned about keeping their proprietary model secret, as it has been shown that unlimited access to a query interface for a prediction model can allow an attacker to extract and reconstruct a model (Tramèr et al., 2016). In this case, while the third party would have unrestricted ability to query the model by individual identifiers, and thus learn the distributions of outputs for each protected characteristic, they would not be able to reverse-engineer the model without access to the other, non-protected characteristics.

The disadvantage of this variation is that it only provides the first party with evidence of the disparate impact of their model. Disparate impact is a blunt measure of discrimination, because some disparities may be 'explicable', in the sense that the disparities might be accountable by reference to attributes which are legitimate grounds for differential treatment (Zafar et al., 2016; Žliobaitė et al., 2011). Furthermore, measures of disparate impact may not be sufficient for the first party to actually change their model to prevent it from being discriminatory. For instance, to remove bias from the training data, the first party would have to know which data points to relabel, massage or re-weight – i.e. the protected characteristics of the specific individuals, which they would lack. More generally, without the ability to check for redundant encoding of protected characteristics by non-protected attributes, it will be difficult for the first party to revise their model.

Nevertheless, the mere ability to detect disparate impact may be valuable in allowing third parties to flag up problems, which can then be dealt with by allowing the first party access to the necessary

additional data to investigate and transform their model accordingly. Separating out detection of disparate impact and prevention could thus prevent unnecessary sharing of sensitive attributes and enable the third party to perform continuous monitoring.

Variation 2: Third party as ex ante discrimination mitigator. Alternatively, the third party could collect both the protected attributes and the other features used to train the model. This would enable the third party to play a more significant role, not only detecting disparate impact in model outputs but also helping to ensure the disparities are attributable to disparate mistreatment (i.e. that they are not explainable), and also to ensure that the model can be bias-free.

Third party as redlining detector. In this approach, the third party has both the sensitive and potentially non-sensitive characteristics, and puts them through a common framework to produce summary information that aims to flag obvious issues that might occur during model building. Upon acquisition of a cleaned dataset, the third party calculates and returns a set of redundant encodings and their strengths. The returning document might note that 'race is correlated to zip code by 0.8'; 'gender is correlated with aspects of profession by 0.2', and so on. The first party could use this knowledge to make trade-offs in the model – removing certain features, or engaging in further discussions with the third party about potential procedures to scrub unwanted correlations from a model.

Naturally, such a framework could suffer from flaws which made it unsuitable for some types of data or problems, particularly highly contextual ones. Yet this approach would create a focal point for the improvement of discrimination detection methods for certain contexts and data types, which would foster active discussion and debate about best practices and processes that could be translated into on-the-ground practice with relative ease.

Third party as data pre-processor. Another approach would see the third party pre-process the training data in such a way as to preserve anonymity and remove bias, before handing it over to the first party. This could be achieved by modifying the data to preserve degrees of anonymity (using techniques such as statistical disclosure control (Hundepool et al., 2012; Willenborg and de Waal, 2012), and privacy-preserving data mining (Agrawal and Srikant, 2000), which allow the statistical properties of the data to be maintained), followed by applying one of a range of anti-biasing techniques described in the DADM/FATML literatures (e.g. Feldman et al., 2015; Hajian and Domingo-Ferrer, 2013; Kamiran et al., 2012).[6] It would even be possible, if it were desired, to introduce *positive discrimination* at this point, and some methods have been proposed for how this could be achieved (Verwer and Calders, 2013). As mentioned above, more recently proposed techniques aim to render datasets both k-anonymous and non-discriminatory in a single procedure with limited loss of accuracy (Hajian et al., 2014; Hajian and Domingo-Ferrer, 2012). Having transformed the data to increase privacy and remove bias, the third party could then hand it over to the first party for model development.

The advantage of this variation is that the first party can develop whatever kind of model they like, without the risk of it learning biases from the training data. It also limits the involvement of the third party to a single step, after which the data could be deleted. Finally, it encourages the development of expertise on the part of the specialist third party and doesn't require the first party to have in-house knowledge about fairness-aware machine learning. The disadvantage of this approach is that the anonymisation techniques only provide a degree of (quantifiable) anonymity. There is a clear trade-off between degrees of anonymity and utility of the dataset (Loukides and Shao, 2008), such that useful datasets will still likely carry re-identification risks. To the extent that such risks persist, the first party could learn more about individuals' sensitive characteristics in this variation than it could in the other variations.

Who could act as a third party? We have thus far assumed the existence of a suitable trusted third party, but it is worth considering what kinds of organisations might fulfil this role. This will likely depend on which of the variations are adopted. Each might pose different requirements of trustworthiness, technical expertise and incentivisation. In the case of a third party whose role is merely to detect disparate impact, relatively little technical expertise would be required, making it suitable for organisations with fewer resources and technical skills. The fact that disparate impact is already the focus of many civil society groups' research activities may make them well situated to take on this role. Many potentially affected minority groups already have active representatives who could benefit from more formal auditing roles. Depending on the application context, it may be appropriate to involve different organisations; for instance, trade unions might be more equipped to address the fairness of algorithmic models deployed in human resources decisions.

If the third party is expected to be an ex ante discrimination mitigator, they will require more data collection and particular expertise in fairness-aware techniques. It may therefore need to be a specialist organisation, potentially working in collaboration

with appropriate civil society organisations. It could be anticipated that consultancy or accountancy firms might provide these services to corporate clients, as they do with other forms of social auditing.[7]

Another option might be statutory or chartered bodies whose remit includes monitoring discrimination, promoting equality, or enforcing law. For instance, the Equality and Human Rights Commission in the UK, or the Equal Employment Opportunity Commission in the US, are statutory bodies responsible for enforcing equalities laws. While traditionally involved in reviewing of individual cases for litigation, providing legal assistance and intervening in proceedings, these bodies could also take on more ongoing, data-driven monitoring of data-driven discrimination. Bodies more linked to data governance might help here too, such as the *Conseil national du numérique* (French Digital Council) or the data stewardship body recently recommended by the Royal Society and the British Academy (2017). State-sponsored API frameworks such as GOV.UK Verify, where the public sector certifies companies to provide verification services to third parties, might also serve as a framework to allow auditors to query trusted bodies for protected characteristics.

