EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS

HELPING TO MAKE FUNDAMENTAL RIGHTS
A REALITY FOR EVERYONE IN THE EUROPEAN UNION

Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights

FRA Focus

Algorithms used in machine learning systems and artificial intelligence (AI) can only be as good as the data used for their development. High quality data are essential for high quality algorithms. Yet, the call for high quality data in discussions around AI often remains without any further specifications and guidance as to what this actually means. Since there are several sources of error in all data collections, users of AI-related technology need to know where the data come from and the potential shortcomings of the data. AI systems based on incomplete or biased data can lead to inaccurate outcomes that infringe on people's fundamental rights, including discrimination. Being transparent about which data are used in AI systems helps to prevent possible rights violations. This is especially important in times of big data, where the volume of data is sometimes valued over quality.

Contents

1. Data quality, artificial intelligence and fundamental rights

Artificial intelligence (AI) and big data continues to be a topic of high priority for policy, science, business and media throughout the world. Developments in the area are of high relevance, as new technologies impact on all spheres of life and hence also impact on fundamental rights. Ethical implications of AI are the topic of many discussions. At the same time, these discussions need to acknowledge that there is a human rights framework setting binding legal obligations around AI, which should be seen as a starting point for any evaluation of the opportunities and challenges brought by new technologies. The European Union's strong fundamental rights framework, as enshrined in the Charter of Fundamental Rights and related case law, provides guidance for the development of guidelines and recommendations for the use of AI.

This paper sets out to contribute to the many ongoing policy discussions around AI and big data by highlighting one aspect that needs attention from a fundamental rights perspective; namely the awareness and avoidance of poor data quality. It does not aim at explaining how to use high quality data, but how to become aware of and avoid using low quality data.

Data quality for building algorithms and AI-related technologies is one of the concerns for the fundamental rights compliant use of data. This is because an algorithm in its application can only be as good as the data it uses. Following the often quoted 'garbage in – garbage out' principle, low quality data lead to low quality outcomes produced by algorithms which in turn can lead to a violation of fundamental rights. Most obviously, privacy and data protection is an important area that the use of low quality data can affect. Other rights are also affected; for example, an automated system used in the justice system, which is based on poor quality data, can negatively impact on the right to a fair trial and effective remedy, as well as on the principle of good administration. As referred to in FRA's previous focus paper on AI, algorithms that are biased can, as a result, lead to discrimination against women, ethnic minorities, the elderly, and other groups based on protected grounds.

For example, if a voice recognition system is mainly trained on male voices, the system may not perform accurately when used by women – due to a data quality problem. The quality of the data used leads to unequal performance of the services for different groups in the population. This has been identified as one of the potential problems that can result in discrimination in data supported decision making, which FRA explored in its earlier focus paper.[1] In addition, data can reflect existing bias and discriminatory behaviour, which is then taken up and potentially reinforced by

What is artificial intelligence (AI)?

There are many definitions of AI. One definition of AI is included in the European Commission communication on Artificial Intelligence for Europe: "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications). We are using AI on a daily basis, e.g. to translate languages, generate subtitles in videos or to block email spam. Many AI technologies require data to improve their performance. Once they perform well, they can help improve and automate decision making in the same domain. For example, an AI system will be trained and then used to spot cyber-attacks on the basis of data from the concerned network or system."*

It is important to accept that the term AI does not refer to 'one thing' but to current technological developments and processes in general. Most of what is discussed under the heading of AI refers to the increased automation of tasks through the use of machine learning and automated decision making. At the core of current AI discussions and machine learning applications lies the use of algorithms. Algorithms are rules followed by a computer, as programmed by humans, which translate input data into outputs.**

* *European Comission (2018).*
** *See further on a definition of algorithms: FRA (2018a). A useful discussion of the definition of artificial intelligence can be found in United Nations (UN), Human Rights Councl (2018).*

[1] FRA (2018a); Barocas, S. and Selbst, A. D. (2016).

an AI-system. Data quality is a broad concept. Two generic concepts related to data quality, as used in social sciences and survey research, can be highlighted:

- errors of representation, which means that the data do not cover well the population it should cover;
- measurement errors, which means that the data do not measure what they intend to measure.

Data quality: a key concern in policy discussions on AI

Discussions and research on the assessment of algorithms often focus on the need to explain how complex algorithms work. However, when assessing algorithms, a focus on the type and quality of data used by algorithms is of equal importance and should be included in any assessment of algorithms. Recently, academic research on data quality in AI and machine learning has received increased attention.[2] However, many text books and articles dealing with data science and machine learning still overlook the crucial aspect of data quality or only scratch the surface of this topic.[3] This paper contributes to the discussion of fundamental rights implications of AI by highlighting concepts of data quality in the development of algorithms and AI-related technologies, as used in social sciences and survey research. It focuses on one potential source of impact on fundamental rights, and serves to provide guidance for ongoing policy discussions on the use of AI and its relation to fundamental rights.

The topic of data quality is mentioned repeatedly in policy documents in relation to the use of AI. The European Council Conclusions of 28 June 2018 acknowledged that "high-quality data are essential for the development of Artificial Intelligence."[4] The European Group on Ethics in Science and New Technologies (EGE) refers to discriminatory biases in datasets used to train and run AI systems, which "should be prevented or detected, reported and neutralised at the earliest stage possible".[5]

Furthermore, the Committee on Civil Liberties, Justice and Home Affairs of the European Parliament (LIBE Committee) in its Opinion on the motion for a European Parliament resolution on a comprehensive European industrial policy on artificial intelligence and robotics[6] highlighted the importance of the "quality and accuracy, as well as representative nature of data used in the development and deployment of algorithms". More specifically, the LIBE Committee noted that "the use of low-quality, outdated, incomplete or incorrect data at different stages of data processing may lead to poor predictions and assessments and in turn to bias, which can eventually result in infringements of the fundamental rights of individuals or purely incorrect conclusions or false outcomes".[7]

Moreover, the European Commission's High Level Expert Group on AI published its ethics guidelines on AI. These include data governance as one of the requirements of trustworthy AI. They highlight the importance of biases in datasets used for machine learning, which could include human misjudgement, errors and mistakes, and the need to keep record of data that are fed into AI systems.[8]

The "Toronto Declaration", initiated and signed by several rights groups, technologists and researchers, issues statements for avoiding bias and discrimination in machine learning systems. One such statement calls on the private sector to take account of risks commonly associated with machine learning systems, including "incomplete or unrepresentative data, or datasets representing historic or systemic bias".[9]

The European Commission for the Efficiency of Justice (CEPEJ) of the Council of Europe has adopted the European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment. This charter puts forward principles including the principle of quality and security, which includes the use of certified sources and intangible data.[10]

2 See for example: Gebru, T. et al (2018); Holland, S. et al. (2018); Richardson, R., Schultz, J. and Crawford, K. (2019).

3 A few text books on the use of big data, machine learning and AI discuss data quality issues in more detail. See for example: Salganik, M. J. (2017); Cabitza, F. et al. (2018); Foster, I. et al. (2016).

4 European Council (2018).

5 European Group on Ethics in Science and New Technologies (2018).

6 European Parliament (2019).

7 Committee on Civil Liberties, Justice and Home Affairs (2018).

8 European Commission, High-Level Expert Group on Artificial Intelligence (2018).

9 RightsConn Canada (2018).

10 Council of Europe (2018).

2. How artificial intelligence and machine learning algorithms use data

One of the main drivers behind technological developments in the area of AI is the unprecedented availability of data.[11] Vast amounts of data are being collected, analysed and used, at an ever increasing pace – a phenomenon referred to as Big Data. Often, but not exclusively, data are gathered over the internet and smartphones. Such data are considered an important asset that provides the basis for many AI applications and progress in the field. In this sense, questions and concerns related to data quality should be at the core of discussions on AI developments. Data constitute the basis for many technological developments in the area.

