

#BigData: Discrimination in data-supported decision making

FRA Focus

We live in a world of big data, where technological developments in the area of machine learning and artificial intelligence have changed the way we live. Decisions and processes concerning everyday life are increasingly automated, based on data. This affects fundamental rights in various ways. The intersection of rights and technological developments warrants closer examination, prompting the Fundamental Rights Agency to research this theme.

This focus paper specifically deals with discrimination, a fundamental rights area particularly affected by technological developments. When algorithms are used for decision making, there is potential for discrimination against individuals. The principle of non-discrimination, as enshrined in Article 21 of the Charter of Fundamental Rights of the European Union (EU), needs to be taken into account when applying algorithms to everyday life. This paper explains how such discrimination can occur, suggesting possible solutions. The overall aim is to contribute to our understanding of the challenges encountered in this increasingly important field.

Contents

1.	Big data and fundamental rights implications	2
2.	Data-supported decision making: predictions, algorithms and machine learning	3
3.	Computers "learning to discriminate"	4
4.	Detecting and avoiding discrimination	6
5.	Possible ways forward: addressing fundamental rights and big data	10
Pof	References	

Big data and fundamental rights implications

In the past decades, technological advancements have changed the way we live and organise our lives. These changes are inherently connected with the proliferation and use of big data.

Big data generally refers to technological developments related to data collection, storage, analysis and applications. It is often characterised by the increased volume, velocity and variety of data being produced ("the three Vs"), and typically refers (but is not limited) to data from the internet. Big data comes from a variety of sources, including social media data or website metadata. The Internet of Things (IoT) contributes to big data, including behavioural location data from smartphones or fitness tracking devices. In addition, transaction data from the business world form a part of big data, such as providing information on payments and administrative data.² The increased availability of data has led to improved technologies for analysing and using data - for example, in the area of machine learning and artificial intelligence (AI).

Arguably, big data can enhance our lives – for instance, in the health sector, where personalised diagnosis and medicine can lead to better care.3 However, the negative fundamental rights implications of big data-related technologies have only recently been acknowledged by public authorities and international organisations.4 The use of new technologies and algorithms, including machine learning and AI, affects several fundamental rights. These include, but are not limited to: the right to a fair trial, prohibition of discrimination, privacy, freedom of expression, and the right to an effective remedy, as outlined by the Council of Europe in their 2017 report.⁵ In addition, the European Parliament adopted a 2017 resolution highlighting the need for action in the area of big data and fundamental rights implications.6 In this resolution, the European Parliament uses the strongest language when referring to the threat of discrimination through the use of algorithms. This serves to underpin the particular focus of this paper.

Other FRA reports have highlighted the fundamental rights challenges posed by technologies that are built on big data. This includes FRA's reporting on the oversight of surveillance by national intelligence authorities, as requested by the European Parliament. In addition, FRA has published reports on the use of biometric and related data in the EU's large-scale IT databases. Starting with the present paper, FRA explores the implications of big data and AI regarding fundamental rights, focusing in the first instance on discrimination as a key area of EU legal competence, and also an area where FRA has undertaken extensive research to date.

European Parliament resolutions on big data and artificial intelligence

To use big data for commercial purposes and in the public sector, the European Parliament "calls on the European Commission, the Member States and the data protection authorities to identify and take any possible measures to minimise algorithmic discrimination and bias and to develop a strong and common ethical framework for the transparent processing of personal data and automated decision-making that may guide data usage and the ongoing enforcement of Union law".

When it comes to the use of big data for law enforcement purposes, the Parliament "warns that [...] maximum caution is required in order to prevent unlawful discrimination and the targeting of certain individuals or groups of people defined by reference to race, colour, ethnic or social origin, genetic features, language, gender expression or identity, sexual orientation, residence status, health or membership of a national minority which is often the subject of ethnic profiling or more intense law enforcement policing, as well as individuals who happen to be defined by particular characteristics".*

Moreover, the European Parliament has highlighted the need for ethical principles concerning the development of robotics and artificial intelligence for civil use. It points out that a guiding ethical framework should be "based on [...] the principles and values enshrined in Article 2 of the Treaty on European Union and in the Charter of Fundamental Rights, such as human dignity, equality, justice and equity, non-discrimination, informed consent, private and family life and data protection", among other principles.**

^{*} European Parliament (2017a).

^{**} European Parliament (2017b).

¹ Additionally, terms such as increased "variability" with respect to consistency of data over time, "veracity" with respect to accuracy and data quality, and "complexity" in terms of how to link multiple datasets can be added to the list of characteristics of big data (Callegaro, M. and Yang, Y. (2017)).

² Callegaro, M. and Yang, Y. (2017).

³ Seitz, C. (2017), pp. 298-299.

⁴ European Parliament (2017a); Council of Europe (2017a); European Data Protection Supervisor (EDPS) (2016); The White House (2016).

⁵ Council of Europe (2017b).

⁶ European Parliament (2017a).

Focus on discrimination

Direct or indirect discrimination through the use of algorithms using big data is increasingly considered as one of the most pressing challenges of the use of new technologies. This paper addresses selected fundamental rights implications related to big data, focusing on the threat of discrimination when using big data to support decision making.

Previously, decisions and processes were undertaken with little support by computers. Nowadays, there is an increase in the use of sophisticated techniques of (statistical) data analysis to facilitate these tasks. However, this can lead to discrimination. For example, this may include the automated selection of candidates for job interviews based on predicted productivity. Another example is the use of risk scores in assessing the credit worthiness of individuals applying for loans. Furthermore, in the course of a trial, the use of risk scores in the decision-making process on sentencing can lead to discrimination.