Knowledge bases about fairness in data and models

Experiential knowledge concerning the construction or attempted construction of ethical algorithmic systems has been largely neglected in the DADM and FATML communities. This has created a not insignificant knowledge gap that we believe has problematic consequences on-the-ground. This neglect is surprising for several reasons.

As data governance tools move increasingly towards ex ante prevention and anticipation of harms, particularly through data protection and privacy impact assessments (Binns, 2017; Wright and de Hert, 2012), relying solely on in-data analysis of unfairness appears not just at tension with on-the-ground regulatory needs – it could even be described as paradoxical. It certainly seems problematic to have to link the data and train a system before you can decide whether you should even be doing either of those things. Many organisations cannot legally or practically proceed with any data work, even basic data access, cleaning, linking or exploration, until this stage is passed. Yet DADM and FATML approaches often implicitly assume that all the ingredients are on the table to build the tool, and the only decision to be made is whether to deploy or not.

Machine learning is a generic technology with sector-specific applications. High profile, consequential domains have included anticipating the geospatial distribution of crime (Azavea, 2015; Perry et al., 2013; Wetenschappelijke Raad voor het Regeringsbeleid (WRR), 2016), the need for child protection (Vaithianathan et al., 2013) and the detection of tax fraud (Khwaja et al., 2011; Sharma and Kumar Panigrahi, 2012). Some ethical issues are sector- or even location-specific, but others are likely to be shared. Highly problematic issues might only appear rarely, limiting their propensity to capture with in-data analysis.

Limited implementation and education surrounding DADM and FATML technologies threatens our ability to cope with pressing issues in today's machine learning systems. Even though this research field has some history (Andrews et al., 1995; Custers et al., 2013; Hajian, 2013; Pedreshi et al., 2008; Vedder, 1999), usable software libraries remain largely unavailable, and little training exists. Given the current lack of practical ethics education in computer science curricula, rapid change seems unlikely (Goldweber et al., 2011, 2013; Spradling et al., 2008). A stopgap is sorely needed.

Diagnosing and addressing social and ethical issues in machine learning systems can be a high capacity task, and one difficult to plan and execute alone or from scratch. Ethical challenges or appropriate methods to tackle them might lurk within aspects of envisaged that are easy overlooked, such as hyperparameters, model structure, or quirks in data formatting or cleaning. Some issues that might arise might also not have their origins in the models or the data, but surrounding social, cultural and institutional contexts. Issues such as automation bias (Skitka et al., 1999), where individuals either place too much trust or too little trust in decision support systems, might be a synergistic result of both the model and the user interface. Other issues might have their origins in a model but likely solutions elsewhere. For example, for fairness grievances which are particularly difficult to detect or anticipate, better systems for decision subjects to feedback to decision-makers might be required. These issues might not have one-size-fits-all answers, but they are also unlikely to need to be treated as fresh each and every time they arise.

Issues of changing data populations and correlations are both currently under-emphasised in DADM/FATML work and appear difficult to fully address with in-data analysis. Concept drift or dataset shift refers to either real or virtual (differently sampled) changes in the conditional distributions of model inputs and outputs (Quiñonero-Candela et al., 2009) – for example, how changes in law might qualitatively affect prison population or the strategies of fraudsters. Fairness and transparency are not static but moving targets, and ensuring their reliability is important. But anticipating change is technically difficult. Knowledge around rates and causes of change can be tacit, obliging

us to carefully consider how best to use expert input (Gama et al., 2013). In particular, these phenomena can be hard to examine when changes are nuanced, or even are a result of the actions of previous machine learning supported decisions themselves. An important key role for domain experts going forward is to explain and record how and why certain types of concept drift occur, rather than just help in their detection (Žliobaitė et al., 2016).

Practical aspects of a knowledge base for fairness. Given the above factors, we propose that a structured, community-driven data resource containing practical experiences of fair machine learning and modelling could serve as a useful resource both in the direct absence of sensitive data, and more broadly in its own right. Such a resource, held online, would allow modellers to record experiences with problematic correlations and redundant encoding while modelling certain phenomena, as well as sociotechnical ethical issues more broadly (such as interpretability, reliability and automation bias), and detail the kinds of solutions and approaches they used or sought to remedy them. It could operate on a relatively open, trust-based model, such as Wikipedia, or have third-party gatekeepers, such as NGOs or sectoral regulators verifying contributions and attempting to instil anonymity where possible or desired. It would create a stepping-stone to enable practical, albeit rudimentary, fairness evaluations to be carried out today.

Linked data technologies have already seen significant adoption in sectors where cross-organisational collaboration around data is necessary (Bizer et al., 2009). This does not necessarily mean an industry-wide, comprehensive, rigid ontology for the purposes of addressing the ethical challenges of machine learning has to be adopted. Rather, a minimal adoption of common practices would enable different organisations to collaboratively annotate and describe the resource.

Several challenges would need to be addressed before such a database could be implemented. Similar variables and entities would need to be aligned in order to make such a dataset structured and navigable. Higher level common identifiers might be needed to group variables even if the levels of such variables were different. Some categorisations might have given individuals the chance to specify non-binary gender identities, or to opt out from this question – but this is unlikely to make any correlations or lessons found completely irrelevant or non-transferable in practice. Database ontologies should incorporate broader parts of the modelling process, such as cleaning or user interfaces, but the best format to do this is unclear. Arriving at it will likely be a result of trial-and-error.

Metadata should also be standardised. What kind of discrimination discovery methods were being utilised? How could effect strength or statistical significance be captured across these? It is likely that a descriptive vignette would also be useful, particularly concerning social processes and organisational context, but should or could this take a standardised format whilst remaining effective?