Machine learning is one broad field of AI, where most discussions focus on learning algorithms that make use of data for establishing patterns that can be applied to new, unseen data. An algorithm learns its rules based on the examples included in the training data. Most discussions, as the one in this paper, focus on so-called supervised machine learning which uses labelled data.[12]

Major advances have been made for image recognition, where an algorithm is built by learning to classify pictures. For instance, many pictures including houses are analysed to create an algorithm for identifying houses. The rules based on the analysis are then tested against a new set of pictures, which also include houses, for accuracy. Other examples include: data from previous crime incidents that help to understand where and when it is most likely that crime occurs, or data from social media posts (for example the use of specific words or combinations of words) are used to predict if someone is likely to click on an advertisement. Data from previous health examinations and their outcomes can help to understand and predict which patients have a certain illness or are at highest risk to suffer from fatal diseases. Data on browsing history could be used to predict the income of people.

These examples make use of so-called labelled data, which already include information on the desired outcome. This information is then used to learn for other situations where it is not yet available. As data labels (e.g. a description of an image) are often created by humans, they are prone to human bias and error.

It is important to note that most learning algorithms cannot (easily) go beyond the data they use to learn patterns from – the so-called training data. While computer scientists make all possible efforts to learn from the data at hand, once a machine learning model is deployed and used in real life, it is not always (easily) possible to verify its success or potential damage.

When using machine learning algorithms, there are at least three different data sets involved:[13]

- The **training data** that are used to build the algorithm. This could be data on internet users, their browsing history, and whether they click on certain advertisements. For example, the demographic characteristics and employment history of unemployed people, and information about whether or not they found a job after some time, could be used to predict when an unemployed will find a job. For supervised machine learning the desired outcome needs to be included in the training data. Data used to learn about the desired outcome are so-called **features**. The desired outcome is often referred to as **labels**. This is the basis of how an algorithm learns patterns.

- When an algorithm is deployed, it is fed with new, unseen features **(input data)**. These are evaluated against the model parameters for taking actions or making decisions. For example, the browsing history of people without yet knowing if they would click on a certain ad.

[11] Data can be loosely defined as more or less structured and standardised information that can be processed by computers – usually in the form of text and numbers. For example, when (raster-) images are analysed, their content can be translated into numbers representing the position and colour of pixels. This paper raises issues that mainly arise with data about people, which is only one subset of all data.

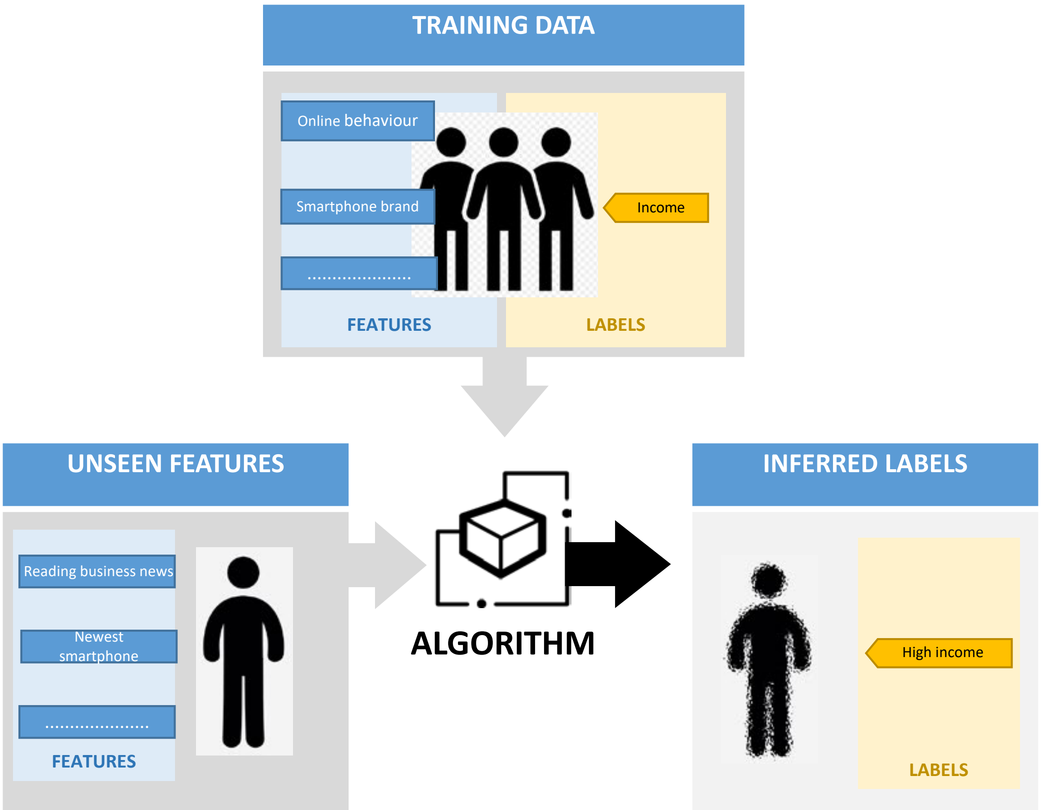
[12] The discussion would similarly apply to unsupervised learning, where no labelled data are available and categorisations of data are being developed. The ways in which algorithms are trained, through splitting training data into training, test and validation sets is relatively complicated and not part of the discussion of the paper. In addition, the use of reinforcement learning, where a component of trial and error is included in the training phase, is not considered in this paper. An accessible description of machine learning can be found in The Future of Privacy Forum (2018) or Alpaydin, E. (2015).

[13] This is a simplified description of one type of machine learning algorithms, which should help to get a better understanding of the type of data involved in often-used cases of machine learning algorithms.

- The algorithm then produces **inferred labels** (or predictions, inferences, deduced actions or output data) that are produced when the unseen data are fed into the machine learning algorithm.

Figure 1 visualises this process. This is a much simplified description of the process which serves to illustrate how data are used in supervised machine learning. In the example shown, the training data contain information on the internet behaviour (for example the browsing history) and data on possible possessions of the user (for example which smart phone a person bought). The training data also includes labels, which could be information on the known income of people (for example obtained through a survey among internet users). The algorithm finds correlations between the features and labels and derives rules. These rules are used for new, unseen features, for example knowing that a person reads certain information online (e.g. business news) and bought the newest smart phone. Feeding this information into the algorithm produces a label, which is inferred from the features; in the example below, high income.

Figure 1: Simplified illustration of types of data used for algorithms



Source: FRA, 2019

An algorithm following the process described above will learn its rules and patterns based on the training data. If the training data are of low quality, most notably structurally different to the new unseen features, the outputs of the algorithms will produce poor results.[14] These could include the reproduction of bias in a dataset and the amplification of existing bias, which can lead to discrimination. Low quality of data can refer to many different aspects, leading to the problem that the algorithm while using new data. If the training data cover a different group of people compared to the new data, where the correlations between the features and labels are different, the results (i.e. the inferred labels) will be more often erroneous. Herein, there are many aspects to data quality. This paper highlights two important errors which are derived from survey research: errors of representation (i.e. different population groups are covered in training data and the new data) and measurement errors (the training data do not include the right information).