The principle of non-discrimination is embedded in EU law. Article 21 of the EU Charter of Fundamental Rights prohibits discrimination based on several grounds, including: sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age, sexual orientation, and nationality. Information about or related to these attributes, once processed as data, is connected to the individual. By

definition, this makes them personal data that are protected under the data protection legal framework. This is particularly important, due to the growing availability and processing of large amounts of data that also include information (potentially) indicating one or more of these characteristics relating to individuals.

At EU level, the General Data Protection Regulation (GDPR)⁷ addresses some of these new technological developments, including the potential for discrimination. The fundamental rights implications of using algorithms in decision making include considerations addressed in the GDPR, but also go beyond.

The issues raised here highlight the potentially problematic nature of the use of data for decision making. However, the use of data to inform decisions is also considered a positive development, as it potentially allows for more objective and informed decisions, in comparison to decisions that do not take into account available data. It also has the potential to limit discriminatory treatment based on human decision making that is derived from existing prejudices. While the limits of data and data analysis need to be taken into account, decisions supported by data are potentially better decisions than those without any empirical support. Algorithms can, in turn, be used to identify systematic bias and potentially discriminatory processes. Therefore, big data also presents opportunities for assessing fundamental rights compliance.

Data-supported decision making: predictions, algorithms and machine learning

With the increased availability and use of data, decisions are increasingly being facilitated or sometimes even completely taken over by using so-called predictive modelling methods – often referred to as the use of algorithms. Using data to predict incidents or behaviour is a major part of developments related to machine learning and Al. A classic example of using algorithms based on data analysis, a tool most people experience every day, is the spam filter. The algorithm has 'learned' – with some level of certainty – to identify whether an email is spam and to block it.

A basic understanding of how algorithms support decisions is essential, to allow experts and practitioners from other fields to enter the discussion

and to increase awareness and technical literacy. Furthermore, it is important to be able to identify and ask the right questions about potential problems that arise when using algorithms, particularly when it comes to discrimination.

Creating algorithms to make predictions may involve different methods, all of which use so-called 'training data' to find out which calculations predict a certain outcome most accurately. For example, a set of several thousands of emails, identified as either 'spam' or 'not spam', is used to identify characteristics that define differences between the two

⁷ General Data Protection Regulation.

groups of emails. In this example, characteristics may include specific words and combinations of words within emails. Through this process, rules for identifying spam are established. Many different calculations (i.e. algorithms) can be used and the best performing one is selected, i.e. the calculation that can categorise most cases correctly.

It is critically important to note that the output of algorithms is always based on probability, which means that there is uncertainty attached to the classifications made. As we can see in our daily lives, the methods can work quite well, but they are not infallible. Sometimes spam passes through to our email inbox, so-called 'false negatives' (i.e. erroneously, it is not identified as spam). Less frequently, a fully legitimate email might be suppressed by the filter, a so-called 'false positive'. The rate of true positives, the rate of true negatives and the tradeoff between these two rates are commonly used to assess a classification problem, such as detecting spam. There are several rates and indicators available that can be analysed for the assessment of how well an algorithm works.

What is an algorithm?

The term 'algorithm' is widely used in the context of big data, machine learning and Al. An algorithm is a sequence of commands for a computer to transform an input into an output. For example, a list of persons is to be sorted according to their age. The computer takes the ages of people on the list (input) and produces the new ranking of the list (output).

In the area of machine learning, several algorithms are used, which could also be referred to as ways of calculating desired predictions with the use of data. Many of these algorithms are statistical methods and most of them are based on so-called 'regression methods'. These are the most widely used statistical techniques for calculating the influence of a set of data on a selected outcome. For example, consider calculating the average influence of drinking alcohol on life expectancy. Using existing data, the average amount of alcohol a person drinks is compared to their life expectancy. Based on these calculations, life expectancy can be calculated and predicted for other persons simply by taking into consideration the amount of alcohol a person drinks, assuming a correlation exists. The algorithm used depends on the way the data are presented (e.g. whether it is numerical or textual data) and the goal of the calculation (e.g. prediction, explanation, grouping of cases). In machine learning, often several algorithms are tested to see which one has the best performance in predicting the outcome.

The creation of algorithms for prediction is a complex process that involves many decisions made by several people who are variously involved in the process. Therefore, it does not only refer to rules followed by a computer, but also to the process of collecting, preparing and analysing data. This is a human process that includes several stages, involving decisions by developers and managers. The statistical method is only part of the process for developing the final rules used for prediction, classification or decisions.

Computers "learning to discriminate"

Using data and algorithms for prediction can considerably facilitate decisions, as it allows for the revelation of patterns that cannot be otherwise identified. However, an algorithm can contribute to discriminatory decision making.

As specified under EU legislation, discrimination is illegal with respect to several conditions related to employment, access to social services and

According to the Racial Equality Directive (2000/43/EU), discrimination occurs "where one person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of racial or ethnic origin".

education. It is also forbidden when it is 'indirect': "where an apparently neutral provision, criterion or practice would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate". Similarly, Article 11 (3) of the Police Directive (Directive (EU) 2016/680) explicitly prohibits any 'profiling' that results in discrimination on the basis of special categories of personal data, such as race, ethnic origin, sexual orientation,

Council Directive 2000/43/EC of 19 July 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180/22. See also FRA (2010); FRA (2011).

political opinion, or religious beliefs.⁹ Whereas the term 'discrimination' was absent from the previous Data Protection Directive (95/46/EC), the need to prevent discrimination as a result of automated decision making is emphasised in the new GDPR. Automated decision making, including profiling, with significant effects on the data subject is forbidden, according to Article 22 of the GDPR, subject to specific exceptions.

"In order to ensure fair and transparent processing in respect of the data subject, (...) the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons (...)."