Such a dataset might benefit from discussion and input from different viewpoints both within the organisations submitting the information, but also externally. Open annotation or discussion technologies might contribute questions and context to the methods and content of dataset entries (Pellissier Tanon et al., 2016; Simperl and Luczak-Rösch, 2014; Vrandečić and Krötzsch, 2014). Technologies such as StackExchange, a question and answer network initially aimed at developers, but recently with wider adoption, have proved practically popular technical and social tools for solving issues around software. Such a database could take inspiration from the factors that make knowledge communities run effectively in these virtual environments. Allowing organisations to trace the sources of the data in such collaborative knowledge bases would also be key; in this respect, much could be learned from proposed solutions to similar challenges in scientific data collaboration (Missier et al., 2010).

Most data scientists are already used to working collaboratively online, through leading technologies in this space such as Git, MediaWiki, or StackExchange. Yet data scientists form only one part of the puzzle. As discussed, fairness issues can concern different parts of the modelling process, and as such viewpoints from others such as user interface developers, project managers and decision subjects would likely be valid and useful. The technologies chosen should be clear and accessible to those who are not used to working in these virtual spaces, whilst incorporating the features and extensibility that more developed solutions bring. If they are not, they are likely to become exclusionary and not see the widespread adoption that would make them most useful.

It is not just modellers who can contribute information to this knowledge base. Quantitative and qualitative findings in the research literature that might be relevant to particular fields or data sources could be added. For example, considerable amounts of research exist on areas such as financial literacy, recidivism or child protection which are carried out with the aims of improving their fields, but not directly to make or inform decision support or decision-making. These forms of evidence could be used to directly inform model structure, or to inform in-data analysis and search for ethical issues and concerns. Many of

these pieces of evidence are currently hard to locate – they are published across disciplines, behind paywalls, or with research questions that do not make clear the correlations that the research also unearths. In the medium term, text mining and natural-language processing might help populate such a database semi-automatically.

DADM/FATML methods, given their own technical opacity to laypersons, come with their own issues of transparency and legitimacy. Individuals are, under the GDPR, entitled to know when automated processing of their personal data is occurring, and for what purposes, although there are practical caveats regarding these rights (Edwards and Veale, 2017). Yet for them to understand the potential harms that could accrue to them by consenting is much trickier. Both they and trusted independent third parties usually lack the source data for investigative purposes. Even if they had it, it is unclear that it would be hugely useful or revealing given the rapidly changing nature of these datasets and the patterns within and the ample possibilities for data linkage that usually exist. Yet what they are (usually) interested in is not the data themselves, but the potentially problematic patterns the data support. An evidence base might help individuals or organisations understand what insights are held in different forms of data.

Potentially confounding issues. The proposal is largely grounded on the idea that organisations would be willing to spend time and money on cooperating to create a common resource. Primarily, this is a collective action problem, as there are great incentives to free ride and let others provide the information, which could result in non-provision (Olson, 1971). This is compounded by intellectual property concerns. If insights from data are viewed through an IP or a trade secrets lens, this could make organisations reticent to share.

Yet sharing of data for ethical purposes between firms is far from unheard of, particularly in other sectors facing similarly tricky societal challenges. Social and environmental issues in the global clothing sector are pervasive due to uncertainties around the environmental impact of processes, materials and chemicals, and uncertainties in the on-the-ground production systems characterised by multi-layered subcontracting. The Sustainable Apparel Coalition (SAC) emerged as a data-sharing body in 2010, now with over 180 members representing well over a third of all clothing and footwear sold on the planet. Together with the US Environmental Protection Agency (EPA), and with several large data donations and collection projects involving members, they have been developing the open-source Higg Index to give designers tools to better and more rigorously anticipate potential products' sustainability further upstream. In some ways, withholding data about ethical concerns and potentially salient social issues could itself be seen as a controversial, reputational risk.

Furthermore, the institutional field of the technology sector does not seem unamenable to this form of cooperation. Institutional fields create like-minded communities of practice through three main mechanisms – coercive pressure, where influence from actors or actants enforces homogeneity; mimetic pressures, which stem from standard, imitative responses to uncertainty; and normative pressures, which stem from how a field coalesces and becomes professionalised (DiMaggio and Powell, 1983). Some promising normative pressures can be seen across the machine learning modelling field that give hope for this – communities of voluntary support on question–answer networks such as Cross Validated[8] (which themselves support mimetic pressures); pro-bono data science for non-profits on the weekends through growing organisations like DataKind; virtual discussions and events from field leaders on /r/MachineLearning and Quora; expectations of contributions to open source software, to name a few. Proposed coercive pressures, such as professional bodies, charters or certification for data scientists might also play a role here in the future.

Identifying and creating databases of 'good' or 'best' practices is a common but also a problematic policy approach to complex socio-technical challenges. This approach can mislead, as practices are usually assumed to lead to good outcomes rather than being treated as hypotheses subject to serious monitoring and evaluation. Even where evidence suggests good practices work in one context, they may fail elsewhere (Cartwright and Hardie, 2012). Instead of prescribing 'good practice', a database of experiences would serve a more exploratory function. Several organisations are well positioned to start or collaborate on such initiatives: private think-tanks such as Data and Society in the United States, proposed bodies such as the national data stewardship body described in a recent report by the Royal Society and the British Academy (2017), or one of many interdisciplinary collaborative melding computer science and social science in universities across the world. It might also connect individuals facing similar challenges across the globe, creating creative, discussion-enabling support networks that help like-minded individuals share advice, strategies and even code to tackle the trickiest challenges together.

Exploratory fairness analysis (EFA)

The situations above assume that information on protected characteristics are either possible to obtain, or available in parallel cases. Yet there may be situations

where such data is restrictively difficult to obtain at all. Ambient computing, for example, judges people based on rather disembodied and abstracted features that environmental sensors can pick up, rather than through a data-entry method. Yet these systems might also exhibit fairness concerns; fairness concerns which might be particularly tricky to deal with.