[14] Sessions, V. and Valtorta, M. (2016).

3. Beware of the bias – case study on the use of data from the internet

The internet is one important – though not the only – source for data generation and collection that AI draws on. Data can come from a variety of sources, cover different types of data and various topics. While data from different sources can impact on bias, this section focuses only on data from the internet as a generic case study to illustrate the potential for errors of representation.

Many businesses try to harness information from the internet, as such data are often freely available. There is not much information available about which data are exactly used for which applications; however, the internet and social media is one of the frequently used sources. The following describes comparative data on the use of internet data by companies and highlights the bias in internet data at a general level in the EU. This should give a general sense of coverage issues of data from the internet and its potential bias. While increasing coverage, data from the internet may only reflect a subset of the entire population, which is related to limited access to the internet and different levels of participation in online services, such as social media. Often data are collected through items, referred to as the Internet of Things (IoT). Although many AI systems do not use data from the internet, several AI applications use data from the internet. This could include insurance companies using data from social media to create risk scores of potential customers[15] or the development of facial recognition algorithms based on images from the internet.[16]

At the beginning of 2018, more than one in ten businesses with more than 10 employees in the EU indicates to use big data analytics. The percentage of businesses using big data analytics is the highest in Malta (24 %) and the Netherlands (22 %), followed by Belgium, Ireland and Finland (19-20 %). Big data analytics might be more important and more easily applied by larger companies due to potentially more resources and more data at hand. Overall, one in three larger enterprises in the EU (33 %) uses big data analytics – particularly large enterprises in Belgium (55 %), the Netherlands (53 %), Malta (48 %), Ireland (47 %), Denmark (46 %) and Finland (44 %).[17]

Among those enterprises using big data, the most important source is geolocation data of portable devices, which is mostly information on where people are and how they move measured through information from smartphones. Every second, enterprise of those using big data make use of such data (49 %). Similarly, 45 % of big data using enterprises use social media data. Other data sources include enterprises' smart devices or sensors, which are used by 33 % of all businesses that use big data analytics.[18]

These data exemplify that smartphone and social media data are important sources for big data analytics, which can potentially be used for the development of machine learning algorithms and business decisions. For example, in the area of insurance, these unconventional sources of types of data are increasingly used.[19]

The use of data from the internet raises many questions in relation to who is included in the data and to what extent the information included is fit for purpose.

First of all, not everyone has access to the internet or social media, and coverage of different applications varies as well. Secondly, not everyone wants to access the internet or – in particular – social media or other applications. As a result, certain groups are not covered by data gathered over the internet. In the same way, location data is only representative of those who make this information available for use, for example on their portable devices such as smartphones (depending on how individuals manage their location settings). The enormous growth in the use of the internet almost lets one forget how many people do not have access to the internet, and that data is often biased as it only represents a particular group in the population. This can lead to invalid

15 See for a discussion: Internet Governance Forum (2018).

16 Wang J. et al. (2014). For instance Google provides a list of datasets that can be used for AI development, including many internet generated data: https://ai.google/tools/datasets/.

17 Larger enterprises include those with 250 persons employed or more. Statistics on all enterprises include those businesses with at least 10 persons employed (not considering the financial sector). Data based on Eurostat (2019a).

18 Eurostat (2019a).

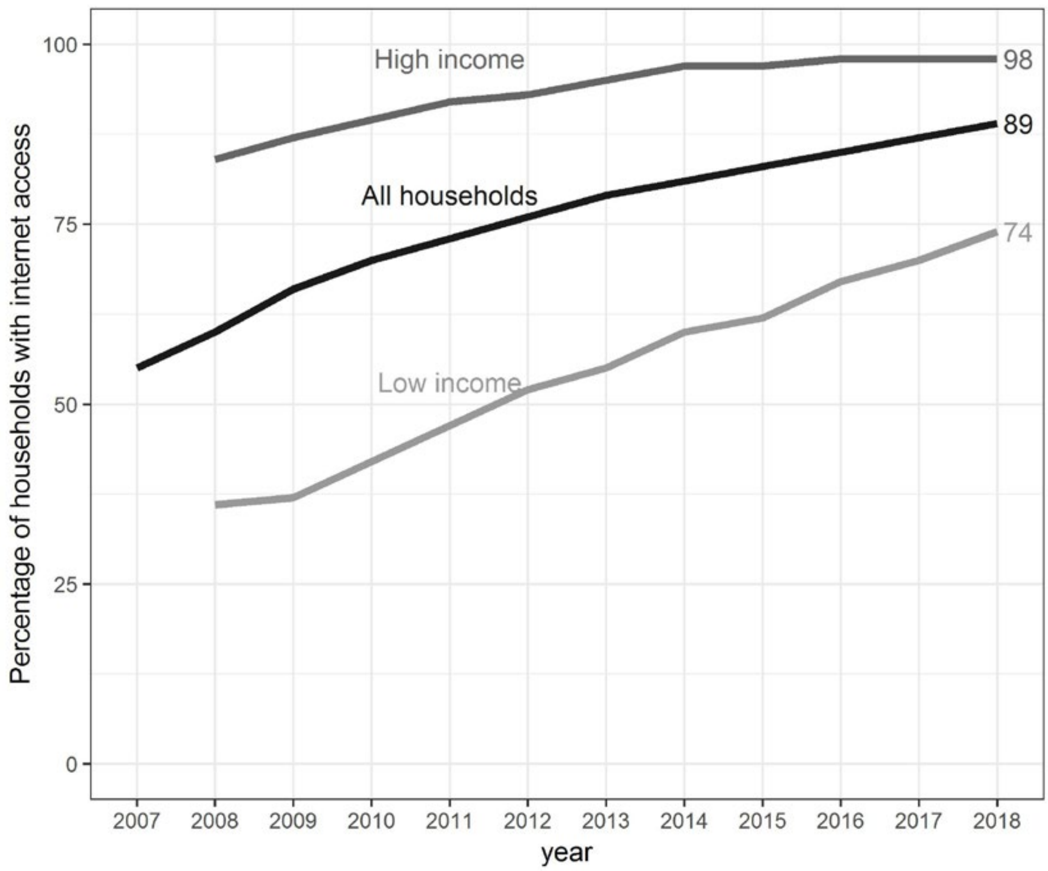
19 For example, the Life Bureau of the New York State Department of Financial Services has issued a circular to all insurers authorised to write life insurance in New York, advising on the use of external, 'non-conventional' data sources with reference to social media and internet data. The circular highlights the potential of discrimination and the challenges of potential issues with the reliability and accuracy of such data. See: Life Bureau of the New York State (2019).

applications if applied to other groups who were not included in the training data for the AI application. The biased coverage of internet data in comparison to the total population can be exemplified through official statistics from Eurostat (Figure 2).

While there has been a strong increase in the proportion of households in the EU with internet access, at the beginning of 2018, only 89 % of all households in the EU-28 had internet access in their household. This means that more than one in ten households did not have a connection to the internet in their households at the beginning of 2018. However, general access to the internet varies considerably across countries, regions within countries and different groups of the population.

Figure 2 exemplifies what is often referred to as the digital divide by showing the different proportion of households with an internet connection for higher income households compared to lower income households. Among the richest quarter of households (the fourth quartile) almost all had internet access at 98 %. Contrary to that, only 74 % of the poorest quarter of households (the first quartile) had internet access. It can be observed, however, that the gap has been closing over time. While the proportion of richer households with internet access was more than twice the proportion of poorer households in 2008 (130 % larger), the proportion of richer households was 32 % larger in 2018.