General Data Protection Regulation, Recital 71

Academics and practitioners are increasingly researching ways to detect and repair algorithms that can potentially discriminate against individuals or certain groups on the basis of particular attributes - for example, sex or ethnic origin. 10 This happens when the predicted outcome for a particular group is systematically different from other groups and therefore one group is consistently treated differently to others. For example, in cases where a member of an ethnic minority has a lower chance of being invited to a job interview because the algorithm was 'trained', based on data where their particular group performs worse, i.e. has worse outcomes than other groups. As a result, they may not be invited to a job interview. This can occur when the data used to train the algorithm include information regarding protected characteristics (e.g. gender, ethnicity, religion). Furthermore, so-called 'proxy information' is sometimes included in the data. This may include the height of a person, which correlates with gender, or a postcode, which can indirectly indicate ethnic origin in cases of segregated areas in cities, or more directly, a person's country of birth. Unequal outcomes and differential treatment, especially relating to proxy information, need to be assessed to see if they amount to discrimination.

Moreover, discrimination might not only be based on differences in the outcomes for groups, but the choice of data to be used might not be neutral. If the data used for building an algorithm are biased against a group (i.e. systematic differences due to the way the data are collected), the algorithm will replicate the human bias in selecting them and learn to discriminate against this group. This is particularly dangerous if it is assumed that the machine-learning procedure and its results are objective, without taking into account the potential problems in the underlying data. Data can be biased for several reasons, including the subjective choices made when selecting, collecting and preparing data. To give a real-life example: an automated description of images was trained, based on thousands of images that were described by humans. However, humans do not neutrally describe images. Namely, a baby with white skin colour was described as a 'baby', but a baby with a black skin colour was described as a 'black baby'. This is biased data because it assigned additional attributes only to a certain group, while objectively either both cases should be described including the colour or none of them. If such biased information is included in training data that is used for the development of algorithms, it will be used for resultant predictions and is therefore not neutral.11

In addition to the problematic use of data in reinforcing bias against groups, low quality data as such can lead to poor predictions and discrimination. Data can be either poorly selected data or incomplete, incorrect or outdated.¹² Poorly selected data might include 'unrepresentative data', which do not allow generalising about other groups. For example, if an algorithm was created using data on a certain group of job applicants, then the predictions for another group might not hold true.13 The quality of data and potential bias is particularly important in the age of big data as data is generated, often quickly, over the internet without any quality control. In statistics, this is often referred to as 'garbage in, garbage out': even the most well-developed methods for predictions cannot function effectively using low quality data.14 In this sense, data quality checks and the appropriate documentation of data and metadata is essential for high quality data analysis and the use of algorithms for decision making.

⁹ Directive (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

¹⁰ There is increasing literature on discrimination by algorithms. See, for example, Zliobaite I., Clusters B. (2016); Kamiran, F., Žliobaitė, I. and Calders, T. (2013); Sandvig C. et al. (2014).

¹¹ Miltenburg E. (2016).

¹² The White House (2016).

¹³ For example, it was found that speech recognition did not work as well for women as for men. This difference might come from the algorithms being trained using datasets which include more men than women. See an article on this issue.

¹⁴ For example, in a number of the EU's large-scale IT databases that are used in the field of border management and asylum, incorrect or poor quality data has been highlighted as an area that needs to be addressed; see FRA's report on data interoperability: FRA (2017c) and FRA's report on biometrics, EU IT-systems and fundamental rights: FRA (2018).

4. Detecting and avoiding discrimination

Auditing algorithms

Detecting discrimination in algorithms is not an easy task - just as detecting forms of discrimination in general can be difficult. Algorithms used in modern applications of machine learning and artificial intelligence are increasingly complex. As a consequence, results are difficult or almost impossible to interpret. in terms of which information in the data influences predictions and in which way. This is related to the fact that huge amounts of data can be used to predict certain outcomes. For example, frequently-used algorithms are based on so-called 'Neural Networks', which work with hidden layers of relationships and combinations of all different characteristics in the data. This makes it difficult to assess whether or not a person is being discriminated against on grounds of their gender, ethnic origin, religious belief or other grounds. However, if a predictive algorithm is fed with information on different groups and it finds a difference according to this information, it may potentially provide an output that discriminates.

Despite the complexity, algorithms need to be 'audited' to show that they are lawful. In other words, they must be shown to not process data in such a way that leads to discrimination. In this context, the term 'auditing' comes from 'discrimination testing' (or 'situation testing') in real-life situations. For example, in an experiment, two identical, fictional job applications are sent to employers and only the group membership of interest (e.g. ethnic origin) varies. Often simply the name of the applicant is changed by using names typically indicating ethnic origin.¹⁵ In this way an experimental situation is created and differences in call-back rates for a job interview can be directly interpreted as discrimination.

Similar approaches can also be used for detecting discrimination in the use of algorithms. For example, recent advances in research on algorithmic discrimination have suggested different ways to detect discrimination on internet platforms. Algorithms can be audited by obtaining full access to the computer software and code used for the algorithms, which could then be evaluated by a technical expert. However, this might not always be straightforward because the discrimination is not directly encoded in the syntax of the computer software. Other methods could include the creation of profiles on platforms which are randomised as 'testers' and sent repeatedly to see if the outcome differs according to characteristics

that could influence discrimination. However, if such tests are undertaken on real data platforms, this could lead to legal problems if the service in question is harmed – for example, through overloading the server with requests. Hence such tests have to be created in a way to avoid such problems. To One example of such a test was used to detect how well gender classification algorithms work for different groups by using preselected images. The results of the study showed that gender recognition works considerably less well for darker skinned females, compared to white males.