These situations, where the protected data are not known, pose a difficult challenge for computational fairness tools. Yet we propose that there are concrete methods for these issues that while imperfect, could prove useful practices to both explore and develop in the future.

Building ex ante unfairness hypotheses with unsupervised learning methods. Before building the model, data can be examined for patterns that might lead to bias. Exploratory data analysis is a core part of data analysis, but teaching, research and practice into it has been historically marginalised (Behrens, 1997; Tukey, 1980). Results of previous research, such as DCUBE-GUI or D-Explorer, have shown how visual tools might help with the understanding of potentially discriminatory patterns in datasets (Gao, 2015; Gao and Berendt, 2011), even for novice users (Berendt and Preibusch, 2014). Still, as with other methods, these tools broadly come with the assumption that the sensitive characteristics are available in the dataset, which we have argued is often unrealistic.

If we assume that immediately sensitive data are unavailable, simply understanding the correlations in the dataset is of less use. Instead, the exploratory challenge can be seen primarily an unsupervised learning problem. Unsupervised learning attempts to draw out and formalise hidden structure in datasets. Through unsupervised learning, we can hope to build an idea of the structure of correlations within data. As we do not have the sensitive characteristics, confirmatory analysis is difficult. This does not mean there is nothing to be done. Exploratory data analysis has much to contribute in the building of hypothesis and the directing of future data and evidence collection as part of a broader process of due diligence.

A relevant subset of unsupervised learning methods we zoom in on here attempt to understand dataset structure through estimating latent variables that appear to be present. Some methods, such as principal component analysis (PCA), try to create a lower dimensional version of the data that captures as much variance as possible with a smaller number of variables. Some social science methods such as Q-methodology (McKeown and Thomas, 2013) use this approach to try and pick up latent dimensions such as subjective viewpoints. Other methods, such as Gaussian mixture models, assume that datasets are generated from several different Gaussian distributions, and attempt to locate and model these clusters.

These forms of analysis can be used to build hypotheses about fairness in datasets. For example, upon clustering or identifying subgroups within a dataset (which may or may not be related to any protected characteristics), these groups can be qualitatively examined, described and characterised. Experimental and sampling techniques might be used to gain more contextual information about the individuals in these clusters – for example, if their sensed or captured behaviour correlates with any sociodemographic attributes. These clusters can be used before or during the model building process to understand performance on different subgroups present in the data.

Building ex post unfairness hypotheses with interpretable models. A second approach to in-data analysis without access to protected characteristics examines trained models, rather than the input data alone. Once models have been trained, even complex models, there are several methods that are available for trying to understand their core logics in human-interpretable ways.

The literature on understanding models such as neural networks has traditionally distinguished between decompositional interpretation and pedagogical[9] interpretation (Andrews et al., 1995; Tickle et al., 1998). Decompositional approaches focus on how to represent patterns in data in a way that both optimises predictive performance whilst the internal logics remain semantically understandable to designers. Proponents of pedagogical systems on the other hand noted that not only was it difficult to get a semantically interpretable logic from models such as neural networks, although some try (Jin et al., 2006). The tactic they have adopted, which is broadly the domain of most current research in interpreting complex systems, is to see the interpretation as a separate optimisation problem to be considered.

The concept of pedagogically interpretable models is relatively simple to explain. The basic idea is to wrap a complex model with a simpler one, which through querying the more complex model like an oracle, can estimate its core logics. Candidates include logistic regression or decision trees. Increasingly, proposals for the analysis of more complex models acknowledge that the gap between the logics that can be represented by the simpler model and the logics latent in a more complex model are too vast to translate appropriately. Image recognition is a case in point. Instead, proposals in this area have tried to estimate the logics that locally surround a given input vector – such as an image – to understand why it was classified as it was (Ribeiro et al., 2016b).[10]

Exploratory fairness analysts might manually examine mechanisms behind a model's core logics and ask if they made sense. Specifically, analysts might wish to consider whether they would be happy publishing such information behind a model, or whether the public might take issue with the way and reasons behind decisions being made as they were. Some recent research that has highlighted gender bias in word embedding systems, which place words in relation to each other in high dimensional spaces to attempt to map different dimensions of their meaning, has gathered attention: and the methods of bias identification in this area are related to what we discuss here (Bolukbasi et al., 2017; Caliskan et al.. 2017). Future research should tangibly explore whether meaningful and relevant information about datasets or models known to be somehow biased can be discerned through this type of analysis.

Discussions and directions

Three approaches, three purposes

The three distinct approaches we have outlined in this paper point to three possible avenues for exploration in the research and practice of fairer machine learning. Each of them is suited for different purposes.

The third-party approach, where another organisation holds sensitive characteristics that they use to detect and potentially mitigate discrimination from data and models, is primarily useful where trust in the organisation interested in model building is low, or potential reputational risk is high. Insurance or hiring seem like prime cases here, particularly as they are areas historically associated with bias over protected variables. A challenge with this approach is that it is not easy to set up in low-resourced situations, or unilaterally.

The collaborative knowledge base approach, where linked databases featuring fairness issues noted and experienced by global researchers and practitioners, could be useful in a broad array of situations. It might provide benefit where general uncertainty is acute, risk assessment must be undertaken pre-emptively, or risks are complex, changing and socio-technical. Yet this requires a change of mindset. Organisations involved in modelling should overcome a reluctance to openly discuss their models, and will need to dedicate time and money to give to as well as take from such a shared resource. Anonymous contributions could work as a model, but issues of who verifies provenance of the information given, and how easily it is to re-identify organisations based on modelling purpose would abound.

The exploratory approach requires the least organisational set-up, as it can be undertaken unilaterally on data where sensitive characteristics are not held. Yet while this approach enables the construction of questions and the probing of certain types of anomalous or potentially problematic patterns in the data, on its own it provides by far the least assurance that fairness issues have been comprehensively identified, assessed and mitigated. Further work should seek to formalise methods of exploring data for these kinds of patterns, and test modellers and processes for their efficacy in identifying a range of synthetically induced issues.