Figure 2: Households with internet access in the EU-28, between 2007 and 2018



Source: FRA, 2019 [based on Eurostat (isoc_ci_in_h)]

Additionally, there is a strong geographical disparity in terms of levels of access to the internet. The percentage of households with access to the internet ranges in the EU from a low of 72 % in Bulgaria up to 98 % in the Netherlands.[20] The level of access to the internet is regionally clustered with declining access from North to South and from West to East.

[20] Eurostat (2019).

When looking at all individuals in the EU-28 (not households), as many as 11 % said, at the beginning of 2018, that they never use the internet. The percentage of those never using the internet differs across occupational and age groups, education levels and gender. Internet use is very widespread among students, non-manual workers, ICT professionals, women and men with high education and among younger people. High proportions of individuals never using the internet are found among older people – 27 % of those aged 55 to 74. Finally, the proportion of individuals never using the internet is very high among people with low formal education, at 24 % among men and even higher among women at 30 %. This points to a strong disadvantage among women with low education in terms of using the internet – a gender gap that is not so pronounced among other levels of education.

Consequently, data generated over the internet are necessarily unrepresentative with respect to certain groups in the population. This includes southern and eastern countries in the EU, poorer families, older people and people with low education – particularly women.

In addition, the use of social media among the same groups is even lower than general internet use and is also a source of biased data. The apparently easy access to smartphones and social media is distracting from the fact that many individuals do not (want to) use social media. In the EU-28, only slightly more than half of all individuals (56 %) use the internet to participate in social networks, such as creating user profiles, posting messages or other content on Facebook, Twitter, etc. Again, there are stark differences across EU Member States and population groups using social media. Social media use is higher among students (88 %), younger people (aged 16 to 24 years, 88 %), ICT professionals (76 %) and women and men with high formal education (69 % and 62 %).

These statistics underpin that data from the internet and social media are limited in terms of coverage of the population. These shortcomings potentially limit the use of data from the internet for developing machine learning models that are applied to the general population and for specific groups. However, the potential harm of using biased data for AI systems or algorithmic decision making depends on the purpose of an application.

4. Low data quality's impact on fundamental rights

This section provides a brief overview of some fundamental rights that may be affected by using low quality data in AI-related technologies and algorithms. FRA is currently carrying out a project that examines concrete 'use cases' of AI-related technologies from a fundamental rights perspective (more information can be found in the box), which will develop this point further.

The most obvious and well-studied impact is on the right to **non-discrimination** (Article 21 of the Charter of Fundamental Rights of the EU). Several reports and studies have highlighted that the use of either non-representative or biased data can lead to unequal treatment of people based on characteristics such as sex, age, disability, sexual orientation, ethnic origin, and religion, which are among grounds protected by law.[21] If there are structural differences in the training data for protected attributes, such as gender, ethnic origin, or political opinion, the output data of machine learning algorithms can discriminate against individuals based on these attributes. Examples of discrimination as a result of AI using inadequate data are growing. A hiring algorithm was found to generally prefer men over women.[22] An online chatbot became racist within a couple of hours.[23] Machine translations showed gender bias[24] and face recognition systems worked well for white men, but not for black women.[25] Sentiment analysis, a method where a piece of text is assigned a score of how positive or negative it is, provides sexist and racist results by reproducing human-like biases.[26] The reasons for these outcomes are mainly based on the data that were used to train the machine learning systems.

Discrimination can also affect the enjoyment of **economic and social rights**, as it relates to access to services. For example, algorithms and automation

21 FRA (2018a); Barocas S. and Selbst A. D. (2016).

22 Dastin, J. (2018).

23 Johnston, I. (2017).

24 Prates, M., Avelar, P. and Lamb, L. (2018).

25 Buolamwini, J. and Gebru, T. (2018).

26 Caliskan, A., Bryson, J. J. and Narayanan, A. (2017).

are increasingly used in areas related to access to employment, social services and welfare, with a primary purpose to make allocation of resources more efficient. In this field, the use of algorithms can have a negative impact especially on poor people, as research from the United States shows.[27] Moreover, if health related data exclude certain groups, the application of newly developed treatments might not work for some groups.

Equality between men and women (Article 23 of the Charter) is another related area impacted on by low data quality. If one gender is under-represented or if sexist behaviour is represented in the training data[28], AI can increase inequality between men and women.[29]

Access to a fair trial and effective remedies (Article 47 of the Charter) can also be impacted, particularly if algorithms are used in the area of crime prevention and the criminal justice system. The use of AI and algorithms in the area of justice needs testing, as highlighted in the CEPEJ European ethical charter.[30] One potential problem might be the use of biased data for automated systems.[31] In addition, to enjoy access to a fair trial and effective remedies, in cases where someone claims to have been mistreated by an AI-system or wants to challenge a decision based on an algorithm, information on how the system or algorithm works is essential. For example, if a person is denied access to a service because of the result of an algorithm (be it automated or not), the person has the right to challenge the decision. To be able to do so, information on how the algorithm works is needed. As the assessment of the quality of training data is one important part in the assessment of the quality of algorithms, even though not all of it, accessible information on the quality of the data is required to access effective remedies.[32]

If algorithms are used by public administration, potential problems can arise through low quality data and, hence, impact on the principle of **good administration** as established in EU law. Article 41 of the Charter mirrors this principle as applying to EU institutions and other bodies.[33] Accordingly, every person has the right to have his or her affairs handled impartially, fairly and within reasonable time by the institutions and bodies of the EU. The use of automated means can contribute to that. However, Article 41 (2) also mentions that this right includes, among others, the obligation of the administration to give reasons for its decisions. This can also be interpreted as the need to make transparent what data sources were used to train algorithms and automated systems.

Finally, developments in the area of AI strongly impact on issues related to the respect for **private and family life** (Article 7) and the **protection of personal data** (Article 8). However, the application of data protection law to the question of data quality for building AI-related technologies and algorithms is not clear. Data protection legislation offers minimal guidance on the topic – the principle of data accuracy in the General Data Protection Regulation (GDPR)[34] is related to data quality, but in a very narrow sense as it only focuses on the obligation to keep personal data accurate and up to date. Accuracy is usually interpreted as correctness of personal data for one individual (e.g. is the age of one person in a database correct), although the term accuracy could be interpreted more widely. Additionally, the GDPR covers the important rights of access, rectification and erasure, data minimisation and security of data.

However, data quality means much more when being used for AI-related technologies and is not limited to personal data. The training data might be used in an anonymised way and data subjects might consent to their data (features) being used by systems for new predictions. There is ongoing discussion if the new, inferred labels (see Figure 1) are actually covered by data protection legislation or not. Researchers have started analysing this topic by highlighting the importance to consider inferred data as personal data. Only then the rights in the GDPR could apply, including the right to know about those data, and access, rectify, delete and object to them.[35] More research on this topic is needed.

27 Eubanks, V. (2018).

28 Buolamwini, J. and Gebru, T.(2018).

29 Prates, M., Avelar, P. and Lamb. L. (2018); Crawford K. (2016). For algorithms reproducing gender biases, see Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K. (2017); Bolukbasi, T., Chang, K., Zou, J., Saligrama, V. and Kalai, A. (2016).