Additionally, there are methods to extract information about which data contribute most to the outcome of the algorithm. This way it can be checked if information on protected grounds, e.g. ethnic origin, is important for the predictions. Information on certain characteristics needs to be extracted from the algorithm to understand differences in the outcomes and results of the algorithm. If, for example, a difference in income explains why a person is not offered a loan, this might be reasonable. However, if group membership makes a difference for a decision, e.g. on a loan, it might be discrimination.

The easiest way to detect discrimination is when full transparency is granted, in the sense that the code and the data used for building the algorithm are accessible to an auditor. However, even in this case it is not always straightforward. Some results might appear discriminatory but a closer look shows that they are not.²⁰ For example, a group may appear to be treated differently on an aggregate level, but breaking up the results into explainable differences shows that no discrimination exists.²¹

To avoid violating fundamental rights, it is crucial that the automated tools used for making decisions on people's lives are transparent. However, the data

¹⁶ Sandvig C. et al. (2014).

¹⁷ Buolamwini and Gebru (2018).

¹⁸ See, for example, a blog on the issue.

¹⁹ Wachter, S., Mittelstadt, B. and Russel C. (2017).

²⁰ Kamiran, F., Žliobaitė, I. and Calders, T. (2013).

²¹ This is related to one of the most well-known cases related to admissions to studies at the University of California Berkeley, where it was shown that, overall, women were much less likely to be admitted than men. However, when splitting up the admission rates by type of studies, it was shown that, for each of the different types of studies, women actually had a slightly higher chance of being admitted for all studies. This is a completely counter-intuitive result, which can occur under certain circumstances, related to the different magnitude of applications per type of studies and differing admission rates. Bickel, P. J. et al. (1975). One explanation is that women tended to apply for more competitive studies as compared to men.

¹⁵ See, for example, Bertrand M. and Mullainathan S. (2003).

used for creating the algorithm are often not available to people outside the company or institution creating the algorithm, due to issues such as copyright and competition between businesses. So only the 'owners' of the data can easily make such tests.22 In this case, either methods to detect discrimination without having access to the dataset need to be developed and allowed (e.g. discrimination testing),²³ or ways of auditing algorithms - through a process that protects business secrets - need to be put in place; for example, through external expert auditors, who can access the data and code, similar to financial auditor checks. In this regard, consideration could be given to establishing public bodies that have the right and power to investigate the use of predictive systems, which are used to make decisions that affect people's lives.24

One example of a study on bias in algorithms was carried out by journalists who investigated if there is racial bias in the risk scores used in the US criminal justice system. They analysed the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) scores, including risk scores for recidivism.²⁵ To evaluate how well the risk scores work,

the journalists obtained a dataset of criminal histories for a period of two years from Broward County in Florida and analysed the actual recidivism in comparison to the risk score. Based on the analysis, it was reported that white defendants were more often mislabelled as 'low risk' compared to black defendants, and the risk score was more likely to falsely flag black defendants as 'high risk'. This potentially indicated racially-biased algorithms. However, the company and other researchers contested the result, by pointing to alternative analyses showing that the model does not make any difference when looking at risk scores for each group separately.26 This discussion demonstrates that it is complex to prove that an algorithm discriminates. Interpretation of results takes place at a comparatively advanced level of statistical analysis, but findings are not fully conclusive, due to an absence of standards for evaluation, and the challenge to create 'fair' algorithms in the presence of different outcomes and their interpretation. The situation is further complicated by the absence of case law in this field.

This last point relates to the problem that it is not fully clear when differential treatment between groups

Application of the legal principle of the right to "meaningful information"

Full transparency on processing might not be possible due to a number of legitimate reasons. For example, these issues may include intellectual property rights or issues of national security. However, controllers have to make sure that an automated decision-making process is explained clearly and simply to individuals, in particular when involving decision-making processes based on profiling.

This materialises in the GDPR through the introduction of a right to explanation, which states that data subjects should be provided with "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject".* The European Parliament has clarified its understanding of the concept, by stating that this should include

"information about the data used for training algorithms and allow individuals to understand and monitor the decisions affecting them".**

However, the concrete implementation of this right remains to be seen and is related mainly to automated decision making. As emphasised by several researchers,*** the delimitation of the scope and feasibility of the explanation to be provided is not evident. Notably, the wording "meaningful information" raises several questions: to whom and when should information be deemed meaningful? Should the information cover the rationale behind the technique or the technique itself? This point would benefit from further clarification – for example, by the European Data Protection Board (EDPB) established by the GDPR.

- * General Data Protection Regulation, Recital 71, Art. 13 (2) (f) and Art. 15 (1) (h).
- ** European Parliament (2017a).
- *** See Wachter, S., Mittelstadt, B., Floridi, L. (2017); Godman B., and Flaxman, S. (2016); Tarran B. (2017); Selbst A. and Powles J. (2017).
- 22 This is often for very legitimate reasons for example, for law enforcement purposes, where the data and algorithms often cannot be publicly available. Moreover, the user of the algorithms might not own the data they use and consequently not be able to share the data (e.g. applications from third parties using Twitter data).
- 23 Sandvig C. et al. (2014).
- 24 Tutt A. (2016).
- 25 The study including accompanying documentation is available on the Propublica webpage. The report is cited as: Angwin J. et al. (2016).

26 Flores A. et al. (2016).

actually constitutes discrimination. In the United States, there was a case on hiring practices, which were ruled as unlawful even if the decision was not explicitly determined based on race. The barriers for accessing a job with a test that was not directly related to the job requirements disproportionately put black applicants at a disadvantage.27 In this case, a certain level of differential treatment was reached that was deemed to constitute discrimination. At the same time, the United States Supreme Court acknowledged that there cannot be a general rule on the level at which differential treatment is seen as discriminatory. Instead, each situation and application needs to be assessed separately.28 A detailed analysis of the way a group is discriminated against should be conducted, taking proportionality into account with respect to the impact that certain criteria have on different groups. For example, a specific job requirement – such as having completed training in a particular country – may not be a genuine or occupational requirement for carrying out the job. If such a requirement puts certain groups at a disadvantage (e.g. immigrants), a potential algorithm that sorts job applications should not use such information for determining eligibility for a job.