There are, unsurprisingly, limits to the effectiveness of technological or managerial fixes to contested concepts such as fairness. Unsupervised learning is particularly challenging to evaluate fairness upon, given that groups discovered are latent, although there has been some recent work beginning to explore this space (Chierichetti et al., 2017). Understanding fairness by demographic will also be hard to grasp when those demographics are latent – such as treating individuals holding particular political views similarly in regards to moderating content online (Binns et al., 2017). More importantly, even though the three approaches we outline deal with different levels of formality and different ways of understanding or conceiving fairness, they all remain broadly centred on the software artefacts themselves. We do not suggest that either these approaches or the broad mindsets that underpin them are sufficient for understanding equity or mitigating discrimination in a digital age. We do, however, tentatively suggest that where these software artefacts are used to make and support decisions, tackling technical aspects of these issues is likely a necessary piece of the puzzle – neither more nor less important than others, such as organisational culture, social methods of oversight, or decisions about the intention or direction of deployment. We also would draw attention to larger challenges with predictive systems: that they might not achieve social or policy goals at all by their nature (Harcourt, 2006), or that fairness might not be the most relevant issue as much as ideas of stigmatisation, over-surveillance, or the devaluing of particular cultural notions, such as family units (Blank et al., 2015). Where there are inherent conflicting interests between organisations deploying such systems and those affected by them, co-operation may not be feasible or desirable; affected groups may instead be drawn (understandably) to more adversarial forms of resistance and political action (Brunton and Nissenbaum, 2015; Danaher, 2016; Lyon, 2007).

Directions for empirical research

These three proposals illustrate how alternative institutional set-ups and ways of knowing might help in the

governance of fairness in the context of machine learning. It focuses on one identified practical constraint – the absence of sensitive data. Each approach introduces limitations, caveats and provides few guarantees of performance. This might irritate researchers in this space, yet it reflects the messy reality of many contemporary on-the-ground situations.

We believe there are opportunities amidst the constraints. The practical limitations of fairness-improving approaches, including these three, will only become apparent upon their introduction and reflexive study within real-world settings. In particular, our second and third suggestions, concerning knowledge bases and exploratory analyses, are not amenable to the sort of mathematical guarantees that the DADM literatures may find comforting. In these situation, *process evaluation* is much more important than *outcome evaluation*. Understanding the questions and challenges that these methods do (or do not) address during the real building, deployment and management of predictive systems is key here. Only a small amount of work has been done in this space (see Veale, 2017 for one example), and we argue strongly that this should increase. As we have noted, it is often unrealistic to assume mathematically sound 'debiasing' on-the-ground is possible, and this means it often unhelpful to apply the validity conditions of traditional research in statistics and computer science to discrimination-aware machine learning. New technologies of this type should be at least partially assessed on the extent of new capabilities for responsible practices they afford practitioners – a difficult, transdisciplinary and heavily value-laden task, but a very necessary one.

Without this dimension, designed tools are likely to stumble in surprising and even mundane ways, which will affect their ability to deal with unfairness and discrimination in the wild. It seems unlikely that statistical guarantees of fairness will translate smoothly to individuals feeling that decisions about them were made fairly – something as much a result of process as of outcome. Researchers working in this space should trial their proposed solutions, monitoring their implementation using rich and rigorous qualitative methods such as ethnography and action research, and feed findings from this back into tool revision and rethinking. To adequately address fairness in the context of machine learning, researchers and practitioners working towards 'fairer' machine learning need to recognise that this is not just an abstract constrained optimisation problem. It is a messy, contextually-embedded and necessarily sociotechnical problem, and needs to be treated as such. This requires technical scholars to better grasp the social challenges and contexts; but also for social scholars to grapple more rigorously with the technical proposals placed on the table, and to ensure that critiques with operational implications reach the ears of the computing community.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The UK Engineering and Physical Science Research Council (EPSRC) provided support to both Michael Veale (grant number EP/M507970/1) and to Reuben Binns (SOCIAM: The Theory and Practice of Social Machines, grant number EP/J017728/1).

Notes

1. Some of these measures have obvious shortcomings. In particular, disparate impact has been criticised because it fails to account for discrimination which is explainable in terms of legitimate grounds (Dwork et al., 2012). For instance, attempting to enforce equal impact between men and women in recidivism prediction systems, if men have higher reoffending rates, could result in women remaining in prison longer despite being less likely to reoffend.
2. Some have argued that a principle of data min*mum*isation would enable better governance of these issues, rather than mimalisation (van der Sloot, 2012). In some ways, it could be argued that the existing regulation could be already read through such a lens, but text interpretation is not the focus of this paper.
3. For consideration of this question, see van der Sloot (2012) and Žliobaitė and Custers (2016).
4. See the Web Cryptography API recommendation from W3C https://www.w3.org/TR/WebCryptoAPI/
5. A range of privacy preserving communication solutions could be applicable here, yet we do not seek to treat the modelling organisation as a malicious adversary. The methods here implicitly focus on organisations actively wishing to increase trust and reduce discriminatory outcomes.
6. Many researchers (e.g. Ossia et al., 2017) are currently exploring how to bring parts of model training activities away from cloud servers and onto a user's own device for the purposes of increasing privacy, utilising mathematical tools such as homomorphic encryption or zero-knowledge proofs, and integration of fairness into these more decentralised systems will likely be a research area for future exploration.
7. Some relevant consultancies already offer services in this space, such as Trilateral Research (http://trilateralresearch.com/services/impact-assessment/) or ORCAA (http://www.oneilrisk.com/).
8. Cross Validated is the statistical question–answer site on the StackExchange network, http://stats.stackexchange.com/

9. Pedagogical interpretation has recently been described as 'model-agnostic' interpretation (Ribeiro et al., 2016a).
10. This method is not exclusive to pedagogical methods, and some recent work has shown how decompositional methods, which use components of model structure rather than just treating it like a black box, also display strong promise in this space (Montavon et al., 2017).