30 Council of Europe (2018).

31 Richardson, R., Schultz, J. and Crawford, K. (2019).

32 Wagner, B. (2018). For further details on negative fundamental rights implications, consult Raso, F., Hilligoss, H., Krishnaumurthy, V., Bavitz, Ch. And Kim, L. (2018); Council of Europe (2017); AccessNow (2018).

33 While the Charter article applies to EU institutions and bodies only, established EU law is much broader and also includes administrations of EU Member States.

34 GDPR, Article 5(1)(d); see also FRA (2018b), pp.127-128. Additional EU law relevant on this aspect is the Police Directive if law enforcement authorities use or rely on AI, and the EU's own Data Protection Reg. (2018/1725) when it comes to the use of data by EU institutions.

35 Wachter, Sandra and Mittelstadt, Brent (2018).

The use of non-personal or anonymised data needs to be assessed as well, if used for services that provide decisions about individuals. For example, decisions on access to services can be based on statistical predictions based on group characteristics, such as post code, occupation and other traits that are duly anonymised.

Apart from the data accuracy aspect, the need for an assessment of the quality of data used for real world applications derives directly from the data protection principles of accountability and transparency. However, transparency might be difficult when it comes to questions of copyright, intellectual property and business secrets.[36]

For automated decision-making, the GDPR requires data controllers to provide meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.[37] The quality of the data used for developing algorithms can be seen as one aspect of being able to understand the logic, significance and consequences of (automated) decision making systems based on algorithms. This makes a description and assessment of data used to train an algorithm an essential component of the provision of meaningful information about algorithms.

FRA project on Artificial Intelligence, Big Data and Fundamental Rights

In 2018, FRA launched a research project on Artificial Intelligence, Big Data and Fundamental Rights. This project aims at assessing the positive and negative fundamental rights implications of new technologies, including AI and Big Data. It analyses concrete case studies through carrying out interviews with public administration and businesses in selected Member States, and explores the fundamental rigths impact of AI in selected areas, such as health or insurance. Additionally, the project collects information on awareness of fundamental rights issues among public administration and businesses applying AI-related technologies. Finally, the project aims to explore the feasibility of studying concrete examples of fundamental rights challenges when using algorithms for decision-making through either online experiments or simulation studies.

For more information on the project, see FRA (2018), Artificial Intelligence, Big Data and Fundamental Rights.

5. Assessing data quality

Many criteria can be looked into when assessing the quality of data for AI applications. In general, data quality includes many different issues; for example questions of completeness, accuracy, consistency, timeliness, duplication, validity, availability and provenance.[38]

FRA's research on the use of biometric data in large-scale EU migration databases has highlighted that the quality of data, most notably the accuracy of alphanumerical data and biometric identifiers, can negatively impact on the protection of personal data. In the EU's large scale databases, such as the Visa Information System (VIS) or in the Schengen Information System (SIS II), incorrect data about individuals is rather commonly reported.[39] The trust in new technologies and large databases often lets people forget that the data might be inaccurate. While the discussion is mainly concerned with the accuracy of data in relation to individuals, data quality takes a much broader scope in the era of big data and artificial intelligence. For big data and AI applications, aggregated data are used to learn about patterns in society, to automate processes, and for decision making.[40]

One of the problems of big data is that the sheer size of the data has the tendency to convince us that the findings based on such large-scale data must be accurate. However, if the data quality is not taken into account, this assumption might not hold. Data quantity is only one criteria for the accuracy of measuring or predicting something. To tackle statistical accuracy based on data and determine how well we can represent the real world, it needs to be assessed alongside *data quality*.[41]

36 Levendovski, A. (2018).

37 GDPR, Article 15 (h), but also Article 13 (f) and Article 14 (g).

38 Burt A., et al. (2018).

39 FRA (2018c).

40 Within these systems only the European Travel Information and Authorisation System (ETIAS) foresees the use of algorithms (namely screening rules to create risk indicators).

41 See Meng, Xiao-Li (2018). Formally, there is a third component, which refers to the difficulty of the problem at hand. This component was not mentioned in the text as it does only complicate the discussion unnecessarily in this context. Data quantity, data quality and problem difficulty refer to the three statistical concepts of sample size, bias and variance.

One definition of 'data quality' is whether or not the data used are "fit for purpose".[42] Consequently, the quality of data depends strongly on the purpose of their use. The following gives some guidance on what errors can occur in the production and use of data, based on classical social science literature. It can help understand errors in training data used for machine learning systems. Understanding these errors helps identify potential problems in data driven systems, such as bias and discrimination.

5.1. Measurement and representation

There are two general sources of error that are related to data quality when using data for producing statistics: measurement error and representation error. These have been developed and discussed in the area of classical survey research. Together, these two issues present the so-called total survey error framework.[43] However, they also impact on data quality from data sources other than surveys.[44] Issues related to representation and measurement are also of importance when assessing the quality of algorithms that are based on data, because machine learning algorithms make use of aggregated information in the training data they use to learn.

Measurement error

Measurement error refers to how accurately the data used indicate or reflect what is intended to be measured. This concept is most appropriate for survey data, where the question asked to a respondent must be evaluated in terms of how well the question measures what it should. For example, does asking someone about their view on the impact of immigrants on the economy help to measure xenophobia? Additionally, it is important to know how well the answers to questions measure what is really true (e.g. do respondents answer the question honestly?) and how much does editing and reorganising the data distort the measurement of the concept in mind (e.g. due to reducing answer categories or building an index). This is obviously important for survey data, which are sometimes used in AI applications and machine learning. However, questions of measurement need to be addressed for other data sources as well, such as data from administrative sources or from social media. If one intends to measure credit worthiness based on information on income, the questions to be asked are:

- how accurate is the information on income (e.g. information provided by persons themselves or information from declared income from tax authorities); and
- is income a good indicator for credit worthiness as such.[45]

Often, several pieces of information are used in combination to measure one concept. With data, we often only approximate what we intend to measure. For example, how do you define or measure a 'good employee'. Would it be someone who is rarely coming late? For statistics and machine learning, concepts such as creditworthiness or a good employee need to be defined in ways that can be measured. To give another example, if we would like to measure the country of origin of people in a dataset, and this information is not available, their nationality might be used instead. This is called a proxy. However, this means that we do not have a perfect measurement, because we leave out those who obtained the nationality of another country other than their country of origin. In some countries, there are substantial numbers of those obtaining citizenship, hence nationality is not a good proxy for country of origin. Data always only approximate real world phenomena and there is always an error in measurement. It is important to understand how much error is acceptable.

For machine learning systems, measurement error links to the question of which features are included in the training data. Do these include information that does not measure well what it should measure or what should be used to predict outcomes?

The process of labelling outcome data is of crucial importance, particularly when assessing bias in a dataset.[46] Often data have to be labelled by humans, which means that people are tasked to look at specific input data and record the outcome. For example, people are looking at pictures (input data) and

42 Cal, L. and Zhu, Y. (2015).

43 Groves, Robert M. et al. (2009); Groves, Robert M. and Lyberg, Lars (2010) On how to link the total error framework to big data, see Bierer, P. (2016).

44 It refers, for instance, to the automated classification of documents. In the case *Pyrrho Investments v. MWB Property*, British court endorsed, for the first time, the use of so-called predictive coding in the process of document disclosure in the civil procedure. Paragraphs 19-21 of the judgment contain detailed description of steps to be taken before deploying the algorithm for document classification. Those include, among others, the definition of the data set, sample size, batches, control set, reviewers, confidence level and margin of error. See: United Kingdom, England and Wales High Court.