Data protection impact assessment (DPIA)

Algorithms may be so complex that the characteristics that will influence the outcome might not be easily identifiable. Article 35 of the GDPR has made data protection impact assessment (DPIA) mandatory for all "processing, in particular using new technologies, [which] is likely to result in a high risk to the rights and freedoms of natural persons". DPIAs are complex procedures that may require a combination of technical, legal and sociological knowledge. Institutional* and academic** actors have developed standards and guidance to help data scientists in the development and implementation of these and standards are being developed by industry associations.*** In this context, guidance on the minimum requirements to be inserted in any DPIA, and established at regional level, would likely enhance the effectivity of DPIAs. Using DPIAs could be a way for controllers to increase their accountability and make sure from the outset that their applications are not discriminatory.

- * See, for instance, the Template for Smart Grid and Smart Metering Systems, developed by the European Commission, or the Guidance on Privacy by design and DPIA, developed by the Information Commissioner's office.
- ** See, for instance, the detailed roadmap developed by S. Spiekermann (2016).
- *** As referred to in the Council of Europe study on human rights dimensions of automated data processing techniques (Council of Europe 2017b).

Avoiding discrimination from the outset and repairing algorithms

Further research and discussion are needed to develop methods to ensure that algorithms are not discriminating, and to rectify algorithms that are found to be discriminatory.

There are several ways to assess whether or not a group is treated differently by an algorithm.²⁹ In the presence of different outcomes for different groups (e.g. one group is more likely to be eligible for a job), there are inherent trade-offs in creating fair predictions on all accounts. It can be mathematically proven that it is not possible to have the same overall risk scores for different groups and a balanced prediction model, unless groups are not performing differently on a certain outcome. For example, women may fare worse on a scale used for hiring people (e.g. using income-related information as a proxy for past performance at work). As algorithms should not discriminate against a particular group, there needs to be an assessment of the implications of using different performance measurements (or their proxies). However, precisely balanced predictions for both genders would not be possible. This limits the ability to make fair predictions for groups who, in reality, fare differently on a certain outcome.30

In addition, situations of potential bias or discrimination cannot be easily solved by simply excluding information on the protected group from the dataset (e.g. just exclude information on gender or ethnic origin). There could be additional information that may be related to membership of a protected group. For example, a given post code could indicate ethnic origin (as mentioned above). There are ways to address and detect indirect information on protected attributes (which can result in indirect discrimination) that can be used to repair algorithms. For example, testing whether protected characteristics of individuals can be predicted well from other information included in the dataset.31 However, research on the topic has only just begun and needs to progress further to develop standardised methods for detecting and

²⁷ U.S. Supreme Court, *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971)

²⁸ Feldmann M. et al. (2015).

²⁹ For example, the overall accuracy (e.g. one group is more likely to be hired, when controlling for other explanatory factors), the true positive rate (the algorithm more often selects people from one group among those who should be selected), and the true negative rate (e.g. the algorithm more often rejects people from one group among those to be rejected) are different ways to assess fairness. For a discussion of measuring bias and different fairness criteria, see Chouldechova (2017).

³⁰ Kleinberg J. et al. (2017a). For a more accessible description, see this blog focusing on quantitative issues.

³¹ Alder P. et al. (2016).

removing discrimination. One fundamental challenge to data-supported analysis is the question of whether information on the protected group is available in the data used to create the predictions.

Data protection by design and by default

The mere exclusion of information on protected groups will not be sufficient to avoid discrimination. Non-discriminatory data processing may be further strengthened through the implementation of "appropriate technical and organisational measures which are designed (...) to integrate the necessary safeguards into the processing in order to (...) protect the rights of data subjects".*

As pointed out by the EDPS and some data protection authorities,** ensuring that data protection is embedded in the early conception phase of any algorithm has several positive consequences: it minimises the risks of potential discriminatory results, and increases trust between the data subject and the data controller. This has led programmers to focus on and develop Privacy Enhancing Technologies (PETs).*** Similar methods could be conceptualised and integrated in the development of data-induced decision making to ensure this is non-discriminatory. In this sense, a 'by design' approach would work in conjunction with a risk assessment that also takes potential discrimination into account.

At the same time, further consideration could be given as to whether 'data protection by design and by default' also leads to a situation where information that could be used to detect potential discriminatory practices is being withheld for data analysis to protecting the data subject. Clarification is needed on how data on sensitive personal characteristics can be employed. For example, such data may be used as aggregate data for statistical purposes, to identify potential discrimination in practices such as selection of persons for interviewing.

- * General Data Protection Regulation, Art. 25.
- ** EDPS (2010), pp. 4-6; ICO, Guide on Data Protection Privacy by design.
- *** ENISA, Privacy Enhancing Technologies: Evolution and State of the Art, 9 March 2017.

(Un-)Availability of information on protected characteristics

Most methods developed so far to identify potential discrimination make use of information on the protected group (e.g. information on ethnic origin) in the dataset. Availability of information on protected groups differs considerably across countries and data collection approaches. For example, information on 'race' or ethnicity is frequently collected in the United States and in the United Kingdom. However, in many EU Member States, there was a move away from data collection on ethnicity, and in some countries, data collection of information on ethnic origin is forbidden. This was introduced in some countries due to the abuse of administrative data that included ethnic origin and/or religion before and during the Second World War.32 This means that, when collecting data on ethnic origin (or any other protected attribute), data protection becomes particularly important. This includes the explicit consent of the data subject as well as consideration of the data collection being in the public interest, or some explicit legal obligation to collect the data.