References

Agrawal R and Srikant R (2000) Privacy-preserving data mining. In: *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, New York, NY, 16–18 May, pp. 439–450. Available at: http://dx.doi.org/10.1145/335191.335438.

Andrews R, Diederich J and Tickle AB (1995) Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems* 8(6): 373–389.

Angwin J, Larson J, Mattu S, et al. (2016) Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica* 23 May.

Azavea (2015) *HunchLab: Under the Hood*. Philadelphia, PA: Author. Available at: http://cdn.azavea.com/pdfs/hunchlab/HunchLab-Under-the-Hood.pdf.

Barocas S and Selbst AD (2016) Big Data's disparate impact. *California Law Review* 104: 671–732.

Behrens JT (1997) Principles and procedures of exploratory data analysis. *Psychological Methods* 2(2): 131–160.

Berendt B and Preibusch S (2014) Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. *Artificial Intelligence and Law* 22(2): 175–209.

Berk R, Heidari H, Jabbari S, et al. (2017) Fairness in criminal justice risk assessments: The state of the art. *arXiv* [stat.ML]. Available at: http://arxiv.org/abs/1703.09207 (accessed 20 April 2017).

Binns R (2017) Data protection impact assessments: A meta-regulatory approach. *International Data Privacy Law* 7(1): 22–35.

Binns R, Veale M, Van Kleek M, et al. (2017) Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In: Ciampaglia G, et al. (eds) *Social informatics: 9th international conference*, SocInfo 2017, Oxford, UK, 13–15 September, Proceedings, Part II. Cham: Springer, pp. 405–415. Available at: https://doi.org/10.1007/978-3-319-67256-4_32.

Bizer C, Heath T and Berners-Lee T (2009) Linked data – The story so far. In: Sheth A (ed.) *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. Hershey, PA: IGI Global, pp. 205–227. Available at: http://dx.doi.org/10.4018/978-1-60960-593-3.ch008.

Blank A, et al. (2015) *Ethical Issues for Maori in Predictive Risk Modelling to Identify New-Born Children who are at High Risk of Future Maltreatment*. Wellington, NZ: Ministry of Social Development/Te Manatu Whakahiato Ora. Available at: https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/research/predictive-modelling/00-ethical-issues-for-maori-in-predictive-risk-modelling.pdf (accessed 16 February 2017).

Bolukbasi T, Chang K-W, Zou J, et al. (2017) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, 5–10 December 2016, Barcelona, Spain.

Bowker G and Star SL (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: The MIT Press.

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

Brunton F and Nissenbaum H (2015) *Obfuscation: A User's Guide for Privacy and Protest*. Cambridge: MIT Press.

Burrell J (2016) How the machine ''thinks'': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.

Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.

Cartwright N and Hardie J (2012) *Evidence-based Policy: A Practical Guide to doing it Better*. Oxford: Oxford University Press.

Chierichetti F, Kumar R, Lattanzi S, et al. (2017) Fair clustering through fairlets. Presented as a talk at the 4th Workshop on Fairness, Accountability, Transparency in Machine Learning (FAT/ML 2017), Halifax, Nova Scotia, Canada. Available at: http://www.fatml.org/media/documents/fair_clustering_through_fairlets.pdf (accessed 30 August 2017).

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Presented as a poster at FAT/ML 2016, New York. Available at: http://arxiv.org/abs/1703.00056 (accessed 16 February 2017).

Custers B, Calders T, Schermer B, et al. (eds) (2013) *Discrimination and Privacy in the Information Society*. Berlin: SpringerAvailable at: http://dx.doi.org/10.1007/978-3-642-30487-3.

Danaher J (2016) The threat of algocracy: Reality, resistance and accommodation. *Philosophy and Technology* 29(3): 245–268.

Dieterich W, Mendoza C and Brennan T (2016) COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe, July 2016. Available at: http://www.northpointeinc.com/northpointe-analysis (accessed 2 April 2017).

DiMaggio PJ and Powell WW (1983) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review* 48(2): 147–160.

Dwork C, Hardt M, Pitassi T, et al. (2012) Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, Cambridge, MA, USA, 8–10 January, pp. 214–226. Available at: http://doi.acm.org/10.1145/2090236.2090255.

Edwards L and Veale M (2017) Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law and Technology Review* 15(2), http://doi.org/10.2139/ssrn.2972855.

Feldman M, Friedler SA, Moeller J, et al. (2015) Certifying and removing disparate impact. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Sydney, NSW, Australia, 10–13 August. Available at: http://dx.doi.org/10.1145/2783258.2783311.

Gama J, Žliobaitė I, Bifet A, et al. (2013) A survey on concept drift adaptation. *ACM Computing Surveys* 1(1): 1–35.

Gao B (2015) *Exploratory visualization design towards online social network privacy and data literacy*. PhD Thesis, KU Leuven. Available at: https://lirias.kuleuven.be/bitstream/123456789/512067/1/thesis.pdf (accessed 16 February 2017).

Gao B and Berendt B (2011) Visual data mining for higher-level patterns: Discrimination-aware data mining and beyond. In: *Benelearn 2011. Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning (Benelearn 2011)*, The Hague, 20 May, pp. 45–52. Available at: https://lirias.kuleuven.be/handle/123456789/330089.

Gellert R, de Vries K, de Hert P, et al. (2013) A comparative analysis of anti-discrimination and data protection legislations. In: Custers B, Calders T, Schermer B, et al. (eds) *Discrimination and Privacy in the Information Society*. Berlin: Springer, pp. 61–89. Available at: http://dx.doi.org/10.1007/978-3-642-30487-3_4.

Goldweber M, Barr J, Clear T, et al. (2013) A framework for enhancing the social good in computing education: A values approach. *ACM Inroads* 4(1): 58–79.