45 In social science this last question is called operationalisation – how well can we measure the concept we want to measure (e.g. credit worthiness) with the indicators/variables/features at hand (e.g. information on income).

46 Barocas, S. and Selbst, A. D. (2016); Borgesius, F. (2018).

provide descriptions of the pictures (output data). This is then used for an algorithm to learn patterns in the data. During the labelling process measurement errors can arise, especially if there is no quality control in the labelling process included.

Representation error

Statisticians are very much concerned with the question of how representative their sample of the population is – for good reason. If the data do not cover well the population it should cover, the resulting statistics will be incorrect (i.e. biased). If the entire population which is intended to be covered with an AI application is not included in the input data used for building the application, there is an error of representation. Hence, it needs to be assessed what the impact of this error is.

The gold standard in statistics is *random selection* of the sample, which means a controlled way of selection of the data units which are used in the analysis. The simplest and a very efficient way of selecting data is through *simple random sampling*, which means that every person in the total population to be covered in the application has an equal likelihood to be selected. When this is the case, only a small sample of persons/units can be used for accurate representations. There is always a certain amount of error that remains, but researchers decide on an acceptable level of accuracy.[47] However, it is often not possible to achieve a situation where the likelihood of selection would be equal for all units. Additionally, even if persons are selected in a representative way, not all of them might agree that their data are used or not all of them provide any information (so-called non-response). This is a real problem when typical big data sources are used for making predictions for the general population. Such data are often from the internet and those using the internet or certain applications online constitute a specific group of the population – i.e. do not represent the total population. For online applications that only target those who use them, the question is less important. Yet, still the question needs to be asked if the data used for building the application can accurately represent future users. For example, if the model is built using data from some time ago and is outdated, the application will not work well due to a representation error.

For automated systems, timeliness of data seems to be of particular importance in this respect, where data should be analysed in near real time. If algorithms are trained based on historical data to predict behaviour or occurrences, the timeliness of the training data is a crucial aspect to assess. One needs to ask if the coverage of the data used for building the application has changed over time and how this would impact on future predictions.[48] For example, if users of an online portal from two years ago are still comparable to and behaving in a similar way as those using the portal today. This is a question of major importance for algorithms used online, since the data and behaviour is being constantly updated and changing.

Another point, which is sometimes overlooked, is that data are often incomplete. Although a data collection might be representative for the target population, some parts of the data are often missing for some people in the data set. For example, information on all forms of income is not available for everyone, when just relying on income data from tax authorities. To give another example, there could be a considerable percentage of online users for a certain application that opted out of having their data used for further use or other purposes. To assess data quality, information on missing data (partial data), and how it is being dealt with, is necessary as well.

Example: Using an algorithm developed in one state in another state

In the case *Wisconsin v. Loomis*, the defendant claimed that the court's reference to the risk assessment report at sentencing violated his constitutional right to due process. The report included scores estimating the risk of recidivism that were calculated by the proprietary algorithm. The defendant alleged, among others, that the software used at sentencing by the Wisconsin authorities had not been cross-validated on (i.e. tested against) a Wisconsin population. Although no violation was found in this case, the Wisconsin Supreme Court ruled that any pre-sentence investigation report must contain information on the software's limitations, including notification that the algorithm compares defendants to a national sample, and not to the population of Wisconsin. Moreover, the court noted that the software must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations (para. 100).

Source: United States, Supreme Court of Wisconsin, State of Wisconsin v. Eric L. Loomis, No. 2015AP157–CR, 13 July 2016.

47 This referes to the level of confidence – or confidence interval – that is deemed acceptable by researchers, based on the number of observations and sampling procedure.

48 In academic research, this is referred to as concept drift. Patterns in data evolve over time, which can render data obsolete over time. See Zliobaite, I., Pechenizkiy, and Gama, J. (2016).

5.2. Reliability and validity

The two sources of errors above – measurement and representation error – are related to two important conditions for data quality and analysis often used in social sciences and survey research: reliability and validity. These two concepts have traditionally been used to describe measurement errors of latent constructs in social sciences (i.e. measurement of certain concepts through several indicators, for example to build an index; or when direct measurement is not possible, using measurement of related issues).[49]

Reliability refers to how stable and consistent measurements are. Validity refers to the question if the data and prediction actually measure what they intend to measure, thus related to errors of representation and measurement. Data that are not reliable have a high variance (i.e. their results vary a lot) and data that are not valid are biased (i.e. they measure the wrong thing). While invalid data often miss the target systematically (bias), unreliable data might have the right target but show too much variation and uncertainty, which means that they often miss the target despite being right on average.

An important aspect in big data analytics is that large amounts of data can be used, which can mitigate errors through having more comprehensive measurement based on many observations. The number of observations used, which are often very large in big data applications, decreases statistical uncertainty. However, without having high quality data, in terms of low errors of representation and measurement, large volumes of data do not increase the validity of measurements. On the contrary, we could miss the target more consistently. This can happen if partial data are used. For example, only data on a certain demographic group, which is systematically different to other groups in the population.

Data quality in the European Statistical System – lessons for AI

Quality of data can be seen as a competitive advantage, particularly in times when increased demand for quick data means that thorough checks on data quality are not undertaken. The European Statistical System (ESS), which is a partnership of Eurostat and national statistical institutes, highlights the need for data quality in its Quality Declaration.* It is based on the European Statistics Code of Practice.** The content of the code can be used as a reference frame for data-based AI applications as well. Principle number 12 of the code of practice makes explicit reference to "accuracy and reliability", requiring that the source data, intermediate results and outputs are regularly assessed and validated, sampling and non-sampling errors are measured and systematically documented, as well as analysis of potential impacts of any revision in the data production to improve the process. Principle number 4 of the code of practice "commitment to quality" can also guide AI applications' data quality management. It refers to four indicators of quality including:

1. The quality policy is made public.
2. Procedures are in place to plan and monitor the quality of the data production process.
3. Product quality is regularly monitored, assessed and reported.
4. The outputs are regularly and thoroughly reviewed, including by external experts where appropriate.

These criteria come from long standing experiences in the production of statistics, which are also related to the use of data for AI systems based on machine learning methods. However, the questions for machine learning are still slightly different and need to be adapted depending on the specificities and goals of each 'use case'. One important difference is that (official) statistics aim at describing the population – such as persons, companies, or countries in the case of statistics on the national economy – according to selected characteristics, potentially identifying correlations and sometimes aiming at causal explanations.

In comparison, machine learning is mainly concerned with predicting the characteristics of one unit, such as one person, one company or one country. This has slightly different implications, because accuracy of the prediction becomes even more important compared to general statistics about population groups.

Developments in the area of machine learning can draw on experiences from other disciplines such as statistics, economics and other social sciences such as psychology and sociology – all of which have developed quality criteria for the purpose of their studies.

* *European Statistical System (2016).*
** *European Statistical System and Eurostat (2011).*

[49] See for example Carmines, E. G. and Zeller, R. A. (1979) or Jackman, S. (2008).