As a general rule, the GDPR has established that the processing of sensitive data should be prohibited, and has listed the ten exceptional cases when processing of such data may be allowed. These include, but are not limited to, the obtention of the data subject's explicit consent, the protection of their vital interests, or the protection of public interest in the area of public health.³³

In contrast to a number of EU Member States, in the United Kingdom there was a movement in favour of collecting data on ethnicity. This started before the 1990s, for the very purpose of being able to detect discrimination. There are strong arguments for the use of ethnic identifiers in data collection in order to be able to detect discriminatory treatment and outcomes.³⁴ In fact, including this information is not only needed to detect discrimination in algorithms, it is also needed to correct the algorithm.³⁵ The discussion in this area is also related to the interplay between the right to privacy and the right to fair processing of personal data.³⁶

Future research should assess the extent to which proxies (i.e. other information highly correlated with protected characteristics, such as citizenship

³² Simon, Patrick (2007).

³³ General Data Protection Regulation, Art. 9.

³⁴ Chopin I, Farkas L. and Germaine C. (2014).

³⁵ Zliobaite I., Clusters B. (2016).

³⁶ Hoboken J. (2016).

for ethnic origin or height for gender) can be used instead of more direct information on protected characteristics. This discussion extends to other grounds as well. There is an increased amount of information in datasets, including potentially thousands of different characteristics being collected. With this in mind, there should be further assessment of the types of characteristics being profiled by algorithms for the purpose of prediction. Considering the list of protected characteristics in Article 21 of the Charter of Fundamental Rights, a lot of information can easily relate to any of these grounds. For example, this protection could apply to political opinions on Facebook or Twitter posts. It is important to note that the use of algorithms and profiling – for example, when using social media data – can easily predict or relate to special categories, as forbidden under Article 9 in the GDPR. For example, combining 'likes' on social media with other data can be used - quite accurately - to determine a person's sexual orientation, ethnic origin or religion.³⁷

Feedback loops and more efficient decision making

Algorithms used for daily decision making shape people's lives. If they are applied for a longer period of time, they can replicate decision making, because the decisions based on algorithms shape the future data for the algorithm to learn from. In other words, decisions are based on data that exists in a feedback

loop. This might not be problematic if the algorithm is fair and balanced. However, if such models are based on biased data or algorithms, discrimination will be replicated, perpetuated and potentially even reinforced. Discriminatory predictions create the (allegedly neutral) basis for future data and therefore the basis for further development of the algorithms used in the future. It could even lead to continuously increasing discrimination due to 'more efficient' decisions being taken on the basis of data, which puts certain groups at a disadvantage and consequently makes it more difficult for members of disadvantaged groups to obtain fairer opportunities.³⁸ This is exacerbated when automated decisions are created with algorithms.

For example, statistical learning procedures for more efficient law enforcement, which adapt policing (or surveillance) practices based on historical and assumed 'neutral' data, can lead to unequal outcomes even if the underlying levels of criminality are the same among certain sub-populations. This can have the negative effect of criminalising specific sub-populations.39 The problem of using algorithms is that future assessments are only based on results of the algorithms - for example, "hiring algorithms only receive feedback on people who were hired and predictive policing algorithms only observe crime in neighbourhoods they patrol".40 This means that feedback loops of algorithms need to be assessed in detail, to identify whether it is necessary to 'repair' the algorithms to avoid discrimination and, ultimately, to produce more accurate results.

5. Possible ways forward: addressing fundamental rights and big data

The use of data and research for evidence-based policies and decision making has gained popularity in the past decades. This is a trend that also needs to embrace the use of big data and new technologies. Current developments in the area of big data call for action at several levels of society, given the impact of the use of big data across multiple spheres. This focus paper has highlighted potential threats to fundamental rights – focusing on discrimination – when building algorithms based on big data to support decisions. The use of data and more advanced methods for prediction can improve the way we make decisions, as long as the basis for those decisions is fully understood and does not negatively affect fundamental rights. The efficiency of

37 See the examples and explanation given in Article 29 Data Protection Working Party (2017).

algorithms also needs to be assessed against basic fundamental values, such as equal treatment and non-discrimination.⁴¹

Since decision making based on predictive models and other related methods is relatively new, the need for further safeguards in this area needs to be considered. In comparison, the production and selling of food or drugs is highly regulated, due to the dangers related to unregulated use⁴² (acknowledging the different reasons and needs for regulation in

³⁸ The White House (2016); O'Neil C. (2016).

³⁹ Osoba, Osonde and Welser IV William (2017).

⁴⁰ Ensign D. et al. (2018)

⁴¹ See for example the discussion in Kleinberg J. et al. (2017b) or Menon and Williamson (2018).

⁴² This comparison was taken from Tutt A. (2016).

these two areas). However, the use of algorithms in making decisions and building automated processes may have a significant impact on people's lives.

This makes it important to consider the need to review the impact of existing legislation and oversight mechanisms – where these are in place. This would help prevent any possible negative consequences for people resulting from the use of algorithms. Potentially, lessons could be learned from existing fields that are highly regulated, such as medicine.

Consideration could be given to how the oversight of algorithms, which are used for different purposes, might be undertaken in practice. This is reflected in FRA's opinions related to oversight in the area of surveillance, which are attached to the agency's 2017 report on this subject. Effective oversight would require detailed assessment by experts, which would need to be provided for by law in each EU Member State.⁴³ Independent oversight by experts in different subject fields - including technical fields - is crucial for safeguarding the legality of data-based applications. One way of potentially ensuring effective accountability is by setting up dedicated bodies with an exclusive mandate to provide oversight of big data-related technologies, similar to the role of Data Protection Authorities (DPAs). Developing guidance for the use of algorithms calls for strong collaboration between statisticians, lawyers, social scientists, computer scientists, mathematicians and subject area experts, to obtain further clarity on the presence of discrimination and other fundamental rights breaches. In sum, this is a field that warrants further careful analysis.