Goldweber M, Davoli R, Little JC, et al. (2011) Enhancing the social issues components in our computing curriculum: Computing for the social good. *ACM Inroads* 2(1): 64–82.

Hacking I (1995) The looping effects of human kinds. In: Premack D, Sperber D and Premack AJ (eds) *Causal Cognition: A Multidisciplinary Debate*. Oxford: Clarendon Press, pp. 351–383.

Hajian S (2013) *Simultaneous discrimination prevention and privacy protection in data publishing and mining*. PhD Thesis, Universitat Rovira I Virgili. Available at: https://arxiv.org/abs/1306.6805.

Hajian S and Domingo-Ferrer J (2012) A study on the impact of data anonymization on anti-discrimination. In: *IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012)*, 10 December 2012, Brussels, Belgium. pp. 352–359. Available at: http://dx.doi.org/10.1109/ICDMW.2012.19.

Hajian S and Domingo-Ferrer J (2013) A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering* 25(7): 1445–1459.

Hajian S, Domingo-Ferrer J and Farràs O (2014) Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery* 28(5–6): 1158–1188.

Harcourt BE (2006) *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago, IL: University of Chicago Press.

Hardt M, Price E and Srebro N (2016) Equality of opportunity in supervised learning. In: Lee DD, Sugiyama M, Luxburg UV, et al. (eds) *Advances in Neural Information Processing Systems 29*, Red Hook, NY: Curran Associates, Inc., pp. 3315–3323. Available at: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf (accessed 16 February 2017).

Hellman D (2008) *When is Discrimination Wrong?* Cambridge, MA: Harvard University Press.

Hildebrandt M (2015) *Smart Technologies and the End(s) of Law*. Cheltenham: Edward Elgar.

Hildebrandt M and Gutwirth S (eds) (2008) *Profiling the European Citizen: Cross-Disciplinary Perspectives*. Dordrecht: Springer.

Huang L, Joseph AD, Nelson B, et al. (2011) Adversarial machine learning. *AISec'11*. Available at: http://dx.doi.org/10.1145/2046684.2046692.

Hundepool A, Domingo-Ferrer J, Franconi L, et al. (2012) *Statistical Disclosure Control*. London: John Wiley & Sons.

Japkowicz N and Shah M (2011) *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge: Cambridge University Press.

Jin Y, Sendhoff B and Körner E (2006) Simultaneous generation of accurate and interpretable neural network classifiers. In: Jin Y (ed.) *Multi-objective Machine Learning*, Dordrecht: Springer, pp. 291–312. Available at: http://dx.doi.org/10.1007/3-540-33019-4_13.

Kamiran F, Calders T and Pechenizkiy M (2012) Techniques for discrimination-free predictive models. In: Custers B, Calders T, Schermer B, et al. (eds) *Discrimination and Privacy in the Information Society*. Berlin: Springer, pp. 223–240.

Keats Citron D and Pasquale F (2014) The scored society: Due process for automated predictions. *Washington Law Review* 89(1): 1–33.

Khwaja MS, Awasthi R and Loeprick J (2011) *Risk-Based Tax Audits: Approaches and Country Experiences*. New York, NY: World Bank.

Kitchin R and Dodge M (2011) *Code/space: Software and Everyday Life*. London: MIT Press.

Kleinberg J, Mullainathan S and Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv* [cs.LG]. Available at: http://arxiv.org/abs/1609.05807 (accessed 16 February 2017).

Kroll JA, Huey J, Barocas S, et al. (2016) Accountable algorithms. *University of Pennsylvania Law Review* 165: 633–705.

Kusner MJ, Loftus JR, Russell C, et al. (2017) Counterfactual fairness. *arXiv* [stat.ML]. Available at: http://arxiv.org/abs/1703.06856 (accessed 16 April 2017).

Latour B (1999) *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge, MA: Harvard University Press.

Loukides G and Shao J (2008) Data utility and privacy protection trade-off in k-anonymisation. In: *Proceedings of the 2008 International Workshop on Privacy and Anonymity in the Information Society (PAIS)*, March 25, pp. 36–45. Nantes, France: ACM, Available at: http://

dl.acm.org/citation.cfm?doid=1379287.1379296 (accessed 27 February 2017).

Lyon D (2007) *Resisting Surveillance*. *The Surveillance Studies Reader*. London: McGraw-Hill Education, pp. 368–377.

McDaniel P, Papernot N and Berkay Celik Z (2016) Machine learning in adversarial settings. *IEEE Security & Privacy* 14(3): 68–72. http://dx.doi.org/10.1109/msp.2016.51.

McKeown B and Thomas DB (2013) *Q Methodology*. London: SAGE. Available at: http://dx.doi.org/10.4135/9781483384412.

Mantelero A (2016) Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection. *Computer Law & Security Review* 32(2): 238–255.

Missier P, Ludäscher B, Bowers S, et al. (2010) Linking multiple workflow provenance traces for interoperable collaborative science. In: *5th Workshop on Workflows in Support of Large-Scale Science (WORKS)*, 14 November 2010, New Orleans, LA, USA. pp. 1–8. Available at: http://dx.doi.org/10.1109/WORKS.2010.5671861.

Montavon G, Lapuschkin S, Binder A, et al. (2017) Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65: 211–222.

Nissenbaum H (1996) Accountability in a computerized society. *Science and Engineering Ethics* 2(1): 25–42.

Olson M (1971) *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Ossia SA, Shamsabadi AS, Taheri A, et al. (2017) A hybrid deep learning architecture for privacy-preserving mobile analytics. *arXiv*. Available at: https://arxiv.org/abs/1703.02952 (accessed 30 June 2017).

Pasquale F (2010) Beyond innovation and competition: The need for qualified transparency in internet intermediaries. *Northwestern University Law Review* 104(1): 105–174.

Pedreshi D, Ruggieri S and Turini F (2008) Discrimination-aware data mining. *KDD '08*. Las Vegas, NV: ACM, pp. 24–27 August.