5.3. Dataset descriptions for assessing data quality

An assessment of errors in data begins with a basic understanding of where the data come from and what they cover. A proper assessment of training data can be achieved by providing descriptions of datasets used. Experts in the field of AI and machine learning have proposed to use dataset descriptions (referred to as 'datasheets', or 'nutrition labels') for providing information on the content and quality of datasets.[50] This way data quality issues that are important for understanding how a certain algorithm works can be addressed. It is suggested to describe datasets similar to the way hardware components are described to check if such components comply with industry standards. Datasets including information on people could be similarly described by including detailed information on the dataset creation, composition, data collection process, pre-processing and distribution of the dataset. There is currently no standardised way of describing datasets agreed upon in the field of AI. Such a standardisation would have to allow for flexibility to be able to include the variety of possible data formats and collections used in AI applications. This is important because if data are generated for one purpose, it needs to be assessed if they are also fit for another purpose.[51]

A very good way of describing data is pioneered by European data archives and portals. The UK Data Service, for example, describes how data can be documented based on existing standards.[52] Data archives also make use of international standards and schemes for describing datasets. For example, the Data Documentation Initiative (DDI) is a private international standard for the description of data from surveys and other data sources in social, behavioural, economic and health sciences.[53] It provides guidance for a standardised way of describing datasets (i.e. meta data). The standards developed by this initiative can document data in a way necessary for sharing data for reuse. In order to be able to assess data quality, information on the context of the data collection as well as methodology, data and meta-data level descriptions are needed. This includes information on the methodology used for obtaining the data, such as information on the population and observational units covered in the data, the method of data collection (e.g. interviews, online tracking, etc.), sampling procedures, temporal and geographical coverage. With this information, users of applications can better asses the quality and potential errors of the tool that uses a particular data set. These existing practices, which have been developed over time in disciplines encompassing statistics, offer potential avenues for data quality assurance in relation to AI.

Conclusion: asking the right questions

Assessing AI-related technologies and algorithms from a fundamental rights perspective is a complex task. Assessments of data processing need to go beyond traditional data protection impact assessments. Machine learning systems and algorithms that make use of data require broader and more flexible ways for assessing and addressing data quality in relation to several, and interconnected, fundamental rights.[54]

It is important to increase awareness and knowledge among businesses – both public and private – about how algorithms work. As part of an improved understanding of algorithms, it is important to query what the basis for the development of algorithms is: the data.

The use of algorithms in AI can negatively impact on fundamental rights if the data used to build an AI-system measures the wrong thing. Additionally, harm can be done if the data used to build AI-systems do not represent the population it is used for. As outlined above, the concepts of measurement and representation errors can help our understanding of whether certain data sources are problematic or not. These errors can only be assessed if the data are available and documented appropriately.

There are no agreed standards for data quality assessments for machine learning applications; however, promising research is ongoing.[55] Drawing on experiences from the social sciences, the following questions could serve as a minimum guidance for understanding the quality of data. Answering these

50 Gebru T. et al (2018); Holland S., et al. (2018).

51 In January 2019 European Parliament, the Council of the EU and the European Commission have reached an agreement on a revised Directive on Open Data and Public Sector Information. Proposal highlights that wide range of information in many areas of public authorities' activity constitute a vast, diverse and valuable pool of resources (recital 6). In addition, allowing re-use of documents held by a public authorities will allow to improve the quality of information collected (recital 11). See: European Parliament and European Council (2018a). Data of the EU institutions, agencies and bodies are available at the EU Open Data Portal.

52 For more, see the website of the UK Data Service.

53 For more, see the website of Data Documentation Initiative.

54 On human rights impact assessments, see Mantelero, A. (2018); and for flexible ways of testing machine learning systems for potential discrimination, see Veale, M. and Binns, R. (2017).

55 Gebru, T. et al (2018).

questions can help identify if there are potential fundamental rights problems with the use of an algorithm due to data quality:

- Where do the data come from? Who is responsible for data collection, maintenance and dissemination?
- What information is included in the data? Is the information included in the data appropriate for the purpose of the algorithm?
- Who is covered in the data? Who is under-represented in the data?
- Is there information missing within the dataset or are there some units only partially covered?
- What is the time frame and geographical coverage of the data collection used for building the application?

The quality of data can give the rise of discriminatory or otherwise erroneous machine learning systems. Therefore, an understanding of the data quality can help to understand and mitigate potential problems of such systems. The assessment of algorithms, however, does not stop here. Apart from the data, other parts of the development of algorithms, for example, the way an algorithm works ('the black box') and the predictive power of an algorithm, are equally important and should also be considered. New technologies need a holistic assessment of potential fundamental rights challenges.[56] An evaluation of the training data, as outlined in this paper, should be part of such an assessment.

In relation to AI, the focus on data quality is important because many developers, businesses and public authorities might make use of easily available and potentially free data for developing algorithms.[57] In essence, the purpose of using machine learning algorithms is to be more efficient and save resources. Investing in costly data acquisition might not contribute to saving costs, therefore data sources that are easily available are often preferred. Moreover, data that could be used for the development of algorithms are often not accessible due to copyright restrictions; hence the attraction of data sourced from the internet.[58] As a result, data from the internet are used for purposes other than those originally envisaged, and those analysing or using data were not involved in the production of the data.[59] The consequences of using such data, with respect to fundamental rights, are only just emerging.

For an increased understanding of the impact on fundamental rights, interdisciplinary research is required as the topic combines elements from many different areas, including law, computer science, statistics and social science. This focus paper contributes to the discussion by highlighting what is understood as data quality in survey research, and by applying this to machine learning algorithms. This means that an assessment of AI-related systems, based on algorithms, should include an evaluation of data quality, which can draw on established experience and scientific rigour from the social sciences and survey research.

56 For the guidelines on reporting machine learning models see Luo, W. et al. (2016) and Zook, Matthew, et al. (2017).

57 As an example, among the datasets most commonly used for machine learning is the 'Enron e-mails' that contains approximately 200,000 e-mails exchanged between employees of a Texas-based energy company that went into bankruptcy in 2001 due to accounting frauds. Although for some of them the dataset may be adequate (e.g. automatic categorisation of e-mails or spam detection), the content of Enron e-mails demonstrates gender bias. See: Mohammad, S. M., Yang T. W. (2011).

58 Levendovski, A. (2018).

59 See Groves, Robert (2019).

References

Access Now (2018), *Human Rights in the age of Artificial Intelligence*.

Alpaydin, E. (2015), *Machine Learning*, MIT Press.

Barocas, S. and Selbst, A. D. (2016), 'Big Data's Disparate Impact', *California Law Review*, Vol. 104, No. 671.

Bierer, P. (2016), 'Errors and Inference', in Foster I. et al. (eds), *Big Data and Social Science: A Practical Guide to Methods and Tools*, Chapman and Hall/ CRC, pp. 265-298.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V. and Kalai, A. (2016), *'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings'*, paper given at 30th Conference on Neural Information Processing Systems (NIPS 2016), 5-10 December 2016).

Borgesius, F. (2018), *Discrimination, artificial intelligence, and algorithmic decision-making*, Council of Europe.

Buolamwini, J. and Gebru, T. (2018), *'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification'*, Proceedings of Machine Learning Research, PMLR 81:77-91, pp. 1-15.

Burt, A., et al. (2018), 'Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models'. In, *The Future of Privacy Forum White Paper*.

Cabitza, F. et al. (2018), 'A giant with feet of clay: on the validity of the data that feed machine learning in medicine', , in: Cabitza, F., Magni, M. and Batini, C. (eds), *Organizing for the Digital World, Lecture Notes in Information Systems and Organisation*, Springer, pp. 113-128.

Cal, L. and Zhu, Y. (2015), 'The Challenges of Data Quality and Data Quality Assessment in the Big Data Era', *Data Science Journal*, Vol. 14, No. 2.

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

Carmines, E. G. and Zeller, R. A. (1979), 'Reliability and Validity Assessment,; Thousand Oaks, Vol. 17.