The GDPR does include safeguards to address nondiscriminatory automated decision making before, during, and after processing data. Data science experts agree that big data-related technologies are currently not being held accountable, and that the need for oversight is urgent and real.44 Transparency is a key element to achieve effective remedies, and this is all the more so when complex algorithmic decision making and/or large datasets are part of the elements legal practitioners will have to review.45 Here again, the GDPR has increased the powers of the data protection authorities in the European Union, making them competent to hear complaints, conduct investigations, and promote awareness. However, in practical terms, further progress is required to ensure that the use of algorithms is in compliance with the law.

When developing and using algorithms for decision making, their lawfulness needs to be assessed. Data

protection principles provide guidance to come to legal decisions concerning the use of algorithms. However, there is more that needs to be considered. A fundamental rights approach encompasses an analysis of legality based on data protection, but at the same time it goes beyond this. As stated in FRA's Opinion on the proposed Regulation on the European Travel Information and Authorisation System (ETIAS), when defining risks, the threat of discrimination through the use of algorithms is not negligible. Without testing algorithms and showing that their application and risk assessments are necessary, proportionate and do not result in discriminatory profiling or decision making, we may not be certain that their application is in compliance with the law.⁴⁶

With this in mind, the following are some examples of how we can move towards fundamental rights compliance in the development and use of algorithms.

- Being as transparent as possible: opening up for scrutiny how algorithms were built supports the further development of these tools and allows others to detect, and therefore rectify, any erroneous applications.
- Conducting fundamental rights **impact assessments**: to identify potential biases and abuses in the application of and output from algorithms. These include, among others, an assessment of the potential for discrimination in relation to different grounds such as gender, age, ethnic origin, religion and sexual or political orientation. At the same time, impact assessments could assess the potential discrimination bias of using 'proxy information' such as addresses with respect to protected grounds in the area of discrimination.
- Checking the quality of data: given the amount of data generated and used, it remains a challenge to assess the quality of all data collected and used for building algorithms. However, it is essential to collect metadata (i.e. information about the data) and make quality assessments of the correctness and generalisability of the data.
- Making sure the way the algorithm was built and operates may be meaningfully explained: this would help facilitate access to remedies for people who challenge data-supported decisions and also relates to the principle of transparency. The challenge of understanding the mathematical background of a statistical method or an algorithm does not prevent a general description of the process and/or rationale behind the calculations feeding the decision making, most notably, which data were used to create the algorithm.

⁴³ FRA (2017a).

⁴⁴ Pew Research Center (2017), pp.77-83.

⁴⁵ See FRA (2017a) and FRA (2017b).

References

Alder P. et al. (2016), Auditing Black-box Models for Indirect Influence, Proceedings of the IEEE International Conference on Data Mining (ICDM).

Angwin J. et al. (2016), Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, ProPublica.

Article 29 Data Protection Working Party (2017), Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, WP251rev.01.

Bertrand M. and Mullainathan S. (2003), Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination, NBER Working Paper Series, Working Paper No. 9873.

Bickel, P. J. et al. (1975), 'Sex Bias in Graduate Admissions: Data from Berkeley', *Science*, Vol. 187, No. 4175, pp. 398-404.

Buolamwini J. and Gebru T. (2018), 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Proceedings of Machine Learning Research*, Vol. 81, pp. 1-15, Conference on Fairness, Accountability, and Transparency.

Callegaro, M. and Yang, Y. (2017), 'The Role of Surveys in the Era of "Big Data"', in Vannette, D.L. and Krosnick, J.A. (eds.), The Palgrave Handbook of Survey Research, pp. 175-192.

Chouldechova, A. (2017), Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, February 2017, arXiv:1703.00056 [stat.AP].

Chopin I., Farkas L. and Germaine C. (2014), Ethnic origin and disability data collection in Europe – Comparing discrimination, Migration Policy Group for Open Society Foundations.

Council of Europe (2017a), Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data, T-PD(2017)01, 23 January 2017.

Council of Europe (2017b), Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, Committee of experts on internet intermediaries, MSI-NET(2016)06 rev6.

European Data Protection Supervisor (EDPS) (2016), EDPS Opinion on coherent enforcement of fundamental rights in the age of big data, Opinion 8/2016.

EDPS (2015), Meeting the challenges of big data, Opinion 7/2015, 19 November 2015, p.7.

EDPS (2010), Opinion of the European Data Protection Supervisor on Promoting Trust in the Information Society by Fostering Data Protection and Privacy, 18 March 2010.

Ensign D. et al. (2018), 'Runaway Feedback Loops in Predictive Policing', *Proceedings of Machine Learning Research*, Vol. 81, pp. 1-12, Conference on Fairness, Accountability, and Transparency.

European Parliament (2017a), Fundamental rights implications of big data, P8_TA-PROV(2017)0076.

European Parliament (2017b), Civil Law Rules on Robotics, P8_TA(2017)0051.

Feldmann M. et al. (2015), 'Certifying and removing disparate impact', In *Proceeding KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Flores A. et al. (2016), 'False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks', Federal Probation, Vol. 80 (2), pp. 38-46.

FRA (European Union Agency for Fundamental Rights) (2010), *Towards More Effective Policing, Understanding and preventing discriminatory ethnic profiling: A guide*, Luxembourg, Publications Office of the European Union (Publications Office).

FRA (2011), Handbook on European non-discrimination law, Luxembourg, Publications Office.