Pellissier Tanon T, Vrandečić D, Schaffert S, et al. (2016) From freebase to Wikidata: The great migration. In: *Proceedings of the 25th International Conference on World Wide Web*, WWW 2016, Montreal, Canada, 11–15 April, Committee, pp. 1419–1428. Available at: http://dx.doi.org/10.1145/2872427.2874809.

Perry WL, McInnis B, Price CC, et al. (2013) *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Washington, DC: RAND Corporation. Available at: http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf (accessed 16 February 2017).

Pinsker J (2015) What your choice of browser says about you as an employee. *The Atlantic*. Available at: http://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/ (accessed 25 February 2017).

Quiñonero-Candela J, Sugiyama M, Schwaighofer A, et al. (2009) *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press.

Ribeiro MT, Singh S and Guestrin C (2016a) *Model-agnostic interpretability of machine learning*. *Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York, NY. Available at: https://arxiv.org/abs/1606.05386 (accessed 16 February 2017).

Ribeiro MT, Singh S and Guestrin C (2016b) "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Sydney, NSW, Australia, 10–13 August, 2015, pp. 1135–1144. Available at: http://dx.doi.org/10.1145/2939672.2939778.

Rouvroy A (2011) Technology, virtuality and utopia. In: Hildebrandt M and Rouvroy A (eds) *Law, Human Agency and Autonomic Computing*. London: Routledge, pp. 119–141.

Sandvig C (2015) Seeing the sort: The aesthetic and industrial defence of 'the algorithm'. *Media-N* 11(1): 35–51.

Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. Paper presented to "*Data and Discrimination: Converting Critical Concerns into Productive Inquiry*," a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA. Available at: https://pdfs.semanticscholar.org/b722/7cbd34766655dea10-d0437ab10df3a127396.pdf (accessed 16 February 2017).

Sharma A and Kumar Panigrahi P (2012) A review of financial accounting fraud detection based on data mining techniques. *IJCAI* 39(1): 37–47.

Simperl E and Luczak-Rösch M (2014) Collaborative ontology engineering: A survey. *Knowledge Engineering Review* 29(1): 101–131.

Skitka LJ, Mosier KL and Burdick M (1999) Does automation bias decision-making? *International Journal of Human–Computer Studies* 51: 991–1006.

Spradling C, Soh L-K and Ansorge C (2008) Ethics training and decision-making: Do computer science programs need help? In: *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education*, SIGCSE 2008, Portland, OR, USA, 12–15 March, ACM, pp. 153–157. Available at: http://dx.doi.org/10.1145/1352322.1352188.

The Royal Society and the British Academy (2017) Data management and use: Governance in the 21st century. Available at: http://royalsociety.org/topics-policy/projects/data-governance/ (accessed 16 July 2017).

Tickle AB, Andrews R, Golea M, et al. (1998) The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks* 9(6): 1057–1068.

Tramèr F, Zhang F, Juels A, et al. (2016) Stealing machine learning models via prediction APIs. Available at: https://www.usenix.org/sites/default/files/conference/protected-files/security16_slides_tramer.pdf (accessed 16 February 2017).

Tukey JW (1980) We need both exploratory and confirmatory. *The American Statistician* 34(1): 23–25.

Tutt A (2016) An FDA for algorithms. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2747994 (accessed 16 February 2017).

Vaithianathan R, Maloney T, Putnam-Hornstein E, et al. (2013) Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine* 45(3): 354–359.

van der Sloot B (2012) From data minimization to data mini*mum*mization. In: Custers B, Calders T, Schermer B, et al. (eds) *Discrimination and Privacy in the Information Society*. Berlin: Springer, pp. 273–288.

Veale M (2017) Logics and practices of transparency and opacity in real-world applications of public sector machine learning. Presented as a talk at the 4th Workshop on Fairness, Accountability, Transparency in Machine Learning (FAT/ML 2017), Halifax, Nova Scotia. Available at: https://arxiv.org/abs/1706.09249 (accessed 16 August 2017).

Vedder A (1999) KDD: The challenge to individualism. *Ethics and Information Technology* 1(4): 275–281.

Verwer S and Calders T (2013) Introducing positive discrimination in predictive models. In: Custers B, Calders T, Schermer B, et al. (eds) *Discrimination and Privacy in the Information Society*. Berlin: Springer, pp. 255–270. Available at: http://dx.doi.org/10.1007/978-3-642-30487-3_14.

Vrandečić D and Krötzsch M (2014) Wikidata: A free collaborative knowledgebase. *Communications of the ACM* 57(10): 78–85.

Wetenschappelijke Raad voor het Regeringsbeleid (WRR) [Dutch Scientific Council for Government Policy] (2016) *Big Data in een vrije en veilige samenleving* [*Big Data in a free and safe society*]. Den Haag: WRR. Available at: http://www.wrr.nl/publicaties/publicatie/article/big-data-in-een-vrije-en-veilige-samenleving/ (accessed 16 February 2017).

Willenborg L and de Waal T (2012) *Elements of Statistical Disclosure Control*. Berlin: Springer.

Wright D and de Hert P (2012) *Privacy Impact Assessment*. Dordrecht: SpringerAvailable at: http://dx.doi.org/10.1007/978-94-007-2543-0.

Zafar MB, Valera I, Rodriguez MG, et al. (2016) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *arXiv* [stat.ML]. Available at: http://arxiv.org/abs/1610.08452.

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

Žliobaitė I, Kamiran F and Calders T (2011) Handling conditional discrimination. In: *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW 2011)*, 11 December 2011, Vancouver, Canada. pp. 992–1001. Available at: http://dx.doi.org/10.1109/ICDM.2011.72.

Žliobaitė I, Pechenizkiy M and Gama J (2016) An overview of concept drift applications. In: Japkowicz N and Stefanowski J (eds) *Big Data Analysis: New Algorithms for a New Society*, *Studies in Big Data*. Berlin: Springer, pp. 91–114. Available at: http://dx.doi.org/10.1007/978-3-319-26989-4_4.