Committee on Civil Liberties, Justice and Home Affairs (2018), *Opinion of the Committee on Civil Liberties, Justice and Home Affairs for the Committee on Industry, Research and Energy on a comprehensive European industrial policy on artificial intelligence and robotics*, (2018/2088(INI)), 11 December 2018.

Council of Europe (2017), *Algorithms and human rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*.

Council of Europe (2018), *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment*.

Crawford K. (2016), 'Artificial Intelligence's White Guy Problem', *The New York Times*, 26 June 2016.

Dastin, J., 'Amazon scraps secret AI recruiting tool that showed bias against women', Thomson Reuters, 10 October 2018.

DDI Alliance, Main page, accessed on 30 April 2019.

Eubanks, V. (2018), *Automating Inequality. How high-tech tools profile, police, and punish the poor*; , St. Martin's Press.

European Commission (2018), *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe*, SWD(2018) 137 final, 25 Arpil 2018.

European Commission, High-Level Expert Group on Artificial Intelligence (2018), *Ethics Guidelines for Trustworthy AI*, version published on 26 April 2019.

European Council (2018), *Council's conclusions from 28 June 2018*, EUCO 9/18.

European Group on Ethics in Science and New Technologies (2018), *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*.

European Parliament (2019), *Resolution of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics*, 2018/2088(INI), P8_TA(2019) 0081, Strasbourg, 12 February 2019.

European Parliament and European Council (2016), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119 (*General Data Protection Regulation*).

European Parliament and European Council (2018a), Proposal for a Directive of the European Parliament and of the Council on the re-use of public sector information (recast), COM/2018/234 final - 2018/0111 (COD), 25 April 2018.

European Parliament and European Council (2018b), Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC, OJ L 295.

European Statistical System (2016), *Quality Declaration of the European Statistical System*.

European Statistical System and Eurostat (2011), *European Statistics Code of Practice for the national and community statistical authorities*.

Eurostat (2019a), table isoc_eb_db, accessed 16 January 2019.

Eurostat (2019b), table "Households - level of internet access (isoc_ci_in_h)", accessed 30 April 2019.

Foster, I. et al. (2016), *Big Data and Social Science: A Practical Guide to Methods and Tools*, Chapman and Hall/CRC.

FRA (2018a), *#BigData: Discrimination in data-supported decision making*.

FRA (2018b), *Handbook on European data protection law – 2018 edition*, Luxembourg, Publications Office.

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

Gebru, T. et al (2018), 'Datasheets for Datasets', arXiv:1803.09010 [cs.DB].

Groves, R. (2019), 'Derivative inquiry: dangers facing fields that use data without producing data', *Provosts's Blog*, Georgetown University.

Groves, R. M. and Lyberg, L. (2010), 'Total Survey Error. Past, present, and future', *Public Opinion Quarterly*, Vol. 74 No. 5, pp. 849-879.

Groves, R. M. et al. (2009), *Survey Methodology. Second Edition*, Wiley.

Holland, S. et al. (2018), 'The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards', Data Nutrition Project.

Internet Governance Forum (2018), *Best Practice forum on Internet of Things, Big Data and Artificial Intelligence*, output report, December 2018.

Jackman, S. (2008), 'Measurement', in Box-Steffensmeier, J. et al. (eds), *The Oxford Handbook of Political Methodology*, Oxford University Press.

Johnston, I., 'AI Robots Learning Racism Sexism and other Prejudices from Humans, Study Finds', *The Independent*, 13 April 2017.

Levendovski, A. (2018), 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem', Washington Law Review.

Life Bureau of the New York State (2019), Insurance Circular Letter No. 1, 18 January 2019..

Luo, W. et al. (2016), 'Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View', *Journal of Medical Internet Research*, 18(12): e323.

Mantelero, A. (2018), AI and Big Data: A blueprint for a human rights, social and ethical impact assessment, *Computer Law and Security Review*, Vol. 34, pp. 754-772.

Meng, X. (2018), 'Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election', *The Annals of Applied Statistics*, Vol. 12, No. 2, pp. 685-726.

Mohammad S. M. and Yang T. W. (2011), *'Tracking Sentiment in Mail: How Genders Differ on Emotional Axes'*, Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, Association for Computational Linguistics., arXiv:1309.6347 [cs.CL].

Prates, M., Avelar, P. and Lamb, L. (2018), *'Assessing Gender Bias in Machine Translation – A Case Study with Google Translate,'* CoRR abs/1809.02208, pp. 1-31.

Raso, F., Hilligoss, H., Krishnaumurthy, V., Bavitz, Ch. and Kim, L., Berkman Klein Center for Internet & Society at Harvard University (2018), *Artificial Intelligence & Human Rights: Opportunities & Risks*, 25 September 2018.

Richardson, R., Schultz, J. and Crawford, K. (forthcoming), 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice', *New York University Law Review Online*.

RightsCon Toronto (2018), The Toronto Declaration.

Salganik M. J. (2017), *Bit by Bit: Social Research in the Digital Age*, Princeton, NJ: Princeton University Press.

Sessions V. and Valtorta M. (2016), 'The Effects of Data Quality on Machine Learning Algorithms', ICIQ 2006, pp. 485-498.

The Future of Privacy Forum (2018), *The Privacy Expert's Guide To Artificial Intelligence and Machine Learning*.

UK Data Service, Document your data, , accessed on 30 April 2019.

United Kingdom, England and Wales High Court, *Pyrrho Investments Limited v. MWB Property Limited*, EWHC 256 (Ch), 16 February 2016.

United Nations (UN), General Assembly (2018), *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, 29 August 2018.

United States, Supreme Court of Wisconsin, *State of Wisconsin v. Eric L. Loomis*, No. 2015AP157–CR, 13 July 2016.

Veale M., Binns R. (2017), 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data,' *Big Data & Society*, pp. 1-17.

Wachter S. and Mittelstadt B. (2019):), 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI', *Columbia Business Law Review*, No. 1.

Wagner, B. (2019), 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems', *Policy and Internet*, Vol. 11, No. 1, pp. 1-18.

Wang J. et al. (2014), *'A Poodle or a Dog? Evaluating Automatic Image Annotation Using Human Descriptions at Different Levels of Granularity'*, proceedings of the 25th International Conference on Computational Linguistics.

Zhao J., Wang T., Yatskar M., Ordonez V. and Chang K. (2017), *'Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints'*, paper given at the 2017 Conference on Empirical Methods in Natural Language Processing, 7-11 September, pp. 2979–2989.

Zliobaite 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*, Springer, pp. 91-114.

Zook Matthew, et al. (2017),: *Ten simple rules for responsible big data research*. PLoS Comput Biol, 13(3): e1005399.

Further information:

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

- #BigData: Discrimination in data-supported decision making (2018)
  http://fra.europa.eu/en/publication/2018/big-data-discrimination
- 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 safeguards 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
- Handbook on European data protection law - 2018 edition (2018)
  https://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law
- Handbook on European law relating to access to justice (2016)
  https://fra.europa.eu/en/publication/2016/handbook-european-law-relating-access-justice
- 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
facebook.com/fundamentalrights
linkedin.com/company/eu-fundamental-rights-agency
twitter.com/EURightsAgency

Publications Office
of the European Union

Print: ISBN 978-92-9474-605-4, doi:10.2811/615718
PDF: ISBN 978-92-9474-606-1, doi:10.2811/546219

TK-01-19-330-EN-C (print); TK-01-19-330-EN-N (PDF)