FRA (2017a), Surveillance by intelligence services: fundamental rights safeguards and remedies in the EU, Luxembourg, Publications Office.

FRA (2017b), Fundamental rights and the interoperability of EU information systems: borders and security, Luxembourg, Publications Office.

FRA (2017c), The impact on fundamental rights of the proposed Regulation on the European Travel Information and Authorisation System (ETIAS), FRA Opinion 2/2017, Vienna, FRA.

FRA (2018), *Under watchful eyes: biometrics, EU IT systems and fundamental rights*, Luxembourg, Publications Office.

Goodman B., and Flaxman, S. (2016), EU Regulations on Algorithmic Decision-Making and A "right to Explanation".

van Hoboken, J. (2016), 'From Collection to Use in Privacy Regulation? A Forward Looking Comparison of European and U.S. Frameworks for Personal Data Processing', In Van Der Sloot, Broeders and Schrijvers (eds.), *Exploring the Boundaries of Big Data*, Netherlands Scientific Council for Government Policy, pp. 231-259.

Information Accountability Foundation (the) (2016), Effective Data Protection Governance Project: Improving Operational Efficiency and Regulatory Certainty in a Digital Age, July 2016

Information Commissioner's Office (2017), *Big Data, artificial intelligence, machine learning and data protection*, 1 March 2017.

Kamiran, F., Žliobaitė, I. and Calders, T. (2013), 'Quantifying explainable discrimination and removing illegal discrimination in automated decision making', Knowledge and Information Systems, Vol. 35 (3).

Kleinberg J. et al. (2017a, forthcoming), 'Inherent Trade-Offs in the Fair Determination of Risk Scores', Proceedings of Innovations in Theoretical Computer Science (ITCS).

Kleinberg J. et al. (2017b), *Human Decisions and Machine Predictions*, NBER Working Paper No. 23180.

Menon, Aditya K. and Williamson Robert (2018), 'The Cost of Fairness in Binary Classification', *Proceedings of Machine Learning Research*, Vol. 81, pp. 1-12. Conference on Fairness, Accountability, and Transparency, 28 March 2018.

Miltenburg E. (2016), Stereotyping and Bias in the Flickr3oK Dataset. Proceedings of the Workshop on Multimodal Corpora: Computer vision and language processing (MMC-2016), 24 May 2016, Portoroz, Slovenia.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter S. and Floridi L. (2016), 'The Ethics of Algorithms: Mapping the Debate', *Big Data & Society*, Vol. 3 (2).

O'Neil, C., (2016), Weapons of Math Destruction, New York, Crown. Osoba, Osonde and Welser IV William (2017), An Intelligence in Our Image. The Risks of Bias and Errors in Artificial Intelligence, Santa Monica, RAND Corporation.

Pew Research Center (2017), Code-Dependant: Pros and Cons of the Algorithm Age.

Simon, P. (2007), "Ethnic" statistics and data protection in the Council of Europe countries, Strasbourg, Council of Europe.

Sandvig C. 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", 22 May, Seattle, WA, USA.

Seitz, C. (2017), 'Big data in the pharmaceutical sector', in Vermeulen, G. and Lievens, E. (eds), *Data Protection and Privacy under Pressure*.

Selbst A. and Powles J. (2017), 'Meaningful information and the right to explanation', *International Data Privacy Law*, Vol. 7 (4), pp. 233-242.

Spiekermann, S. (2016), Ethical IT Innovation, A value-based system design approach, CRC Press.

Tarran B. (2017), 'A "right to explanation" for Al decisions?', Significance magazine, April 2017, p. 9.

The White House (2016), *Big Data: A report on Algorithmic Systems, Opportunity and Civil Rights.*

Tutt A. (2016), 'An FDA for Algorithms', Administrative Law Review, Vol. 67.

Wachter, S., Mittelstadt, B. and Floridi, L. (2017), 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*.

Wachter, S., Mittelstadt, B., and Russel C. (2017), Counterfactual explanations without opening the black box: automated decisions and the GDPR, last revised 25 Dec 2017.

Zliobaite I. and Clusters B. (2016), "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", Artificial Intelligence and Law, Vol. 24 (2).

Further information:

The following FRA publications offer further information relevant to the topic of the paper:

- Handbook on European data protection law 2018 edition (2018), http://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law
- Under watchful eyes: biometrics, EU IT systems and fundamental rights (2018), http://fra.europa.eu/en/publication/2018/biometrics-rights-protection
- Fundamental rights and the interoperability of EU information systems: borders and security (2017), http://fra.europa.eu/en/publication/2017/fundamental-rights-interoperability
- Surveillance by intelligence services: fundamental rights safeguards and remedies in the EU Volume II: field perspectives and legal update (2017), http://fra.europa.eu/en/publication/2017/surveillance-intelligence-socio-lega
- Surveillance by intelligence services: fundamental rights safequards and remedies in the European Union -Mapping Member States' legal frameworks (2015), http://fra.europa.eu/en/publication/2015/surveillance-intelligence-services
- The impact on fundamental rights of the proposed Regulation on the European Travel Information and Authorisation System (ETIAS) (2017), http://fra.europa.eu/en/opinion/2017/etias-impact
- Interoperability and fundamental rights implications (2018), http://fra.europa.eu/en/opinion/2018/interoperability
- Handbook on European non-discrimination law 2018 edition (2018), http://fra.europa.eu/en/publication/2018/handbook-european-law-non-discrimination

FRA - EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS

Schwarzenbergplatz 11 - 1040 Vienna - Austria Tel: +43 158030-0 - Fax: +43 158030-699 fra.europa.eu – info@fra.europa.eu facebook.com/fundamentalrights linkedin.com/company/eu-fundamental-rights-agency twitter.com/EURightsAgency





© European Union Agency for Fundamental Rights, 2018